Talanta 117 (2013) 483–491

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Automated pipeline for classifying Aroclors in soil by gas chromatography/mass spectrometry using modulo compressed two-way data objects

Mengliang Zhang, Peter de B. Harrington *

Center for Intelligent Chemical Instrumentation, Clippinger Laboratories, Department of Chemistry and Biochemistry, Ohio University, Athens, OH 45701-2979, United States

ARTICLE INFO

Article history:
Received 14 August 2013
Received in revised form
26 September 2013
Accepted 27 September 2013
Available online 7 October 2013

Keywords:
Modulo compression
Polychlorinated biphenyl
Fuzzy rule-building expert system
Partial least-squares discriminant analysis
Fuzzy optimal associative memory
Solid phase microextraction

ABSTRACT

Seven polychlorinated biphenyls (PCBs) commercial mixtures, Aroclor 1016, 1221, 1232, 1242, 1248, 1254, and 1260, were analyzed by gas chromatography/mass spectrometry (GC/MS) combined with solid phase microextraction (SPME). Three pattern recognition methods: a fuzzy rule-building expert system (FuRES), partial least-squares discriminant analysis (PLS-DA), and a fuzzy optimal associative memory (FOAM) were used to build classification models. Modulo compression was introduced for data preprocessing to extract the characteristic features and compress the data size. Baseline correction and data normalization were also applied prior to data processing. Four GC/MS data set configurations were constructed and used to evaluate the classifiers and data pretreatments including two-way modulo compressed, two-way data, one-way total ion current and one-way total mass spectrum. The results indicate that modulo compression and baseline correction methods significantly improved the performance of the classifiers which resulted in improved classification rates for FuRES, PLS-DA, and FOAM classifiers. By using two-way modulo compressed data sets, the average classification rates with FuRES, PLS-DA, and FOAM were $100 \pm 0\%$, $94.6 \pm 0.7\%$, and $96.1 \pm 0.6\%$ for 100 bootstrapped Latin partitions of the Aroclor standards. The classifiers were validated by application to Aroclor samples extracted from soil with no parametric changes except that the calibration set of standards and validation set of soil samples were individually mean centered. The classification rates for the GC/MS modulo 35 compressed data obtained from the Aroclor soil samples with FOAM, FuRES, and PLS-DA were 100%, 96.4%, and 78.6%, respectively. Therefore, a chemometric pipeline for SPME-GC/MS data coupled with chemometric analysis was devised as a fast authentication method for different Aroclors in soil.

1. Introduction

Polychlorinated biphenyls (PCBs) are dioxin-like compounds (Fig. 1) and have 209 possible congeners, about 133 of which were produced as marketable products [1,2]. Aroclors, the trade name of products that are complex mixtures of PCBs were major commercial products of the United States [3]. PCBs had been produced from 1930 to 1975 for the use as insulating oil, lubricating oil and heat medium, and different Aroclors were named as 'Aroclor 12xx', in which the last two digits represent the percentage of Chlorine by mass in the mixture with the exception of Aroclor 1016 that contains 41% Chlorine by mass [4]. In 1977, the manufacture of Aroclors was banned in North America because PCBs are toxic compounds that bioaccumulate and resist degradation in the environment [5]. PCBs were found to have potential carcinogenicity [6], reproductive toxicity [7], adverse growth and development effects [8], irritation and sensitization effects to skin [9], and so on.

Conventionally, capillary columns and highly selective detectors were used to classify different Aroclors based on similarity of PCB peaks or relative response factors (RRFs) of several 'marker' PCB peaks between the sample and Aroclor standards [10]. Capillary gas chromatography with electron capture detectors (GC/ECD) and capillary gas chromatography/mass spectrometry (GC/MS) have been in common use since the 1980s [11]. For classification, the mass spectrometer is more powerful because of

Abbreviations: PCBs, polychlorinated biphenyls; GC/MS, gas chromatography/mass spectrometry; SPME, solid phase microextraction; FuRES, fuzzy rule-building expert system; PLS-DA, partial least-squares discriminant analysis; FOAM, fuzzy optimal associative memory; RRFs, relative response factors; GC/ECD, gas chromatography with electron capture detectors; SIMCA, soft independent modeling by class analogy; PLS-DA, partial least-squares discriminant analysis; KNN, *K*-nearest neighbor; PNN, probabilistic neural network; PDR, projected difference resolution; SVD, singular value decomposition; PDMS, polydimethylsiloxane; TMS, total mass spectrum; TIC, total ion current; mod-35, compressed mass spectra using a modulo divisor of 35; PCA, principal component analysis.

* Corresponding author. Tel.: +1 740 994 0265; fax: +1 740 593 0148.
E-mail addresses: peter.harrington@ohio.edu, harring10@msn.com (P.d.B. Harrington).

0039-9140/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.talanta.2013.09.050

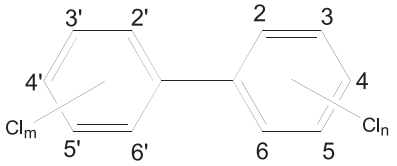

Fig. 1. Chemical structure of polychlorinated biphenyls (PCBs).

its greater informing power compared to GC/ECD [12]. However, many PCB isomers exist in different Aroclors, so lengthy gas chromatographic programming was required to provide enough chromatographic resolution for the identification of unique PCB congeners assigned to a specific Aroclor [13]. For Aroclor identification, manual selection of PCB peak patterns for classification of Aroclors is tedious and error prone because it depends on the skill and experience of the analyst as well as the precision of the measurement.

Chemometric methods especially pattern recognition including soft independent modeling by class analogy (SIMCA) [14], partial least-squares discriminant analysis (PLS-DA), *K*-nearest neighbor (KNN) [15], probabilistic neural network (PNN) [13], and a sequential classification method [16] for classification of Aroclor samples with GC/ECD and GC/MS data have been used because of their superb capability to obtain categorical information from complex data sets. None of these methods have been applied to soil samples. Of pattern recognition methods, the fuzzy rule-building expert system (FuRES) has been developed and utilized to construct reliable and robust classifiers for many applications [17–20]. The FuRES algorithm was described in the early 1990s [18]. PLS-DA is used as a second reference method because of its widespread use for classification. FuRES has three key advantages: (1) interpretation of classification results from inductive classification trees; (2) no adjustable parameters that require optimization; (3) the nonlinear fuzzy logistic function accommodates overlapping classes and fits better with the target categories than other methods that attempt to model discrete data with a continuous linear function [18,21]. The fuzzy optimal associative memory (FOAM) is another powerful fuzzy classifier [22]. The difference between FuRES and FOAM is that, FuRES is a classification method which is better at tweezing out the differences of the features among objects that belong to different classes; but FOAM is a modeling method which exploits the similarities of the features within one class, is softer, and can be used when only one class is known or present [21]. In this current report, three pattern recognition methods, FuRES, FOAM, and PLS-DA were used to classify 7 different Aroclors (i.e., Aroclor 1016, 1221, 1232, 1242, 1248, 1254, and 1260). Using bootstrapped Latin partitions, average classification accuracies are reported with 95% confidence intervals (CI) [21].

Appropriately preprocessing the data before applying the pattern recognition methods is equal or even more important than the selection of the classifier because preprocessing can simplify the classification by reduction of noise, correction of systematic errors, decrease model complexity, and correct for registration errors in the form of retention time drift [23]. For mass spectral features, a modulo compression method, mod-14 feature, was developed and applied for the prediction of molecular substructures in 1968 [24]. 'Condensed mass spectra' so-called by the authors comprised 14 features which indicates the class/type of molecule. These features are characteristic for classes of homologous compounds whose masses differ by multiples of 14 Th accounting for the loss of methylene fragments from the ion. Modulo compression was developed and used prior to constructing and applying the classifiers in this study. The number of features or divisor in the modulo compression was evaluated with the projected difference resolution (PDR) method [25]. The effects of modulo compression preprocessing on classification rates obtained from the FuRES, PLS-DA, and FOAM classifiers and on several data set configurations were evaluated.

The goal of this current study is to develop a classification pipeline for identifying Aroclor 1016, 1221, 1232, 1242, 1248, 1254, and 1260 samples. The Aroclors are complex mixtures comprising many of the same PCB congeners so identifying these mixtures is a challenging problem, however GC/MS has exceptional informing power to meet the demands of this problem. The Aroclor standards were sampled by solid phase microextraction (SPME) of the headspace in the vials. The Aroclor constituents were separated by GC using a 22-min temperature program and detected with an ion trap mass spectrometer. With the application of modulo compression to the two-way GC/MS data, three classifiers (FuRES, PLS-DA, and FOAM) were constructed and used to classify the GC/MS data into one of 7 Aroclors categories. These classifiers were then applied to data that were collected by measuring soil samples that were spiked with the Aroclors.

2. Theory

2.1. Baseline correction

Two baseline correction methods were available from previous studies. Both methods reconstruct a best fitting background mass spectrum by using orthogonal bases constructed from mass spectra of the baselines (i.e., regions where there are no analytical chromatographic peaks). The earlier approach used mass spectra collected at the end of each chromatogram where no chromatographic peaks had eluted and column bleed was the highest [19,25]. A later approach used the mass spectra to construct the basis from the entire chromatogram of a solvent blank [21]. The current project employed the approach of modeling the blank measurements in this case the sampling of an empty vial or a blank soil sample. This approach has the advantage of removing artifact peaks that arose from the SPME fiber, septa in SPME vial cap, column bleed, and from the soil.

Both approaches use singular value decomposition (SVD) to obtain an orthonormal basis that is used to reconstruct the GC/MS baseline. The number of components selected for the basis can significantly affect the baseline correction results especially when the blank GC/MS data sets contain artifact peaks. In the current work, several polydimethylsiloxane (PDMS) chromatographic peaks existed in both the blank and sample GC/MS data sets (Fig. 2), so this baseline correction method was used for all the sample data before any other data preprocessing or processing. To investigate the effect of the number of components on baseline correction for FuRES and PLS-DA classification, different numbers of components for baseline correction were selected and evaluated.

The mass spectra from the chromatogram of a sample were projected onto the bases that were constructed from the mass spectra of the blanks and the basis that had the best fit (i.e., lowest total sum of square error) was selected to correct that sample. Each mass spectrum was projected onto the basis and used to reconstruct a background mass spectrum that was then subtracted from the sample spectrum to correct the background contribution. This procedure was applied to each mass spectrum in the chromatogram.

The baseline correction method used SVD to create the bases. The number of components (vectors) of the basis can affect the baseline correction results. Subtraction of a reconstructed background spectrum from too many components can to a large extent remove background features but also increases the risk of

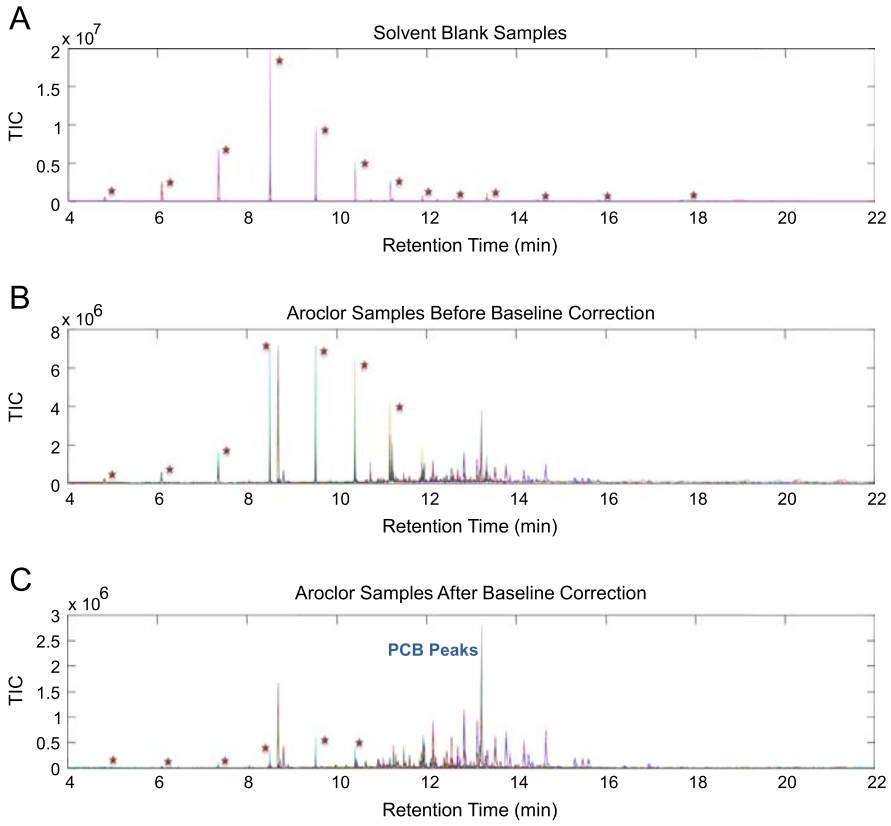

Fig. 2. Total ion current (TIC) chromatograms of solvent blank samples (A), Aroclor samples before (B), and after (C) baseline correction.

overfitting the data and generating negative peaks after baseline correction.

2.2. Normalization

Normalization is used to remove or minimize the effect of variable concentration of the samples and the systematic variations due for instance to varying amounts of analytes or variation in the mass spectrometer sensitivity. All data were normalized to unit vector length [21].

2.3. Modulo compression

Modulo compression was applied after baseline compression. The number and position of Chlorine atoms on the benzene rings are the most important characteristic features for PCB congeners. So the molecular masses of the Chlorine isotopes (35 or 37 Th) would be reasonable to characterize these chlorinated compounds. For two-way GC/MS data sets, 411 mass spectral peaks were compressed to 35 or 37 features using this compression. The features s_j are defined as follows:

$$s_j = \sum_{l=1}^{m} I_{l+mk}, k = 0, 1, 2, \cdots;\ m \text{ is modulo number, e.g., } 35 \text{ or } 37 \quad (1)$$

for which I_{l+35k} is the intensity of m/z $(l+35k)$, m is the divisor and values with same remainder (i.e., modulus) are added together. Different divisors were evaluated by the PDR method, and selected divisors were used to evaluate their effects on classification rates. Because the modulo method sums the ion intensities, a signal noise improvement is obtained as with any other signal averaging approach.

2.4. Projected difference resolution method

Projected difference resolution (PDR) had been used successfully for selecting the optimal parameters for baseline correction [25], comparing the performance of GC/MS, gas chromatography-differential mobility spectrometry (GC-DMS) data [26], and optimizing wavelet filter types and the compression level for the discrete wavelet transform [27].

The PDR method measures the separation of two classes in multivariate data space and gives a figure or merit that resembles chromatographic resolution. First, objects are converted from vectors to scalars by projecting the objects (i.e., vectors) of the two classes onto the difference vector of the two class averages. Then, the absolute value of the difference between the averages of the scalar projections is divided by two times the sum of the standard deviations of the scalar projections [25]. The larger the PDR value, the more separated the two classes are in the multivariate data space. In a previous study, the geometric mean was used for the assessment of overall difference of multiple classes [25]. However in many cases, the two most similar classes among all classes are the most difficult to differentiate and their separations are crucial for the construction of the classifiers rather than classes that are well separated in the data space. Thus, the minimum PDR value of all the pairwise combinations was used in this current study to evaluate the divisors used in modulo compression.

2.5. The FuRES classifier

For the FuRES classifier an inductive classification tree is constructed from fuzzy multivariate rules. Each rule is a branch of the tree that partitions data objects based on a fuzzy logistic value (i.e., consequent of the rule). Each data object is projected onto a normalized weight vector so that the objects are mapped

from a multidimensional space to a single dimension. The scalar projections are then processed by a fuzzy logistic function. The fuzzy logistic values are then used to measure the entropy of classification. The orientation of the weight vector is optimized to achieve the lowest fuzzy classification entropy while constraining the fuzziness through a computational temperature that maximizes the magnitude of the first derivative of the fuzzy entropy with respect to temperature. The objects are partitioned by their fuzzy logistic values until each leaf of the tree consists of objects with the same class designation. Because of the inherent fuzziness of the FuRES classifier, it can be used when classes are overlapped or contain outliers [18].

2.6. The PLS-DA classifier

PLS-DA is commonly used for pattern recognition. The latent variables which are transformed from the independent *X*-data are used for regression with a dependent variable *Y*. The latent variable which has maximum covariance between the *X*- and the *Y*-scores is selected and then the variance of this latent variable is removed by deflation. From the residual matrix, the next latent variable is derived and the variance is removed in the same way. This procedure is continued until the best prediction rates are achieved using an internal boot-strapped Latin partition [21]. The PLS-DA classifier is used as a reference method in this study.

2.7. The FOAM classifier

FOAM is an optimal associative memory (OAM), except fuzzy encoding of the data is used. The data are first encoded as a binary image as opposed to a vector using grid encoding. A fuzzy function is then applied to the binary grid. The FOAM method has three steps, first binary encoding by a gridding function, reconstruction with an orthogonal basis, and then decoding to vector format by reversing the gridding procedure. SVD is used to form the orthogonal basis set from fuzzy encoded data. The basis is built for each class. By comparing the least-squared error between the reconstruction with each basis and the original data an object can be assigned to the best fitting class [22].

The standard grid size is 100 grids between the maximum and minimum intensities of the mean-centered calibration set. The fuzzy function was a triangular membership function with 19 elements that ranged between 0.1 and 1.0 with a 0.1 increment. This function is applied to each variable with respect to intensity although other applications might apply across the variables or use a 2-dimensional function. For forming the basis of grid values all the components are used to fully characterize the data. For the inverse grid process, a data point is assigned to the intensity of the maximum grid value for each variable.

3. Materials and methods

3.1. Reagents

Aroclor 1016, 1221, 1232, 1242, 1248, 1254, and 1260 standards at a concentration of 100 $\mu g\,mL^{-1}$ in methanol were purchased from AccuStandard, Inc. (New Haven, CT). The SPME fiber was coated with polydimethylsiloxane (PDMS, 100 μm film thickness) and used with 20-mL headspace glass vials and crimp seals with PTFE/silicone septa. The SPME fibers, vials, and seals were all obtained from Sigma-Aldrich Co. LLC. (St. Louis, MO). The clean soil was purchased from RT Corp (Laramie, WY).

3.2. Instruments

A Thermo Finnigan PolarisQ quadrupole ion trap mass spectrometer/Trace GC system with a Triplus AS2000 autosampler (San Francisco, CA, USA) was used to collect all the experimental data. The GC/MS system was controlled using the XCalibur software version 2.0.7 provided by Thermo. Analytes were separated using a SHRXI-5 MS capillary column (5% diphenyl/95% dimethylpolysiloxane cross-linked, 30 m × 0.25 mm i.d., 0.1 μm film thickness) from Shimadzu Scientific Instruments Inc. (Columbia, MD). All the data were processed using MATLAB R2012b (MathWorks Inc., Natick, MA).

3.3. Data collection

Aroclor standard solutions at concentrations of 0.3, 1, and 3 $\mu g\,mL^{-1}$ in duplicate were prepared by dilution with methanol from a 100 $\mu g\,mL^{-1}$ stock solution. A 50 μL aliquot of each standard solution was added to a 20-mL headspace glass vial. After the vial was sealed, it was incubated at 100 °C for 5 min. Then a PDMS fiber was exposed to the headspace for 25 min. The fiber was thermally desorbed in the GC injector at 280 °C for 5 min. The analytes were separated using the following oven temperature program at a constant flow of 1 mL min^{-1}: 50 °C, hold for 1 min, ramp at 20 °C min^{-1} to 280 °C, hold for 10 min. The transfer line and ion source temperatures were both maintained at 280 °C. Full scan mode was selected for the mass spectrometer and the scan range was from mass-to-charge ratio (m/z) 140 to 550. Six replicates of an empty vial and each Aroclor sample were prepared and determined individually. A random block experimental design was used so that replicates of each Aroclor sample were distributed with respect to time and would characterize any variation that occurred during the course of the experiment.

Two concentrations at 0.3 and 1.0 $\mu g\,g^{-1}$ of soil samples in duplicate for each Aroclor were prepared by using Aroclor standards and blank soil samples ($4 \times 7 = 28$ soil samples). The PCBs tend to bind to the surface of soil particles strongly because of their lipophilicity. A saturated potassium dichromate solution was used to free the PCBs from the soil and prepared by dissolving an excess of potassium dichromate in 6.0 M sulfuric acid. Soil samples of 0.5 g were added to the SPME vial and 2 mL of saturated potassium dichromate-sulfuric acid solution were added to the SPME vial for the sampling of the PCBs from the soil matrix. After the vial was sealed and vortexed for 10 s, it was incubated at 100 °C for 5 min. Afterwards, a PDMS fiber was exposed to the headspace for 25 min. Two-way data sets were collected using the same GC/MS program as previously described. Five blank soil samples without any Aroclor were treated in the same way. The blank soil sample data were used for correcting the baselines of the Aroclor soil samples.

3.4. Data format

The two-way GC/MS data sets were initially acquired as RAW files. The RAW files were converted to the network common document format (CDF) with the 'File Converter Tool' in the XCalibur Software. The CDF files were read directly into MATLAB using their netcdf tools.

For further data processing, the data sets were binned by retention time from 4.10 to 22.00 min with a 0.01 min increment and binned by mass-to-charge ratios from 140 to 550 Th with a 1 Th increment. Therefore, each two-way GC/MS object comprised 1801 × 411 data points in which 1801 rows corresponded to the retention times and 411 columns corresponded to mass-to-charge ratios. A total of 81 two-way GC/MS data (i.e., 42 for Aroclor

standard samples and 6 blanks and 28 for Aroclor soil samples and 5 blanks) were collected.

3.5. Data preprocessing

The data sets were pretreated by baseline correction (30 components), modulo compression, and normalization. For comparison purposes, one-way total mass spectrum (TMS) and one-way total ion current (TIC) objects were constructed by summing across the complementary way of the two-way GC/MS data object.

No retention time alignment was required in this study. A set of experiments using a polynomial fitting method [21] demonstrated that no improvement was obtained by aligning the standard samples and the soil samples, and analyses of the TIC did not demonstrate any run-to-run retention time variation among the chromatograms.

3.6. Data processing

After data preprocessing, FuRES, PLS-DA, and FOAM classifiers were compared according to their prediction accuracies through validation with the bootstrap Latin partition method. Bootstrapping is a resampling method which is applied to reach a good estimation of prediction performance with limited numbers of samples. The Latin partition method randomly divides the data sets into training and prediction sets and both sets contain equal distributions of classes. Each sample is used once and only once for prediction and the prediction results are pooled for all the objects [28]. Four Latin partitions were bootstrapped 100 times to evaluate the models, so each object was used once for prediction and three times for model building.

To evaluate the effectiveness of modulo compression, four data set representations were evaluated: two-way GC/MS, two-way GC/MS after modulo compression, one-way TIC, and one-way TMS. These different data configurations were used to construct FuRES, PLS-DA, and FOAM classifiers for each Latin partition. The pooled prediction results from FuRES, PLS-DA, and FOAM classifiers were averaged across the 100 bootstraps and were reported with 95% confidence intervals.

4. Results and discussion

4.1. SPME-GC/MS analysis

The primary purpose of this study was to develop an automatic method or pipeline that does not require human expertise to identify different Aroclors. SPME is an efficient technique to extract PCBs from a wide variety of matrices [29] and it integrates sampling, extraction, concentration, and sample introduction into one step. The optimization of SPME conditions will be described in another publication. To summarize, a 100 μm PDMS coating the SPME fiber was selected because of its relatively high extraction efficiency; for a reasonable extraction time and acceptable response, 100 °C as the extraction temperature and 25 min as extraction time were used; to avoid cross-contamination, the SPME fiber was desorbed at 280 °C for 5 min which is the maximum operation temperature for a 100 μm PDMS SPME fiber.

Fig. 2 gives the superimposed TIC chromatograms of the blank and Aroclor samples. The PDMS peaks appear at retention times of 4.83 min, 6.10 min, 7.36 min, 8.51 min, 9.53 min, 10.41 min, 11.19 min, 11.90 min, 12.61 min, 13.43 min, 14.46 min, 15.83 min, and 17.68 min. These mass spectra conformed to the PDMS spectra in the National Institute of Standards and Technology (NIST) database (Version 2.0). The size of the PDMS interference peaks and their effect on classification were reduced after baseline correction.

4.2. Baseline correction

The baseline variations may cause errors in classification. The baseline chromatograms are not always constant so the baseline correction simply by subtracting an average blank spectrum does not work very well, nor does it work well for two-way GC/MS data objects because of column bleed [25]. In the present study, a blank object from each block of experiments was collected. The blank samples were either the SPME of the empty headspace vials or the SPME sample of soils without any Aroclors added. The mass spectra from each blank run were used to construct an orthonormal basis set. Therefore, the bases characterized variations from the septum peaks, the SPME fiber peaks and column bleed, and for the soil the same experimental artifacts along with any peaks that arose from the soil were corrected. The soil and standards were corrected separately so a soil basis was never used to correct the standard and vice versa.

Different numbers of components for the baseline correction were systematically investigated and their effects on classification rates with FuRES and PLS-DA are reported in Fig. 3. Twelve numbers (e.g., 0, 2, 3, 5, 8, 10, 20, 30, 50, 100, 150, and 200) of components were selected arbitrarily to construct the bases for the baseline correction. The models were constructed from four Latin partition and ten bootstraps. Increased numbers of components improved classification rates for both FuRES and PLS-DA. The classification rates with FuRES using modulo compressed data sets was 100% after 30 or more components were used. The quality of the data was significantly improved after baseline correction. Some negative mass spectral peaks arose from the baseline correction which when calculating the total ion current results in a decrease in chromatographic peak intensities. However, when a baseline is properly modeled there should be an equal number of positive and negative residuals about the baseline. For the unprocessed data, all the mass spectral peaks are positive so that when the mass spectra are totaled the total ion current produces larger chromatographic peak intensities.

The benefit of baseline correction can also be visualized by the comparison of TIC chromatograms of Aroclor samples before and after baseline correction in Fig. 2. The PDMS peaks were reduced or even eliminated and the baselines for all the TIC chromatograms were corrected. As a result of this evaluation, 30 components were chosen for baseline correction to treat all the Aroclor data sets before any further data preprocessing or processing.

4.3. Modulo compression

After baseline correction of the two-way GC/MS data sets, the mass spectral features were compressed with different divisors or numbers of features (modulo numbers in Eq. (1)). The PDR method was used to study the effect of the number of features on modulo compression.

For evaluating the choice of divisor for the modulo compression, 21 PDR values among the seven classes/Aroclors (21 pairwise combinations) were obtained for each divisor. The pair that had the minimum PDR value represented the most similar pair of classes. The modulo divisor was evaluated from 1 to 60 and the minimum PDR values are plotted with respect to divisor in Fig. 4. The greatest minimum PDR value was observed for modulo number 35 which indicated that the best separations were achieved for the most abundant isotope of Chlorine. The PDR method is very efficient, and it required 12 min to complete the 60 evaluations.

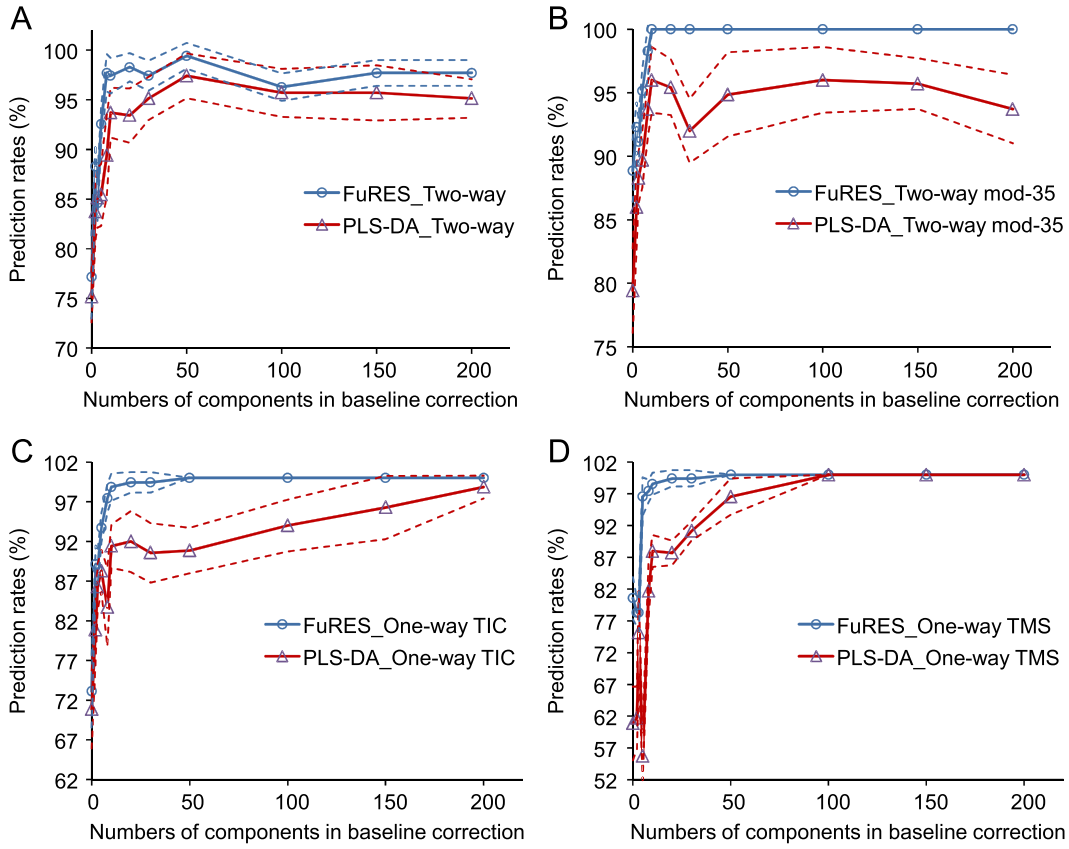

Fig. 3. FuRES and PLS-DA prediction rates with respect to the numbers of components for baseline correction using different data sets representations. The 95% confidence intervals are given in dashed lines.

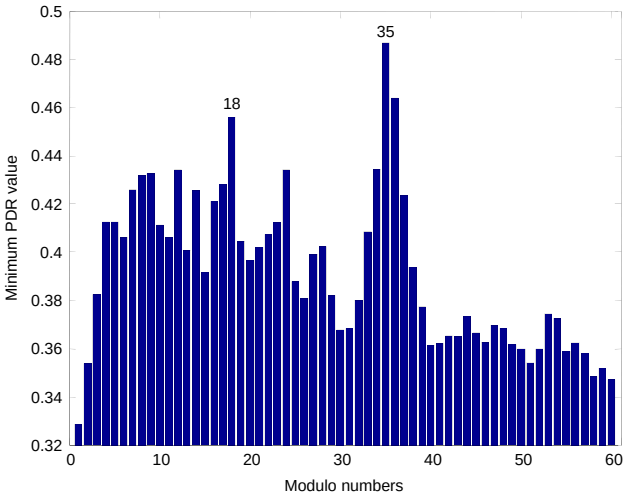

Fig. 4. The minimum PDR value among 7 classes in modulo compression data sets varies with the modulo numbers.

The average classification rates with FuRES, PLS-DA, and FOAM by using two-way modulo compression data with different modulo numbers were also evaluated for 100 bootstrapped Latin partitions. Divisors of 35 and 37 were selected to model the two most abundant isotopes of Chlorine. Smaller and larger divisors 19 and 60 were selected arbitrarily. A divisor of 18 was also chosen because it was half of the average of 35 and 37. The PDR value for the divisor 18 is relatively high compared to the other values in Fig. 4. Smaller divisors achieve greater compressions and a consequent improvement in signal to noise ratio although with a loss of informing power. The choice of divisor significantly affected the classification rates of all three classifiers, see Table 1. Using a modulo divisor of 35 (mod-35) to compress the mass spectra provided the best classification rates, which was expected from the PDR study.

Table 1
FuRES, PLS-DA and FOAM Classification Rates with 95% Confidence Intervals Obtained by Using Different Numbers of Modulo Features in Modulo Compression with Baseline Correction of Training Data Using 3 Components.

Modulo numbers	Average classification rates[a] (%)		
	FuRES	PLS-DA	FOAM
18	90.9 ± 0.6	84.5 ± 0.8	84.4 ± 0.8
19	86.3 ± 0.8	81.2 ± 0.8	75.9 ± 1.0
35	93.0 ± 0.6	86.0 ± 0.7	85.8 ± 1.0
37	90.4 ± 0.7	84.3 ± 0.7	81.0 ± 1.0
39	86.7 ± 0.8	82.9 ± 0.8	75.8 ± 0.9
60	85.8 ± 0.9	82.0 ± 0.8	75.2 ± 0.8

[a] Averages were calculated from 100×4 bootstrapped Latin partitions.

When the mass spectra were compressed with mod-19 or mod-60, the classification rates were worse. Among these tests, using mod-35 yielded the best classification performance which is in accord with the results from the PDR evaluation. Henceforth, mod-35 compression was used. Using classification rates to optimize modulo numbers is clear and straightforward but requires longer evaluation times. At least 1.5 h was needed for just one bootstrap evaluation which would be enough time to complete 450 evaluations using PDR. However, by reducing the number of bootstraps would also decrease the evaluation time.

Table 2
FuRES and PLS-DA Classification Rates with 95% Confidence Intervals Obtained by Using Different Data Representations with Baseline Correction of Training Data.

Classifiers	Average classification rates[a] (%)			
	Two-way mod-35[b]	Two-way	One-way TIC	One-way TMS
FuRES	92.8 ± 0.7	86.3 ± 0.7	88.4 ± 0.7	78.7 ± 0.8
PLS-DA	86.0 ± 0.7	84.0 ± 0.9	83.7 ± 0.8	74.3 ± 1.3

[a] Two-way mod denotes two-way data sets after modulo compression; Two-way denotes two-way data sets; One-way TIC denotes total ion current data sets; One-way TMS denotes total mass spectra data sets. The number of components in baseline correction was 3. Averages were calculated from 100 × 4 bootstrapped Latin partitions.
[b] The mass spectra were compressed to 35 features using modulo compression.

The benefit of modulo compression for the data was evaluated by the classification rates with FuRES and PLS-DA for different data configurations. Among them, the classification rates of FuRES and PLS-DA using mod-35 compressed data were better than the other data representations.

All the data representations were evaluated by the classification rates after baseline correction for which 3 components were used. With 3 components selected for the baseline correction, so that the prediction results were good but not perfect and the classifiers would be more sensitive to different data representations, e.g., 2-way versus 1-way. The average classification rates of FuRES and PLS-DA for the two-way modulo compressed data (mod-35) were 92.8 ± 0.7% and 86.0 ± 0.7%; for two-way data sets were 86.3 ± 0.7% and 84.0 ± 0.9%; for one-way TIC were 88.4 ± 0.7% and 83.7 ± 0.8%; for one-way TMS were 78.7 ± 0.8% and 74.3 ± 1.3%, respectively, for 100 × 4 bootstrapped Latin partitions, see Table 2.

Moreover, the classification results for FuRES were always better than those of PLS-DA when comparing the classification rates by using the same data sets. However, PLS-DA was implemented in an automated fashion [30] so it would function similar to FuRES which is parameter free. The training data was partitioned into 2 and bootstrapped 10 times. The number of components that gave the lowest average prediction error for the bootstrapped training data was selected for building a model that was then used for the prediction data. By tweaking the number of PLS-DA components for each training-prediction set pair, an improvement in prediction results might be obtained, but that would be a prodigious task when implemented in a bootstrap evaluation.

4.4. FuRES, PLS-DA, and FOAM classification

Three classifiers, FuRES, PLS-DA, and FOAM, were constructed and evaluated with four Latin partitions and 100 bootstraps for the classification of seven standard Aroclor sample data sets using the optimized parameters for data preprocessing as previously described. All four data representations were pretreated by baseline correction using 30 components, and normalization. For two-way modulo compressed data, the mass spectra were compressed using a divisor of 35. The classification rates of the three classifiers for the four different data sets are listed in Table 3. The FuRES, PLS-DA, and FOAM predictions were averaged across the 100 bootstraps for the different data representations. Therefore all three classifiers using two-way mod-35 and one-way TIC representations worked well while the classification of two-way modulo compressed data sets by FuRES gave the best classification rate.

Fig. 5 is a principal component analysis (PCA) score plot for the two-way mod-35 data of the 7 Aroclor standard samples. After the data preprocessing of baseline correction, modulo compression, and normalization; Aroclor 1260, 1254, 1221, and 1248 were individually separated very well from the other Aroclors. The similarity of the data among Aroclor 1016, 1232, and 1242 lead to overlapping 95% confidence ellipses in the PCA score plot. By comparing the complete PCB congener distributions for Aroclor 1016, 1232, and 1242, it was found that the compositions of PCB congeners for these Aroclors were very similar [3]. Actually Aroclor 1016 was manufactured by the fractional distillation of Aroclor 1242, which excluded the more highly chlorinated

Table 3
FuRES, PLS-DA and FOAM Classification Rates with 95% Confidence Intervals Obtained by Using Different Data Representations with Baseline Correction Using 30 Components of the Calibration Data.

Classifiers	Average classification rates[a] (%)			
	Two-way mod-35[b]	Two-way	One-way TIC	One-way TMS
FuRES	100 ± 0	97.9 ± 0.4	99.5 ± 0.3	99.5 ± 0.3
PLS-DA	94.6 ± 0.7	96.2 ± 0.6	95.0 ± 0.6	89.9 ± 0.8[c]
FOAM	96.1 ± 0.6	81.5 ± 0.9[c]	97.0 ± 0.5	87.5 ± 0.5[c]

[a] Two-way mod denotes two-way data sets after modulo compression; Two-way denotes two-way data sets; One-way TIC denotes total ion current data sets; One-way TMS denotes total mass spectra data sets. Averages were calculated from 100 × 4 bootstrapped Latin partitions.
[b] The mass spectrum were compressed to 35 features in modulo compression.
[c] indicates that the classification rates smaller than 95%.

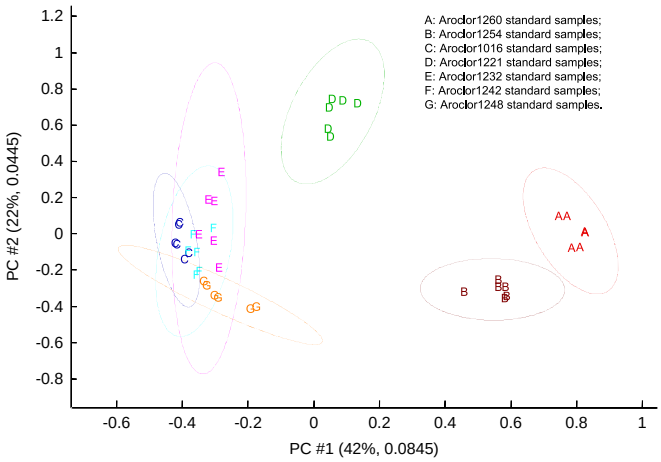

Fig. 5. Principal component analysis score plot for two-way modulo compressed (35 features) data sets of seven Aroclor standard samples. The 95% confidence intervals are represented by the ellipses.

Table 4
Classifier Prediction Rates (%) of Standard Aroclor Soil Samples with No Parametric Changes.

Classifiers	Classification rates[a]				
	Two-way mod-35[b]	Two-way	One-way TIC	One-way TMS	Two-way mod-18[c]
FuRES	96.4	85.7	92.9	85.7	82.1
PLS-DA	78.6	78.6	89.3	71.4	78.6
FOAM	100	78.6	89.3	89.3	82.1

[a] Two-way mod denotes two-way data sets after modulo compression; Two-way denotes two-way data sets; One-way TIC denotes total ion current data sets; One-way TMS denotes total mass spectra data sets. The number of components in baseline correction was 30.
[b] The mass spectrum were compressed to 35 features in modulo compression.
[c] The mass spectrum were compressed to 18 features in modulo compression.

congeners and Aroclor 1232 was a 50:50 blend of Aroclor 1242 and 1221 [3]. Nevertheless, the FuRES classifier successfully classified all 7 Aroclors for 100 bootstraps without a single error and the PLS-DA and FOAM classifiers were able to classify the seven Aroclors with classification rates above 95%.

4.5. Classification of Aroclor soil samples

FuRES, PLS-DA, and FOAM classifiers were constructed using the initial 7 Aroclor standard sample data sets after baseline correction (30 components), normalization, modulo compression (mod-35). The constructed FuRES, PLS-DA, and FOAM classifiers were used on the new data sets of Aroclor soil samples. The prediction rates of the 28 Aroclor soil samples with FuRES, PLS-DA, and FOAM classifiers using four data set representations are listed in Table 4. One Aroclor 1232 was misclassified as Aroclor 1242 by the FuRES classifier built from the mod-35 data (Fig. 6A) and all Aroclor soil samples were accurately classified by the FOAM classifier established by using the mod-35 data (Fig. 6B). Three misclassifications were found when Aroclor soil samples were classified with PLS-DA by using one-way TIC data, and the results were worse for the other data representations. The PCA score plot of the 7 Aroclor standard samples as training sets and seven Aroclor soil samples as prediction sets are given in Fig. 7. Although the prediction rates for the 7 groups of Aroclor soil samples were 100%, the PCA scores of the prediction data were not superimposable with the PCA scores obtained from the training data, see Fig. 7. This disparity could be ascribed to the variations of PCB extraction efficiency and matrix effect differences for the Aroclor standard samples and Aroclor soil samples. However, this study has demonstrated the capability to classify Aroclor soil samples using Aroclor standard samples and thereby eliminating the requirement to collect a variety of soil samples for building classifiers.

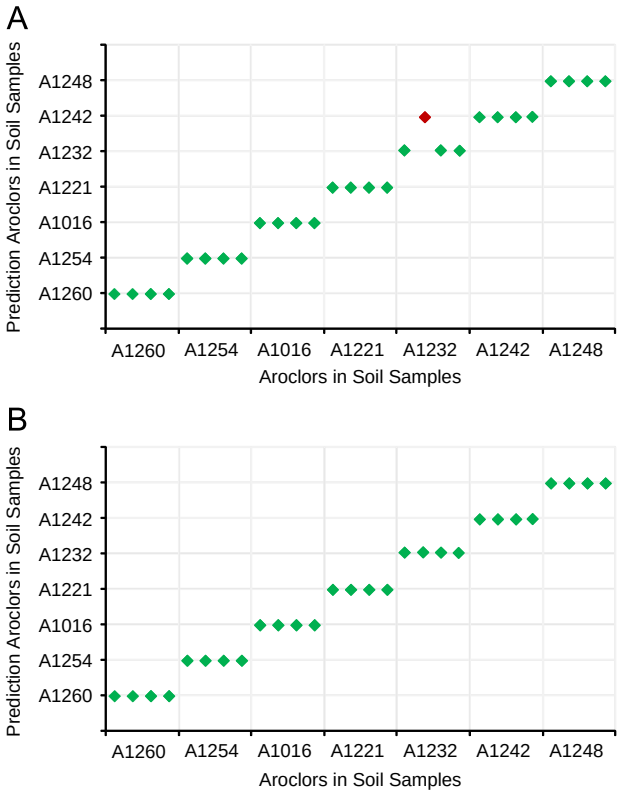

Fig. 6. Prediction plot of Aroclor soil samples using FuRES (A) and FOAM (B) classifiers constructed by modulo compressed (35 features) Aroclor standard data sets.

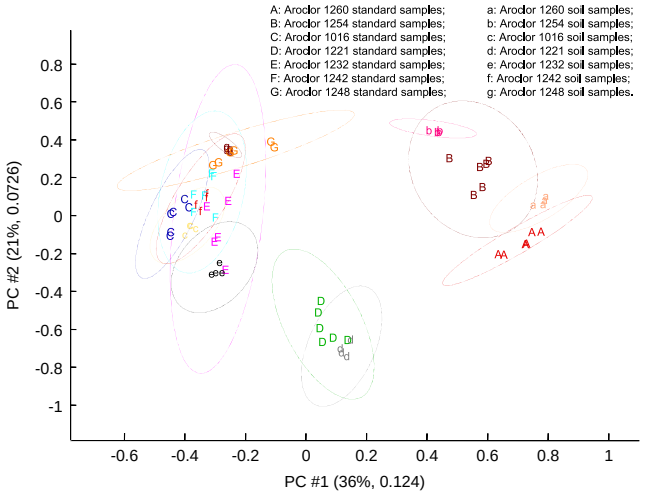

Fig. 7. Principal component analysis score plot for two-way modulo compressed data sets of seven Aroclor standard samples and seven Aroclor soil samples. The 95% confidence intervals are represented by the ellipses.

5. Conclusions

An automatic classification method for 7 Aroclors using SPME-GC/MS and chemometrics was developed in this study. The modulo compression was introduced and evaluated for the classification of complex mixtures of PCBs by GC/MS for the first time. The performance of the classifiers was improved after mod-35 compression was applied. The effect of the number of components in the baseline correction was investigated. The application of a larger number of components in the baseline correction successfully removed the baseline and significantly reduced the influence of artifact peaks that arose from the SPME fiber or the septa (i.e., PDMS peaks). After the Aroclor data sets were treated by baseline correction using 30 components and the mass spectral features were compressed from 411 to 35 by modulo compression, the average classification rates for 100×4 bootstrapped Latin partitions were $100 \pm 0\%$ with FuRES, $94.6 \pm 0.7\%$ with PLS-DA and $96.1 \pm 0.6\%$ with FOAM. This method compressed the data to 8.5% of its original size. Compressed GC/MS data may be beneficial for miniaturized and portable GC/MS units that lack sensitivity and the capacity to store large volumes of data. Modulo peak integration may improve signal-to-noise ratios and sensitivity for miniaturized instruments while maintaining some selectivity. With an appropriate baseline correction method, all the other data representations, two-way data, one-way TIC data, and one-way TMS data, can be used to establish classifiers with satisfactory classification rates without the need to run standards in soil or other complex matrices. Twenty-eight Aroclor soil samples were classified by these classifiers with the classification rates of 100% for FOAM and 96.4% for FuRES. This study proposed a method for the classification of 7 Aroclors at low concentration (1 ppm and below) in soil without using a long GC program for full separation of PCB congeners nor were any soil samples used in building the classifiers or configuring the automatic classification system.

Acknowledgments

The research is funded by a Grant from the US Department of Energy, Office of Environmental Management, Portsmouth/Paducah Project Office.[#] The Center for Intelligent Chemical Instrumentation and Department of Chemistry and Biochemistry at Ohio University are acknowledged for the financial support. Dr. Cevdet

Demir and Zhengfang Wang are thanked for their comments and suggestions.

[#]The project was supported by US Department of Energy. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US Department of Energy Office of Environmental Management Portsmouth/Paducah Project Office, or of the Voinovich School of Leadership and Public Affairs at Ohio University.

References

[1] M.M. Ulaszewska, E. Zuccato, E. Davoli, Chemosphere 83 (2011) 774–782.
[2] D. Consonni, R. Sindaco, P.A. Bertazzi, Environ. Int. (2012) 1989–2010.
[3] G.M. Frame, J.W. Cochran, S.S. Bowadt, Hrc-J. High Res. Chrom. 19 (1996) 657–668.
[4] S. Safe, Crit. Rev. Toxicol. 13 (1984) 319–395.
[5] J. Castro-Jimenez, C. Gonzalez, J. Environ. Monit. 13 (2011) 894–900.
[6] R.D. Kimbrough, M.L. Doemland, J.S. Mandel, J. Occup. Environ. Med. 45 (2003) 271–282.
[7] A. Axmon, L. Rylander, U. Stromberg, L. Hagmar, Int. Arch. Occup. Environ. Health 73 (2000) 204–208.
[8] P. Stewart, J. Reihman, E. Lonky, T. Darvill, J. Pagano, Neurotoxicol. Teratol. 22 (2000) 21–29.
[9] Y.L. Guo, M.L. Yu, C.C. Hsu, W.J. Rogan, Environ. Health Perspect. 107 (1999) 715–719.
[10] P.J. Sather, M.G. Ikonomou, R.F. Addison, T. He, P.S. Ross, B. Fowler, Environ. Sci. Technol. 35 (2001) 4874–4880.
[11] D. Muir, E. Sverko, Anal. Bioanal. Chem. 386 (2006) 769–789.
[12] H. Lohninger, K. Varmuza, Anal. Chem. 59 (1987) 236–244.
[13] G.R. Magelssen, J.W. Elling, J. Chromatogr. A 775 (1997) 231–242.
[14] D.L. Stalling, T.R. Schwartz, W.J. Dunn, S. Wold, Anal. Chem. 59 (1987) 1853–1859.
[15] W.J. Dunn, D.L. Stalling, T.R. Schwartz, J.W. Hogan, J.D. Petty, E. Johansson, S. Wold, Anal. Chem. 56 (1984) 1308–1313.
[16] C.Y. Ma, C.K. Bayne, Anal. Chem. 65 (1993) 772–777.
[17] P.D. Harrington, N.E. Vieira, P. Chen, J. Espinoza, J.K. Nien, R. Romero, A. L. Yergey, Chemometr. Intell. Lab 82 (2006) 283–293.
[18] P.B. Harrington, J. Chemometr. 5 (1991) 467–486.
[19] X.B. Sun, P. Chen, S.L. Cook, G.P. Jackson, J.M. Harnly, P.B. Harrington, Anal. Chem. 84 (2012) 3628–3634.
[20] P.D. Harrington, J. Kister, J. Artaud, N. Dupuy, Anal. Chem. 81 (2009) 7160–7169.
[21] Z.F. Wang, P. Chen, L.L. Yu, P.D. Harrington, Anal. Chem. 85 (2013) 2945–2953.
[22] B.W. Wabuyele, P.D. Harrington, Appl. Spectrosc. 50 (1996) 35–42.
[23] O.E. Denoord, Chemometr. Intell. Lab. 23 (1994) 65–70.
[24] L.R. Crawford, J.D. Morrison, Anal. Chem. 40 (1968) 1469–1474.
[25] Z.F. Xu, X.B. Sun, P.D. Harrington, Anal. Chem. 83 (2011) 7464–7471.
[26] Y. Lu, P. Chen, P.B. Harrington, Anal. Bioanal. Chem. 394 (2009) 2061–2067.
[27] P. Chen, Y. Lu, P.B. Harrington, Anal. Chem. 80 (2008) 7218–7225.
[28] P. Rearden, P.B. Harrington, J.J. Karnes, C.E. Bunker, Anal. Chem. 79 (2007) 1485–1491.
[29] R. Montes, M. Ramil, I. Rodriguez, E. Rubi, R. Cela, J. Chromatogr. A 1124 (2006) 43–50.
[30] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.