analytical chemistry

Article

Classification and quantification of microplastic (< 100 µm) using FPA-FTIR imaging system and machine learning

Vitor Hugo da Silva, Fionn Murphy, Jose Manuel Amigo, Colin Andrew Stedmon, and Jakob Strand

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.0c01324 • Publication Date (Web): 18 Sep 2020

Downloaded from pubs.acs.org on September 20, 2020

is published by the American Chemical Society. 1155 Sixteenth Street N.W.,
Washington, DC 20036

Classification and quantification of microplastic (< 100 µm) using FPA-FTIR imaging system and machine learning

Vitor H. da Silva*[,a], Fionn Murphy[a], Jose M. Amigo[b,c], Colin Stedmon[d], Jakob Strand[a]

[a] Department of Bioscience, Aarhus University, Frederiksborgvej 399, 4000 Roskilde, Denmark.

[b] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain

[c] Department of Analytical Chemistry, University of the Basque Country UPV/EHU, PO. Box 644, 48080 Bilbao, Basque Country, Spain

[d] National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet, 2800 Kgs. Lyngby, Denmark.

* Corresponding author: vhs@bios.au.dk

ABSTRACT: Microplastics are defined as microscopic plastic particles in the range from few µm and up to 5 mm. These small particles are classified as primary microplastic when they are manufactured in this size range, whereas secondary microplastics arise from the fragmentation of larger objects. Microplastics are a widespread emerging pollutant and investigations are underway to determine potential harmfulness to biota and human health. However, progress is hindered by the lack of suitable analytical methods for rapid, routine and unbiased measurements. This work aims to develop an automated analytical method for the characterization of small microplastic (< 100 µm) using micro Fourier Transform Infrared (µ-FTIR) hyperspectral imaging and machine learning tools. Partial least squares discriminant analysis (PLS-DA) and soft independent modelling of class analogy (SIMCA) models were evaluated, applying different data pre-processing strategies for classification of nine of the most common polymers produced worldwide. The hyperspectral images were also analyzed to quantify particle abundance and size automatically. PLS-DA presented a better analytical performance in comparison with SIMCA models with higher sensitivity, sensibility and lower misclassification error. PLS-DA was less sensitive to edge effects on spectra and poorly focused regions of particles. The approach was tested on a seabed sediment sample (Roskilde Fjord, Denmark) to demonstrate the method efficiency. The proposed method offers an efficient automated approach for microplastic polymer characterization, abundance numeration and size distribution with substantial benefits for methods standardization.

Plastic represents a wide range of organic polymeric material that is used in a wide variety of applications in modern society. In 2018, the world plastic production reached 359 million tons reflecting its widespread use and continual demand.[1] The ever-growing uses and production of this material entail an increase in the loss of plastic litter into terrestrial and aquatic ecosystems. As a result of their low degradation rate (up to hundreds of years), plastics can persist and accumulate in the environment and are now widely recognized as an emergent environmental contaminant worldwide.[2,3] These synthetic polymers can fragment into smaller pieces under different degradation factors or purposefully manufactured in a small size range being categorized as microplastic when they achieve a size < 5 mm.[4] Although pollution from large plastic debris is the most visible and publicized, the distribution of microscopic plastic fragments is more widespread reaching far from point sources.

Microplastics have been widely reported in different environments worldwide in an ever-expanding range. The concerns about these plastic particles have gained a new dimension due to the potential harm they can cause to biota that ingests them and also potential human health effects.[5,6] Although this emerging pollutant needs to be addressed, the field is still relatively young and robust standardized analytical approaches for sampling, treatment and analysis are still in development. To some extent, this also hinders process in identifying relevant sources and assessing the occurrence, composition, fate and impacts of microplastics. Sampling and isolation of microplastics from the initial environmental matrices (water, soil, sediments, biota or air) are cumbersome operations; mainly due to their microscopic size and their low concentrations in comparison with others interferences such as naturally occurring inorganic and organic matter.[7]

However, significant advances have been made in harmonizing sampling and sample preparation, where some recommendations and protocols have already been published.[4,8,9] Polymer identification is complex and involves the use of different analytical methods or measurement modes depending on particle size fractions. Moreover, microplastic characterization is often carried out in a very labor-intensive manner, where the samples are initially analyzed first by visual inspection (optical microscopy) followed, in general, by spectroscopic measurement on selected particles and often with a focus on microplastic particles with sizes above 100 µm or even 500 µm.[4,9] Therefore, more systematic, holistic and automated analytical approaches focusing mainly in small microplastic would be beneficial.

Several analytical methods have been described in the literature for microplastic characterization such as vibrational spectroscopy [10], thermogravimetric analysis (TGA) and pyrolysis-gas chromatography-mass spectroscopy (py-GC-MS).[11] Fourier-Transformed infrared (FTIR) is, arguably, one of the most common techniques used for microplastic identification[12–14], providing fast and non-destructive measurements, as well as a spectral profile with defined and characteristic peaks for each polymer.[14] Infrared instruments can also be coupled with microscopes (μ-FTIR) for hyperspectral image acquisitions for sample mapping with high precision and spatial resolution.[15] This type of analysis has the advantage of collecting chemical (spectral) and spatial information of several particles at the same time by automated mapping of a sample, allowing the analysis of small microplastic without manual sorting and the estimation of particle features such as their area and diameters.[13] Thus, several types of information can be simultaneously extracted from hyperspectral images in an automated manner.[16]

Simon *et al.*[17] described a method for quantification of microplastic mass and their removal rates at wastewater treatment plants using μ-FTIR hyperspectral imaging. The authors analyzed microplastic in different size range (10-500 μm) and used library searching to characterize the particle/spectra. Renner *et al.*[18] optimized a library search approach with automatic peak detection to assign better the FTIR imaging spectra of microplastics in environmental samples. A series of studies by Primpke *et al.* [19–21] went one step further and developed analytical methods for automated microplastic and microfiber identifications using μ-FTIR imaging and spectra correlation methods. With these strategies, the sample spectra are compared with a reference library for matching and assign the spectra when the similarity surpasses a given threshold. This approach has been applied worldwide for microplastic characterization, such as by Liu *et al.*[22] that identified plastic particles in storm water treatment ponds in Denmark. The spectra similarity are determined by the hit quality index (HQI). The results often rely only on this number that is a potential source of error due to its dependence on the spectral library and the arbitrary threshold applied. Moreover, it is time-consuming since all spectra need to be compared with the whole reference library. Alternatively, multivariate data analysis applying machine learning strategies is well suited and can easily be applied to develop a fully automated process with little or no dependence on a spectral library. Multivariate characterization techniques use both concepts of spectral similarity and dissimilarity which is applied to a database of representative particles. In addition, the models are validated and a statistical evaluation is performed to reduce bias. Furthermore, multivariate models are faster than library searching with results obtained in few minutes once the models have been developed.

Wander *et al.* [23] performed an exploratory analysis of μ-FTIR imaging applying Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) to reduce the data dimensionality and visualize particle similarity. This strategy reduces significantly the analyzed data removing the background information from the images, but further analysis must be performed for spectra characterization. On the other hand, Hufnagl *et al.* [24] developed a method applying a multivariate model (Radon Forest Classifier) for identification of microplastic in environmental samples using hyperspectral images. The authors developed the model for four plastic types and applied spectral descriptors determined by spectroscopy experts for polymer characterization. Thus, it is appropriate to explore more multivariate models for fully automated analysis by applying data pre-processing strategies, exploratory analysis and multiclass models that embrace the reality of plastics that are more found on the environment.

This work aims to develop an analytical method for small microplastic (< 100 μm) characterization using μ-FTIR imaging and multivariate classification tools. Different types of plastic were applied and some morphological features about the particles were also extracted. We propose that with the correct data pretreatment and appropriate selection of hyperspectral data analysis approaches, we can fuel the development of a method to automate the quantification and identification of microplastic from an environmental sample. The proposed method can speed up data analyses, improve quality and reproducibility in polymer assignments and, subsequently, also provide benefits for method standardization.

MACHINE LEARNING BACKGROUND

A brief theoretical background about the different machine learning multivariate models used in this work is presented here using spectroscopy as an example. A more detailed explanation about the multivariate models can be accessed through the references provided.[25]

Principal Component Analysis (PCA). PCA is, by far the most common and well-known multivariate machine learning tool.[26,27] It is well suited for exploratory analysis (unsupervised model) and aims to extract underlying features in a multivariate dataset (in this case, spectral data). This technique reduces the data dimensionality by creating a new space with orthogonal variables describing in order the maximum variance of the original dataset. The new variables are called Principal Components (PCs) it is a linear combination of the originals variables.[26,28,29] The samples projections in the PCs space are called scores, and it indicates the differences or similarities in composition among the samples. In general, samples with similar chemical composition will group in the score plots. On the other hand, the contribution of the original variables in the new orthogonal axes are called loadings, and they describe the influence of the originals variables to a given PC.[26,28,29] The model for PCA decomposition is shown in Eq. (1):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad (1)$$

where X is the measured spectral data (sample by wavenumber), T is the score matrix (sample by component), P^T is the loading matrix (component by wavenumber), and E is the residuals (unexplained data, sample by wavenumber).

In this work, PCA was applied to select the Region of Interest (ROI) in the hyperspectral images before proceeding with an additional model. The ROI is realized on the hyperspectral image to remove any pixel information not related to the plastic particles (i.e. to automatically discard background data from the disk where the sample is placed on). The ROI was selected using the score frequency histogram obtained from a PCA where there is a separation between background and particles driven by the differing spectral properties.[30]

Partial Least Squares – Discriminant Analysis (PLS-DA). PLS-DA is a multivariate classification technique (supervised model) based on the PLS regression model. This widespread classification model correlates the independent matrix (X matrix, spectral data) mathematically with the parameters of interest (Y matrix, class membership). In this technique, the model is built by reducing the data dimensionality like PCA (explaining the data as a linear combination of underlying "components"). However, the new variables in PLS models are called Latent Variables (LV), and the direction is determined by the maximum covariance between X and Y matrices, i.e. both matrices are used to describe the new space. The parameter of interest is a dummy binary matrix indicating the assignment of a given sample to a class, i.e., each row of Y matrix is a vector describing information about the class membership of a given sample, whereas each column encoding information for a specific class.[31–33]

One common strategy to perform PLS models is by using a variation of the NIPALS algorithm, in such a way that X and Y matrices are decomposed as shown in Eq. (2) and (3) by maximizing the correlation between the scores T and U:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad (2)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \qquad (3)$$

where T and U are the score matrices, P^T and Q^T are the loadings matrices, and E and F are residual matrices.

Class assignment of unknown samples is derived from the Y model estimation ($\hat{Y}$) for each pixel and class using the model's regression coefficient (b) expressed in Eq. (4):

$$\boldsymbol{\hat{Y} = TQ^T + G = XW(P^TW)^{-1}Q^T + G = Xb + G} \qquad (4)$$

where W is the weight matrix calculated for each LV and it takes account the contribution of the original variables to explain the Y matrix; G is the residual matrix.

The PLS-DA model estimates are then used to calculate the probability of a sample belonging to a class. As a discriminant method, PLS is univocal, and the samples have to be assigned to only one class using the highest probability as a classification rule. This limits the approach to having to assign each sample to one of the existing classes in the model.[34] In the case of plastic polymers, a considerable variety of types may make this inappropriate or at least rely on comprehensive coverage of polymer types in the training data. Alternatively, PLS-DA applies a softer classification rule employing a threshold-based prediction on probability theory which can be adjusted as a classification rule instead. Therefore, the highest probability information is used and just the pixels above a determined value are assigned in the class estimated by the model. On the other hand, samples below a determined threshold will be unsigned, and if these samples belong to a plastic class not predefined in the model, it can be updated, and new plastic class included.

Soft Independent Modeling of Class Analogies (SIMCA). SIMCA is focused on the analogies among samples from a specific class (category).[32] SIMCA is a PCA-based model and it also follows the Eq. (1). Basically, a PCA model is developed for each nominated class and the boundaries that define the variability of the samples is determined within the multivariate space. In the end, SIMCA presents a collection of PCA models and the distance of unknown samples projected into the PCA models are the base for samples acceptance or rejection in a given category.[32,35] As SIMCA models are developed individually for each class, and several criteria can be followed in order to assign the belonging of one sample to different classes. The most common criteria involve the calculation of the distance of the sample to the center of each SIMCA model developed (class modelling criteria). The distance is calculated using the Hotelling's T-Squared and Q residual obtained from PCA as described by Eq. (5):

$$d_{i,g} = \sqrt{\left(T_{i,g}^2\right)^2 + \left(Q_{i,g}\right)^2} \qquad (5)$$

where ith indicates the samples and gth the class. In this manner, each sample can be individually tested to belong to one, to more than one or to zero classes, depending on a certain threshold imposed of the value d. This might promote that some overlapped boundaries can occur and samples can belong to different categories at the same time. On the other hand, SIMCA model can reject samples with long distance from the model center and, commonly, not classify them in one of the classes predefined.[32,36,37] Finally, the calculated distance is used to estimate the probability of a sample belonging to the classes and a threshold is set as a classification rule.[31,37]

One of the SIMCA model advantages is the natural recognition of samples statistically uncorrelated with any of the classes and, therefore, updates on the developed models can be performed. Besides, SIMCA models only observe information within the class, which makes it less sensitive to detect differences between classes.

MATERIAL AND METHODS

Samples. The polymers used in this study are highlighted in Table 1. They were obtained from an internal reference library at Aarhus University (Denmark) of various plastic materials with polymer composition identified from various sources such as food packaging and construction materials. This material is kept as a reference to aid with the identification of plastic taken from the environment. The listed plastics were selected based on the majority of the standard plastic that is produced throughout the world[1]. They are also currently recommended as the primary polymers to focus on in marine monitoring of microplastic in the Northeast Atlantic.[38]

Table 1. List of polymers used in this work.

Plastic Material	Abbreviation
Polyamide	PA
Polycarbonate	PC
Polyethylene	PE
Polyethylene Terephthalate	PET
Polymethyl Methacrylate	PMMA
Polypropylene	PP
Polystyrene	PS
Polyurethane	PU
Polyvinyl Chloride	PVC

The polymers were ground into tiny pieces using sandpaper and density separated with $ZnCl_2$ solution (1.6 g/cm^3). Density separation is commonly used in microplastic research to separate the polymers from denser particles, making extraction and identification of the microplastic easier.[17] In order to use the plastic size from 100 to 10 µm, each plastic was sieved using stainless steel filters and stored in an ethanol solution. An aliquot of each microplastic solution was pipetted on to an aluminum oxide filter (Anodisc 25 mm diameter, Whatman) with 0.2 µm of pore size and vacuum filtered. Therefore, one membrane for each plastic type was produced. These filters were selected to allow a sample to be filtered through the membrane directly on it.[14]

Two additional membranes with mixtures of all plastic types used in this work were produced to increase the complexity of the samples. One of these membranes was spiked with sediment from Roskilde fjord (Denmark) that was previously cleaned and fractioned to obtain a similar size fraction to the plastics. The spiked sediment was cleaned to represent remaining sediment particles that could be not removed on the sample purification.

A seabed sediment sample from Roskilde Fjord (Denmark) was also analyzed to evaluate the method in an environmental sample. The ~2 cm top surface layer of the sediment sample was collected using a Van Veen grab sampler, and 100 mg of sediment were analyzed. Sample purification was carried out according to Strand *et al.* [39], and the particles were fractioned similarly to the reference plastics in order to obtain the same size fraction. The analyzed muddy sediment sample was characterized by 36% dry weight content, 46% of fine silt and clay particles determined as the < 63 µm fraction and a content of total organic carbon (TOC) of 4%.

All µ-FTIR measurements were collected directly on the membrane filter.

µ-FTIR images acquisition. µ-FTIR hyperspectral images were collected using a Cary 620 FTIR microscope coupled with a Cary 670 FTIR spectrometer from Agilent Technologies. The microscope is equipped with a Focal Plane Array (FPA) detector with 128 x 128 pixels. The analyses were carried out with a 15x Cassegrain objective in transmission mode with a pixel size of 5.5 µm. Samples were measured in the spectral range of 3,800 – 1,300 cm^{-1} with spectra resolution of 8 cm^{-1} applying 32 scans. Clean membrane was used as background applying 128 scans throughout the spectral acquisition process.

For this instrument set up the final images were a square collection of 25 mosaic tiles (5 x 5 mosaic tiles). Therefore, each image was a square with 640 x 640 pixels and 650 wavenumbers (409,600 pixels/spectra per sample).

Data analysis and software. Region of Interest (ROI) for each sample image was initially selected using PCA models to remove any pixels not related to the MP particles. The spatial pre-processing was made using the score frequency histogram obtained from PCA realized on the hyperspectral images.[30] In sequence, the hyperspectral images (three-dimensional) were unfolded in the spatial direction to two-dimensional matrices and only the pixels retained in the ROI were used to build the multivariate classification models.

Different pre-processing strategies were applied on the spectra to reduce any physical and instrumental artefacts such as noise and baseline offset that are not related to the chemical information of the polymers. Standard Normal Variate (SNV), Asymmetric Least Squares (AsLs), Savitzky-Golay 1st and 2nd derivative were all evaluated in this work.[40–42] Normalization of the data was also investigated to remove the intensity variability in the spectra due to the different thickness of the plastic particle that modifies the light pathway. On the pre-processed data, discriminant analysis and class modeling techniques, PLS-DA and SIMCA, respectively, were developed to classify the different microplastic types.

Pixels from each standard sample were partitioned into calibration and prediction subset containing randomly 600 pixels and 400 pixels, respectively. The same calibration and prediction subset were used for PLS-DA and SIMCA models. Another measured mosaic for each plastic class and the samples with plastic mixtures were also used to test the performance of the models.

PLS-DA regression models were developed using the pre-processed spectra. Random cross-validation was carried out to select the optimal number of latent variables (LV) in the regression models. The number of LVs was determined from the classification error. A dummy matrix was used as a response matrix to describe class membership. For SIMCA models, PCA was performed for each plastic class, and random cross-validation was used to determine the optimal number of principal components (PCs). It is worth to mention that all data were mean-centered before both PLS-DA and SIMCA models development.

The assessment of the classification models was done using the misclassification error, sensitivity and sensibility for cross-validation and prediction steps. Sensitivity (Sn), also called True Positive Rates (Eq. 6), describes the fraction of the pixels from a category of interest that were correctly classified by the model. This measurement is used to estimate the probability of pixels that were genuinely belonging to the correct target category. On the other hand, Specificity (Sp), also called True Negative Rates (Eq. 7), describes the fraction of the pixels not coming from the category of interest that was rejected by the model and it estimates the probability of these pixels to be genuinely identified as aliens.[35,43] Misclassification error is the proportion of samples that were incorrectly classified by the models, and it is described by Eq. (8).[43]

$$Sn = \frac{TP}{TP + FN} \tag{6}$$

$$Sp = \frac{TN}{TN + FP} \tag{7}$$

$$Misclass. Error = \frac{FP + FN}{Total\ instances} \tag{8}$$

Where, TP and TN stand for true positive and true negative, respectively, accounting the number of pixels signed correctly as belonging (TP) or not (TN) for a specific class. On the other hand, FP and FN stands the false positive and false negative, respectively, accounting the number of pixels that were wrongly signed as belonging (FP) or not (FN) to a specific class.[44]

Distribution maps for the external mosaics were obtained by refolding the predicted plastic class at each pixel of the hyperspectral image. For these external predicted mosaics, confusion matrix was used to evaluate the model performance encoding the classification rates per plastic category. Particle features such as the diameter were calculated using the rebuild images and the number of particles determined.

All calculations were performed using Matlab software (Mathworks). PLS Toolbox (Eigenvector Research Inc.) was used for model development, Multivariate Image Analysis Toolbox (Eigenvector Research Inc.) for morphological analysis and HYPER-Tools[45] (freely available in www.hypertools.org) for exploratory analysis.

Figure S1 (Supporting information) shows a flowchart describing the data analysis strategy used in this work.

RESULTS AND DISCUSSION

Microplastic µ-FTIR spectra and exploratory analysis. Figure 1 shows the µ-FTIR spectra of all plastic types used in this work and the corresponding data set used to develop the multivariate classification models. The spectral profile for each plastic type shows characteristic absorption bands with some differences among them. Although some spectral differences are appreciated, a high fraction of the spectral signal is overlapping in this spectral range. These spectral differences are highlighted in the fingerprint region (> 2000 cm^{-1}) due to the unique absorption bands for each compound. This region was not fully used in this work as the aluminum oxide filter was not suitably transparent over 1300 cm^{-1}.[14] The main characteristic organic bands for these polymers can be observed in Figure 1a such as the –C-H stretching which arises from 3100 to 2800 cm^{-1} and around 1500 cm^{-1}. The carbonyl peak presented in some plastic can be observed around 1700 cm^{-1}. The broad bands noticed around 3300 cm^{-1} refers to the –NH stretching observed for PA and PU.[14]

Figure 1b shows the spectra set used in this work to develop the multivariate classification models. In general, μ-FTIR hyperspectral images have some interferences as a consequence of the camera focus, instrumental and samples issues, as well as interference by absorption of atmospheric constituents such as the CO_2. The latter absorbs infrared light from 2600 to 2000 cm^{-1}, and this region was therefore not included in the data analysis. The dataset was also influenced by baseline variation as a result of light scattering and partial light reflection due to plastic particle features. These plastic particles are heterogeneous in size, shape and thickness with rough surface causing variation in the spectra and their quality.[46–48] These spectral interferences must be removed before any data analysis since it is not chemically related to the polymers. Thus, different pre-processing strategies were evaluated to overcome these physical effects on spectral data quality before developing the classification models.

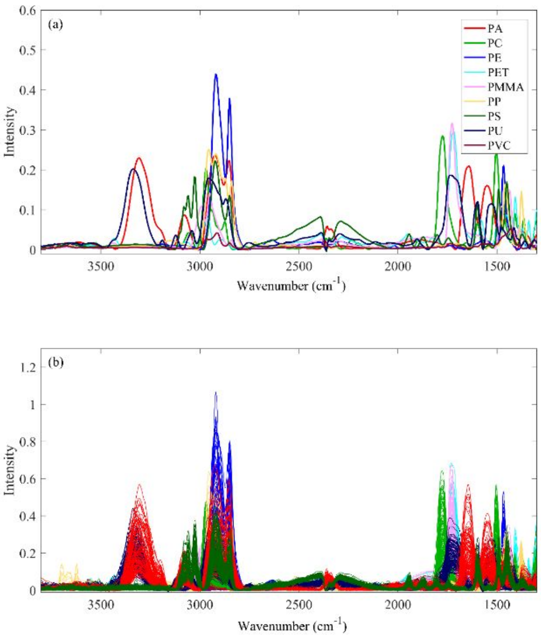


Figure 1. (a) Pre-processed spectra (by AsLs) μ-FTIR spectra of the synthetic organic polymers and (b) the dataset used for PCA.

Exploratory analysis was initially performed on the data set using a PCA model to depict the variability in the dataset (Figure 2). The score plot (Figure 2a) revealed clear differences between the plastics with specific and distinct clustering for each plastic type. The PCA model in which the plastics are more clearly different was obtained using the 2nd derivative spectra with Savitzky-Golay filter applying 2nd order polynomial and 15 window width. In addition, normalization (Inf-Norm) was applied to the data to deal with changes in the spectra intensity related to the optical pathway (plastic thickness). Based on these results, all the subsequent multivariate models were developed using this pre-processing. Figure 2b shows the PCA loadings indicating the main spectral regions used to describe the PCs. These results also demonstrated that the spectral region used in this work provided adequate information to differentiate the plastic types.

Multivariate Classification Models. The SIMCA model was applied to the data set to evaluate its performance to sort different plastic types. Table S1 summarizes the SIMCA model results presenting the statistic assessment of the model for calibration and prediction subset. SIMCA model was built using 3-6 PCs with cumulative explained variance of 75.37 ± 12.03 %. In general, the SIMCA model had an excellent specificity (Sp ≅ 1) for all plastic types, and it demonstrates the model ability to identify the pixels that did not come from the interest category. The results also show the model ability to identify the plastic differences with low misclassification error (1.5%). The average sensitivity was 0.85, which indicates the fraction of pixels that were correctly classified by the model. The remaining pixels were either unclassified or misclassified. As the model had an excellent specificity, these pixels were not categorized by the model with probability value below the acceptance threshold.

These unsigned pixels typically had low quality spectra, and they are originated from areas of a particle that were poorly resolved due to an irregular surface or edge. It is important to note that it is preferential to have unsigned pixels rather than misclassifications to avoid false positive issues. The high similarity between the values obtained for calibration and prediction datasets indicate that the model is not under or over-fitting.

Table S2 shows the confusion matrix for the SIMCA model derived from independent data for each plastic type. These results reiterate the findings based on the calibration and validation datasets. For all plastic types, most of the pixels were classified correctly. These images predictions did not show misclassification issues demonstrating the model ability to sort different plastic types and its ability to discriminate the classes studied.

Some particles from PVC samples were sorted as PC and PP due to sample contamination. These particles were not misclassified because their spectra were analyzed with a spectra library resulting in sample contamination.

Figure 3 shows an example of the prediction image for PA and PU samples presented in Table S2 and demonstrates the issue with pixels with weak quality spectra. These samples should only contain PU and PA, respectively. The center of the particles was correctly classified, but towards the particle edges, the SIMCA model has problems with classification. The same issue was observed for all plastics. The unsigned pixels are typically at the particle edges, and this was explored more below.

PLS-DA was also tested for microplastic classification, and Table S3 summarizes the model results applied for calibration and prediction subset. The model had better sensitivity in comparison with the SIMCA model with an excellent true positive rate for all plastic types. This means that there is a high probability (>95%) of correct classification for every single pixel in its plastic category for PLS-DA model. The PLS-DA approach did not significantly differ from the SIMCA model concerning the ability to avoid false negative classifications, with specificity values also around one. However, the low misclassification rate (< 1% for all plastic types) demonstrates the ability of the PLS-DA model to encompass variability in spectral quality within the classes.

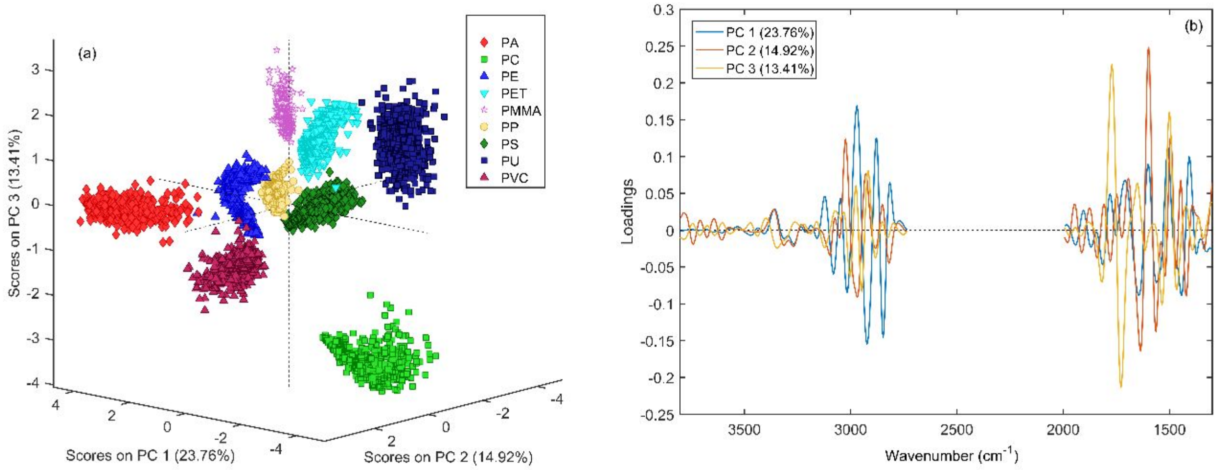


Figure 2. (a) Score plot and (b) loadings obtained from a PCA model developed with all types of plastic.

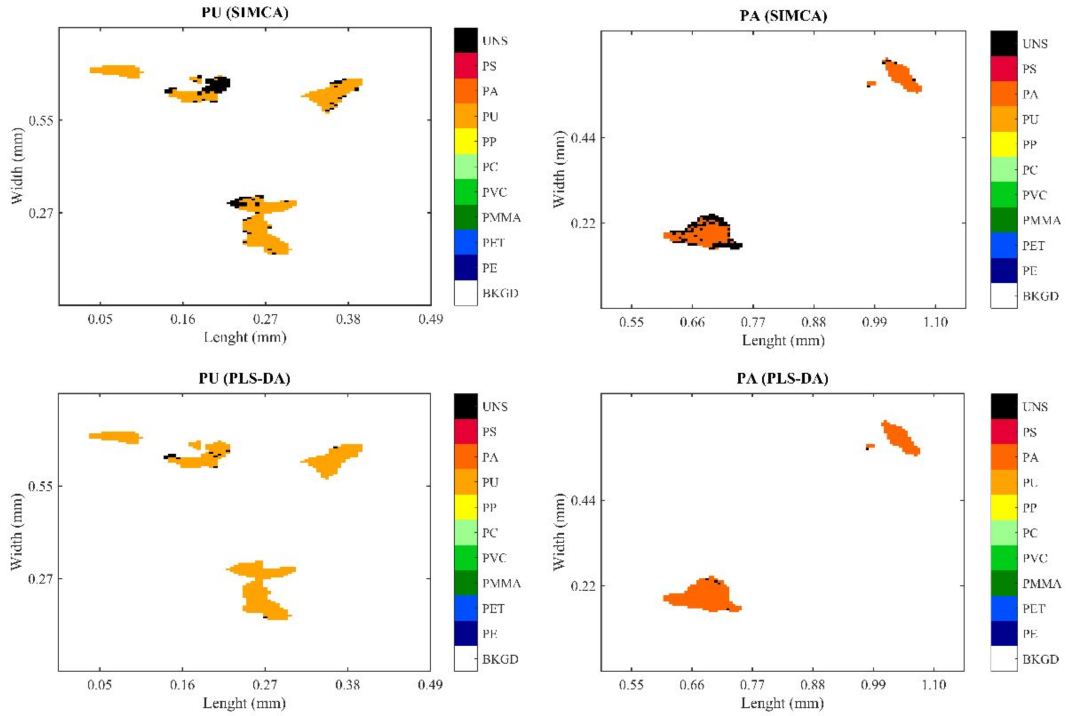


Figure 3. Example of a small region the prediction image obtained from SIMCA (top row) and PLS-DA models (bottom row) for samples containing PU and PA only.

The superior sensitivity of the PLS-DA model can be explained by the nature of the model. PLS-DA models use the difference among the classes to determine the regression coefficient that will define the class features into the multivariate space. The PLS-DA regression coefficients account the spectral information which most describe the class variability and their difference from the other classes. The regression coefficients are orthogonal from each other, and unique information about the class is described. Therefore, the PLS-DA model can better recognize the minor spectroscopic changes caused by the particles features (roughness) and edge effects.

Table S4 shows the confusion matrix for PLS-DA model performed with external sample images for each plastic type. The proportion of unassigned pixels for each plastic type (< 4%) is considerably lower than these obtained from the SIMCA model (5-25%). Figure 3 shows the PLS-DA predicted image for PU and PA, as presented by the SIMCA model. An improvement on the particle prediction is observed for PLS-DA model mainly around the particle edges. This demonstrates the better PLS performance to compute the spectra variability within the particle for all plastics.

Figure 4 shows the mix samples prediction for both PLS and SIMCA models. Both classification methods predicted effectively the core of the particles with 60% and 7% of unsigned pixels for SIMCA and PLS-DA, respectively.

Towards the particle edges, the SIMCA model had lower sensibility due to the spectra quality decay in this direction. The quality of the spectra for each pixel varies mainly for two reasons: (1) the features of the particles such as irregularity, roughness and shape; and (2) the particle thickness that usually changes over the particle length modifying the light pathway and, consequently, the spectra intensity and their quality.[10,17,44] Figure 5 shows the infrared spectra of a single PE particle and the variability in the spectra across the region covered by the particle. In general, the particle edges have poor spectra quality that makes these pixels classification more complex. An additional issue may arise from interference related to the background matrix (Anodisc), which influences the spectra on the particle edges. Although the pre-processing was carefully designed to remove these effects, it is clear that complete removal is not possible and this might influences pixel classification.[44]

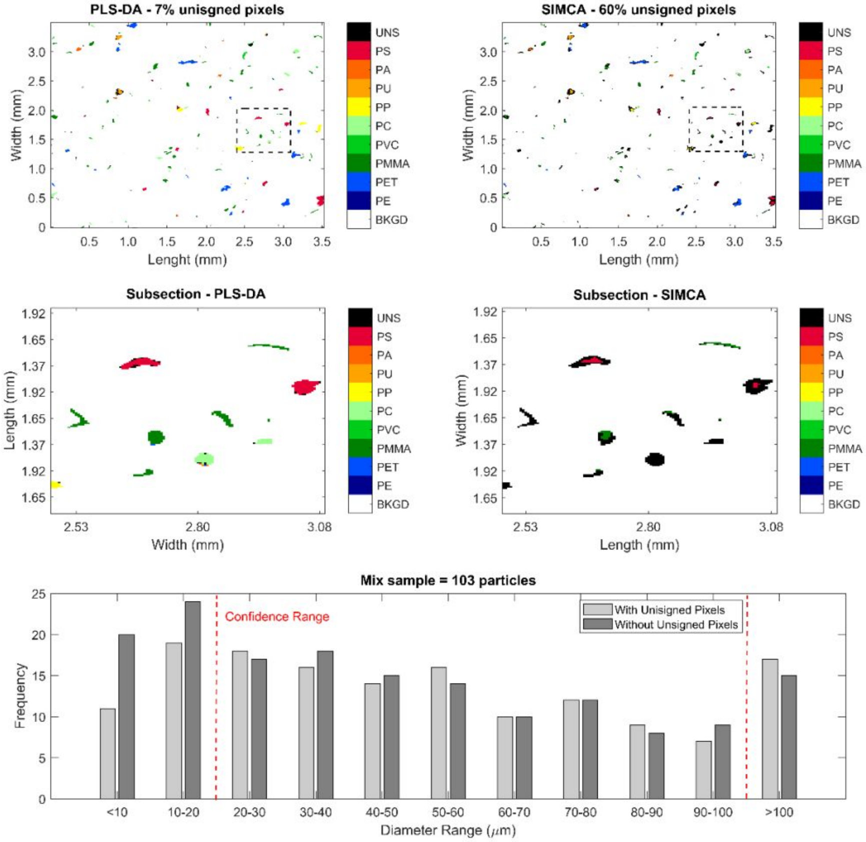


Figure 4. Prediction image obtained for PLS-DA and SIMCA models applied to external Mix sample. Particle abundance numeration and size distribution are provided for the predicted image by PLS-DA model.

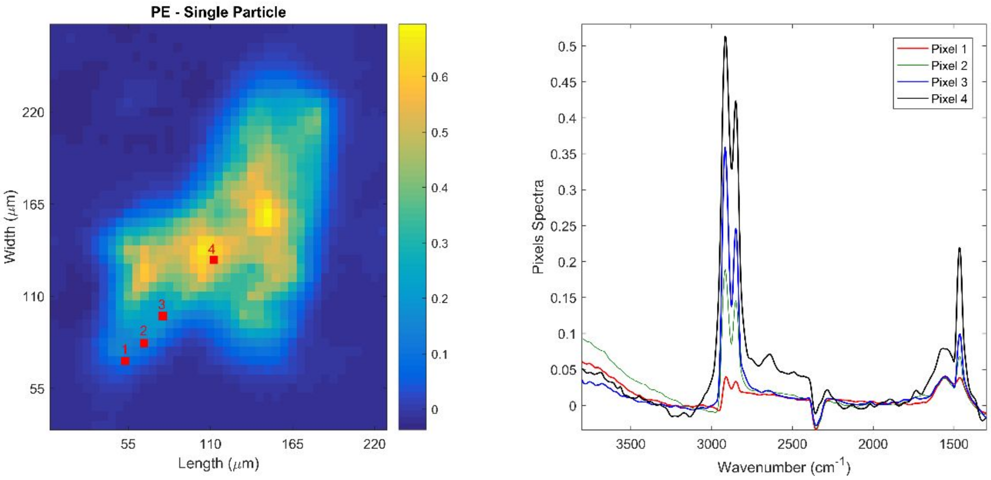


Figure 5. PE single particle with selected pixels and their spectra. The observed spectra are pre-processed by AsLs to obtain a comprehensive observation of the vibration bands.

Particles size distribution was also calculated using the predicted image sample of the plastic mixture by the PLS-DA model (Figure 4). The diameter of the particles was calculated in two different ways, size distribution with (1) and without (2) the unsigned pixels. The differences between these two strategies allow the global estimation of particles fraction that can be reliably estimated. Plastic particles over 100 µm refer to clustering that is a common issue in microplastic determination and it compromises the size distribution estimation. Plastic grouping can be avoided using surfactants such as sodium dodecyl sulfate (SDS).[17] As stated previously, the small particles can suffer from the background removal step and it mainly happened with the particles bellow 20 µm due to the spectra quality in these particles sizes. These small particles are close to the diffraction limit of IR spectroscopy (10 µm) providing poor spectra and low signal-to-noise ration that underestimate the classification and diameter determination.[10,49] The unsigned region also had some pixels classified by the model, and these pixels were counted as new particles after the region removal. For that reason, the smallest size fraction significantly increased the rate. The other size fraction also has rates variation, but it is more related to the new particle category size when the unsigned pixels were removed. Therefore, the confidence range for size distribution was set up between 20 and 100 µm in this work.

Figure S2 shows the predicted mix plastic sample with spiked sediment classified with the PLS-DA model. The sediment was added on the sample to increase its complexity and evaluate the model performance. The sediment particles spiked in this sample have been removed together with the background because their spectra are basically offset and clearly different from the plastic spectra. This result is essential to reinforce the background removal since it reduces the size of the dataset and removes most of the interfering particles. In that way, mainly microplastic spectra are used for multivariate classification, and it also reduces false positives. This strategy is also supported by Wander *et al.* [23] that removed the membrane information to perform the microplastic data analysis in a smaller dataset. The predicted sample image presented 7% of unsigned pixels, and they were mainly pixels on the edge of the particles which is difficult to remove as discussed previously. The size distribution was split by plastic category for individual class distribution.

Evaluation of an Environmental Sample. The approach was also evaluated in a seabed sediment sample (Roskilde Fjord, Denmark) to investigate the model efficiency in an environmental sampling. Figure 6 shows the sediment sample predicted with the PLS-DA model. It worth noting that this approach can be applied in any environmental sample as long as the sample is purified to remove the majority of the interfering particles such as clay, silt and organic matter. This Roskilde Fjord sediment had a high content of organic matter and silt. Although the sediment sample was purified and most of the interfering particles were removed, a significant amount of fine silt particles persisted and were found on the membrane.

The majority of the pixels containing interfering particles were removed together with the background information since their signal is basically offset. The multivariate model predicted PE, PET, PA and PP particles on the sediment sample in the analyzed area (12 mm^2), where from a content for microplastic particles in the sample, 2430 particles per kg of sediment in dry weight could be determined. Figure 6 also shows some few particles not recognized by the model (unsigned), demonstrating its ability to sort particles that are not one of the category classes and were not removed previously together with the background. The model ability to identify unknown particles were also performed using a produced wood sample, and all particles were unsigned by the model (Figure S3). The unsigned particles can be investigated, and new classes can be added on the model in an ever-expanding way. It is also important to remark the importance of the appropriate data pretreatment for correct assignment of microplastic and to avoid false positives results.[50] For instance, the differentiation of animal fur and PA can be obtained with the derivative spectra such as the one applied in this work.

The approach provides a wealth of information that can be useful for characterizing and tracing different sources of microplastics in environmental samples. Having individual size spectra of different polymer types, which can be generated in a few minutes, can potentially transform how microplastics samples are processed in the future. Additionally, the approach removes user bias and offers a method which can be standardized across laboratories. Furthermore, this strategy offers also agility in microplastic analysis with results delivered within few minutes improving time of monitoring programs with a fully automated approach. This proposed strategy is faster than library search since a great part of the spectra information is removed, and the model regression coefficients (one for each category class) are used for microplastic classification. For instance, both prediction and size distribution analyses in this work images (12 mm^2) were performed in around 10-15 minutes using a regular notebook. The analysis can be easily expanded to include the whole filter area, requiring more instrument and computational time.

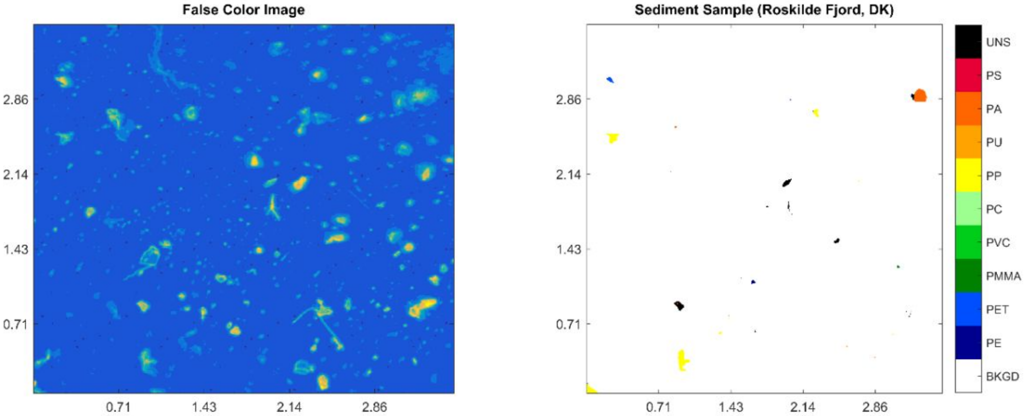


Figure 6. Prediction of sediment sample (Roskilde Fjord, DK) for PLS-DA model.

CONCLUSION

This work derived a machine learning approach for using μ-FTIR hyperspectral images to quantify and characterize microplastic particles. The performance of PLS-DA and SIMCA models were compared on a range of conventional polymers. PLS-DA model performance was superior providing more specific and sensible model for all plastic classes. A mixed plastic sample could be classified with a high degree of precision and was capable of taking into account the effects of deteriorating spectral quality at particle edges. The SIMCA model had difficulties encompassing these artefacts. The derived method also provided particle analysis describing the number of particles and some features such as the diameter for each polymer type. These results demonstrate an approach to fully harness the potential of μ-FTIR hyperspectral for quantifying and qualifying microplastics. The proposed strategy demonstrated efficiency in a sediment sample, the developed model is equally applicable for different environmental samples as long as the samples are purified previously the spectroscopy analysis. Lastly, the proposed method also provided benefits for method standardization sorting the microplastic particles more automated using optimized spectra to extract the most information from the data.

ASSOCIATED CONTENT

Supporting Information

Tables with model results; Flowchart of the applied hyperspectral image analysis strategy; and image prediction results. Contact corresponding author for examples of MATLAB scripts required for carrying out the analysis.

AUTHOR INFORMATION

Corresponding Author

Vitor Hugo da Silva – *Department of Bioscience, Aarhus University, Frederiksborgvej 399, 4000 Roskilde, Denmark.*
Email: vhs@bios.au.dk

ACKNOWLEDGEMENTS

This article was supported by the Danish Center for Environment and Energy (DCE) at Aarhus University and the Danish Center for Research on Plastic Pollution (MarinePlastic) supported by the Velux Foundations. The DTU contribution was supported by the CLAIM (Cleaning Litter by developing and Applying Innovative Methods in European Seas) project which receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 774586.

REFERENCES

(1) PlasticEurope. *Plastics – the Facts 2019: An Analysis of European Plastic Production, Demand and Waste Data*; **2019**.

(2) Wang, W.; Wang, J. Investigation of Microplastics in Aquatic Environments: An Overview of the Methods Used, from Field Sampling to Laboratory Analysis. *TrAC Trends Anal. Chem.* **2018**, *108*, 195–202.

(3) Barnes, D.K.A.; Galgani, F.; Thompson, R.C.; Barlaz, M. Accumulation and Fragmentation of Plastic Debris in Global Environments. *Philos. Trans. R. Soc. B.* **2009**, *364*, 1985–1998..

(4) Kershaw, P.J.; Turra, A.; Galgani, F. Guidelines for the Monitoring and Assessment of Plastic Litter in the Ocean. *GESAMP Reports Stud.* **2019**, *99*, 126.

(5) Zhang, S.; Wang, J.; Liu, X.; Qu, F.; Wang, X.; Wang, X.; Li, Y.; Sun, Y. Microplastics in the Environment: A Review of Analytical Methods, Distribution, and Biological Effects. *TrAC Trends Anal. Chem.* **2019**, 62–72.

(6) Rochman, C.M.; Browne, M.A.; Underwood, A.J.; van Franeker, J.A.; Thompson, R.C.; Amaral-Zettler, L.A. The Ecological Impacts of Marine Debris: Unraveling the Demonstrated Evidence from What Is Perceived. *Ecology.* **2016**, *97* (2), 302–312.

(7) Auta, H.S.; Emenike, C.; Fauziah, S. Distribution and Importance of Microplastics in the Marine Environment: A Review of the Sources, Fate, Effects, and Potential Solutions. *Environ. Int.* **2017**, *102*, 165–176.

(8) González, D.; Hanke, G.; Tweehuysen, G.; Bellert, B.; Holzhauer, M.; Palatinus, A.; Hohenblum, P.; Oosterbaan, L. Riverine Litter Monitoring-Options and Recommendations. MSFD GES TG Marine Litter Thematic Report. *JRC Tech. Report. EUR 28307.* **2016**.

(9) Hanke, G.; Galgani, F.; Werner, S.; Oosterbaan, L.; Nilsson, P.; Fleet, D.; Kinsey, S.; Thompson, R.; Van Franeker, J.; Vlachogianni, T.; Palatinus, A.; Scoullos, M.; Veiga, J.; Matiddi,

M.; Alcaro, L.; Maes, T.; Korpinen, S.; Budziak, A.; Leslie, H.; Gago, J.; Liebezeit, G. Guidance on Monitoring of Marine Litter in European Seas. JRC Sci. *Policy Reports, EUR 2611*. **2013**.

(10) Käppler, A.; Fischer, D.; Oberbeckmann, S.; Schernewski, G.; Labrenz, M.; Eichhorn, K.J.; Voit, B. Analysis of Environmental Microplastics by Vibrational Microspectroscopy: FTIR, Raman or Both?. *Anal. Bioanal. Chem*. **2016**, 408 (29), 8377–8391.

(11) Peñalver, R.; Arroyo-Manzanares, N.; López-García, I.; Hernández-Córdoba, M. An Overview of Microplastics Characterization by Thermal Analysis. *Chemosphere*. 2020, 125170.

(12) Anger, P.M.; von der Esch, E.; Baumann, T.; Elsner, M.; Niessner, R.; Ivleva, N.P. Raman Microspectroscopy as a Tool for Microplastic Particle Analysis. *TrAC Trends Anal. Chem*. **2018**, 109, 214–226.

(13) Serranti, S.; Palmieri, R.; Bonifazi, G.; Cózar, A. Characterization of Microplastic Litter from Oceans by an Innovative Approach Based on Hyperspectral Imaging. *Waste Manag*. **2018**, 76, 117–125.

(14) Löder, M.G.J.; Kuczera, M.; Mintenig, S.; Lorenz, C.; Gerdts, G. Focal Plane Array Detector-Based Micro-Fourier-Transform Infrared Imaging for the Analysis of Microplastics in Environmental Samples. *Environ. Chem*. **2015**, 12 (5), 563.

(15) Salzer, R.; Siesler, H.W. *Infrared and Raman Spectroscopic Imaging, 2^{nd} ed.* Wiley-VCH: Weinheim, 2014.

(16) Prats-Montalbán, J.M.; de Juan, A.; Ferrer, A. Multivariate Image Analysis: A Review with Applications. Chemom. Intell. Lab. Syst. **2011**, 107 (1), 1–23.

(17) Simon, M.; van Alst, N.; Vollertsen, J. Quantification of Microplastic Mass and Removal Rates at Wastewater Treatment Plants Applying Focal Plane Array (FPA)-Based Fourier Transform Infrared (FT-IR) Imaging. *Water Res*. **2018**, 142, 1–9.

(18) Renner, G.; Sauerbier, P.; Schmidt, T.C.; Schram, J. Robust Automatic Identification of Microplastics in Environmental Samples Using FTIR Microscopy. *Anal. Chem*. **2019**, 91 (15), 9656–9664.

(19) Primpke, S.; Dias, P. A.; Gerdts, G. Automated Identification and Quantification of Microfibres and Microplastics. *Anal. Methods*. **2019**, 11, 2138–2147.

(20) Primpke, S.; Lorenz, C.; Rascher-Friesenhausen, R.; Gerdts, G. An Automated Approach for Microplastics Analysis Using Focal Plane Array (FPA) FTIR Microscopy and Image Analysis. *Anal. Methods*. **2017**, 9 (9), 1499–1511.

(21) Primpke, S.; Wirth, M.; Lorenz, C.; Gerdts, G. Reference Database Design for the Automated Analysis of Microplastic Samples Based on Fourier Transform Infrared (FTIR) Spectroscopy. *Anal. Bioanal. Chem*. **2018**, 410, 5131–5141.

(22) Liu, F.; Olesen, K. B.; Borregaard, A. R.; Vollertsen, J. Microplastics in Urban and Highway Stormwater Retention Ponds. *Sci. Total Environ*. **2019**, 671, 992–1000.

(23) Wander, L.; Vianello, A.; Vollertsen, J.; Westad, F.; Braun, U.; Paul, A. Exploratory Analysis of Hyperspectral FTIR Data Obtained from Environmental Microplastics Samples. *Anal. Methods*. **2020**, 12 (6), 781–791.

(24) Hufnagl, B.; Steiner, D.; Renner, E.; Löder, M. G. J.; Laforsch, C.; Lohninger, H. A Methodology for the Fast Identification and Monitoring of Microplastics in Environmental Samples Using Random Decision Forest Classifiers. *Anal. Methods*. **2019**, 11 (17), 2277–2285.

(25) Amigo, J.M. (Ed.). *Data Handling in Science and Technology: Hyperspectral Imaging*, 32. Elsevier: Amsterdam, 2019.

(26) Bro, R.; Smilde, A.K. Principal Component Analysis. Anal. Methods **2014**, 6, 2812–2831.

(27) Granato, D.; Santos, J. S.; Escher, G. B.; Ferreira, B. L.; Maggio, R. M. Use of Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) for Multivariate Association between Bioactive Compounds and Functional Properties in Foods: A Critical Perspective. *Trends Food Sci. Technol.* **2018**, 72, 83–90.

(28) Beebe, K.R.; Pell, R.J.; Seasholtz, M.B. *Chemometrics: A Practical Guide.* Wiley-Interscience: New York, 1998.

(29) Brereton, R.G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant.* Jonh Wiley & Sons Ltd.: Chichester, 2003.

(30) Vidal, M.; Amigo, J.M. Pre-Processing of Hyperspectral Images. Essential Steps before Image Analysis. *Chemom. Intell. Lab. Syst.* **2012**, 117, 138–148.

(31) Biancolillo, A.; Marini, F. Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. *Front. Chem.* **2018**, 6, 576.

(32) Ballabio, D.; Todeschini, R. Multivariate Classification for Qualitative Analysis. In Infrared Spectroscopy for Food Quality Analysis and Control. *Academic Press*. **2009,** 83–104.

(33) Brereton, R.G.; Lloyd, G.R. Partial Least Squares Discriminant Analysis: Taking the Magic Away. *J. Chemom.* **2014**, 28 (4), 213–225.

(34) Calvini, R.; Orlandi, G.; Foca, G.; Ulrici, A. Development of a classification algorithm for efficient handling of multiple classes in sorting eystems based on hyperspectral imaging. *J. Spectr. Imaging.* **2018**, 7, 13.

(35) De Luca, S.; Bucci, R.; Magrì, A.D.; Marini, F. *Class Modeling Techniques in Chemometrics: Theory and Applications.* In Encyclopedia of Analytical Chemistry; John Wiley & Sons, Ltd: Chichester, UK, 2018; 1–24.

(36) Marini, F. Classification Methods in Chemometrics. *Curr. Anal. Chem.* **2010**, 6 (1), 72–79.

(37) Cocchi, M.; Biancolillo, A.; Marini, F. Chemometric Methods for Classification and Feature Selection. *Compr. Anal. Chem.* **2018**, 82, 265–299.

(38) OSPAR Convention for the Protection of the Marine Environment of the North-East Atlantic. *Proposal for a Candidate Indicator on Micro Litter in Sediments.* In Meeting of the Environmental Impact of Human Activities Committee (EIHA); Dordrecht, 2018.

(39) Strand, J.; Lundsteen, S.; Murphy, F. *Microplastic-like Particles in Seabed Sediments from Inner Danish Waters 2015*. Danish Center for Enviroment and Energy. 2019.

(40) Rinnan, Å.; Berg, F. van den; Engelsen, S. B. Review of the Most Common Pre-Processing Techniques for near-Infrared Spectra. *TrAC - Trends Anal. Chem.* **2009**, 28 (10), 1201–1222.

(41) Eilers, P.H.C. A Perfect Smoother. *Anal. Chem.* 2003, 75, 3631–3636.

(42) Boelens, H.F. .; Dijkstra, R.J.; Eilers, P.H.C.; Fitzpatrick, F.; Westerhuis, J.A. New Background Correction Method for Liquid Chromatography with Diode Array Detection, Infrared Spectroscopic Detection and Raman Spectroscopic Detection. *J. Chromatog*r. A. **2004**, 1057, 21–30.

(43) Fielding, A. H.; Bell, J.F.A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. *Environ. Conserv.* **1997**, 24 (1), 38–49.

(44) Amigo, J. M.; Babamoradi, H.; Elcoroaristizabal, S. Hyperspectral Image Analysis. A Tutorial. *Anal. Chim. Acta.* **2015**,

896, 34–51.

(45) Mobaraki, N.; Amigo, J.M. HYPER-Tools. A Graphical User-Friendly Interface for Hyperspectral Image Analysis. *Chemom. Intell. Lab. Syst.* **2018**, 172, 174–187.

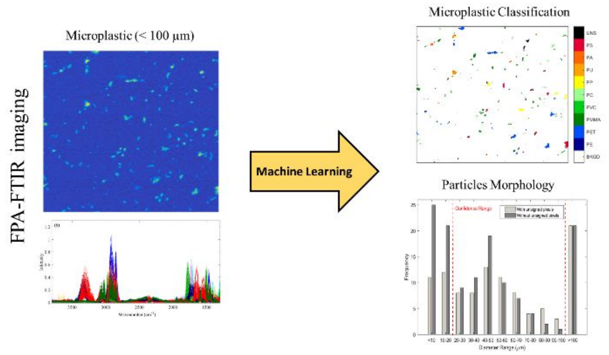
(46) Diem, M. Data *Preprocessing and Data Processing in Microspectral Analysis.* In Modern Vibrational Spectroscopy and Micro-Spectroscopy; John Wiley & Sons, Ltd: Chichester, UK, 2015; 251–282.

(47) Renner, G.; Schmidt, T.C.; Schram, J. Characterization and Quantification of Microplastics by Infrared Spectroscopy. *Compr. Anal. Chem.* **2017**, 75, 67–118.

(48) Lasch, P. Spectral Pre-Processing for Biomedical Vibrational Spectroscopy and Microspectroscopic Imaging. *Chemom. Intell. Lab. Syst.* **2012**, 117, 100–114.

(49) Griffiths, P.R.; Miseo, E.V. *Infrared and Raman Instrumentation for Mapping and Imaging.* In Infrared and Raman Spectroscopic Imaging, 2nd ed. Wiley-VCH: Weinheim. 2014; 1–56.

(50) Renner, G.; Nellessen, A.; Schwiers, A.; Wenzel, M.; Schmidt, T.C.; Schram, J. Data preprocessing & evaluation used in the microplastics identification process: A critical review & practical guide. *TrAC Trends Anal. Chem*. **2019**, 111, 229–238.



For Table of Content Only