Authentication of Organically and Conventionally Grown Basils by Gas Chromatography/Mass Spectrometry Chemical Profiles

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2013

Impact Factor: 5.64 · DOI: 10.1021/ac303445v · Source: PubMed

Zhengfang Wang
U.S. Food and Drug Administration
10 PUBLICATIONS **52** CITATIONS

Liangli Lucy Yu
University of Maryland, College Park
155 PUBLICATIONS **5,061** CITATIONS

Peter B Harrington
Ohio University
180 PUBLICATIONS **2,182** CITATIONS

analytical chemistry

Article

pubs.acs.org/ac

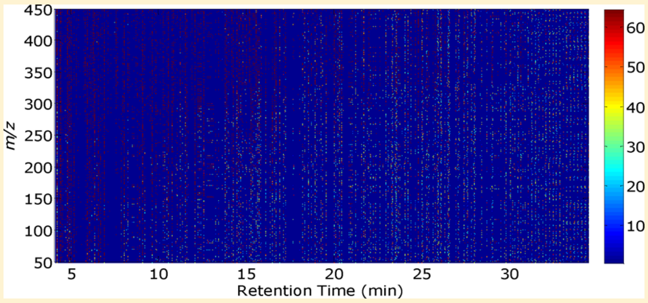
Authentication of Organically and Conventionally Grown Basils by Gas Chromatography/Mass Spectrometry Chemical Profiles

Zhengfang Wang,[†] Pei Chen,[‡] Liangli Yu,[§] and Peter de B. Harrington*,[†]

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

[‡]Food Composition and Methods Development Lab, Beltsville Human Nutrition Research Center, Agricultural Research Services, United States Department of Agriculture, Beltsville, Maryland 20705-2350, United States

[§]Department of Nutrition and Food Science, College of Agriculture and Natural Resources, University of Maryland, College Park, Maryland 20742-7640, United States



ABSTRACT: Basil plants cultivated by organic and conventional farming practices were accurately classified by pattern recognition of gas chromatography/mass spectrometry (GC/MS) data. A novel extraction procedure was devised to extract characteristic compounds from ground basil powders. Two in-house fuzzy classifiers, i.e., the fuzzy rule-building expert system (FuRES) and the fuzzy optimal associative memory (FOAM) for the first time, were used to build classification models. Two crisp classifiers, i.e., soft independent modeling by class analogy (SIMCA) and the partial least-squares discriminant analysis (PLS-DA), were used as control methods. Prior to data processing, baseline correction and retention time alignment were performed. Classifiers were built with the two-way data sets, the total ion chromatogram representation of data sets, and the total mass spectrum representation of data sets, separately. Bootstrapped Latin partition (BLP) was used as an unbiased evaluation of the classifiers. By using two-way data sets, average classification rates with FuRES, FOAM, SIMCA, and PLS-DA were 100 ± 0%, 94.4 ± 0.4%, 93.3 ± 0.4%, and 100 ± 0%, respectively, for 100 independent evaluations. The established classifiers were used to classify a new validation set collected 2.5 months later with no parametric changes except that the training set and validation set were individually mean-centered. For the new two-way validation set, classification rates with FuRES, FOAM, SIMCA, and PLS-DA were 100%, 93%, 97%, and 100%, respectively. Thereby, the GC/MS analysis was demonstrated as a viable approach for organic basil authentication. It is the first time that a FOAM has been applied to classification. A novel baseline correction method was used also for the first time. The FuRES and the FOAM are demonstrated as powerful tools for modeling and classifying GC/MS data of complex samples, and the data pretreatments are demonstrated to be useful to improve the performance of classifiers.

Basil, a common name for the herb *Ocimum basilicum* of the family Lamiaceae, has been cultivated in tropical regions of Asia for more than 5000 years.[1] It has become popular throughout the world because of its culinary use. Current research into the health benefits of basil components have also revealed its potent antioxidant, antiviral, and antimicrobial properties, as well as potential uses in treating cancers.[2−4]

Chemical components of basils of different cultivars have been widely studied. Generally, basils are characterized by their high contents of oxygenated monoterpenes (e.g., eucalyptol and linalool) and phenylpropenes (e.g., eugenol and estragole).[5,6] Extraction techniques (e.g., solid-phase microextraction, steam-distillation, and dialysis)[7,8] for isolating volatile compounds from basils are commonly employed for sample preparation, because the strong flavor of basils arises from a complex aroma profile. Compositional diversity has been found among basil essential oils of different varieties.[9]

Organic basils are preferred by consumers compared to their conventionally grown counterparts. Organic foods are cultivated without using synthetic pesticides and chemical fertilizers, and their processing process does not involve irradiation, industrial solvents, or chemical food additives.[10] Although agronomical practices and environmental conditions have been found to affect the composition of important compounds in plants to a certain extent,[9,11] volatiles profiles of basils have not been reported to be distinctly affected by the farming method.[12] Hence, this study aims to classify organically and conventionally grown basils by their chemical profiles, using pattern recognition.

To obtain an adequate chemical profile from basil components for the purpose of authentication, the sample preparation method must be carefully chosen. The extraction approach should be rapid and effective. Distillation and dialysis are commonly used extraction methods.[8] Although these two

Received: December 18, 2012
Accepted: February 11, 2013
Published: February 11, 2013

methods have been applied to determine essential oil content and composition of basil, both of them are limited by the long operation time (i.e., greater than 1 h) and the requirement of a large amount of sample (i.e., greater than 1 g).[8]

Solid-phase microextraction (SPME) is simple and time-saving, but the high selectivity of SPME fibers toward specific chemicals can be one of its disadvantages. The SPME technique has only been used for the extraction of volatile components from the headspace of basils,[7,12] and no significant differences were found among the aroma profiles of basil cultivars grown under organic and conventional conditions.[12] Sample discrimination, saturation, the lack of robustness, and low reproducibility due to the aging of the SPME fiber increase the cost of this method.[13]

Therefore, direct solvent extraction was chosen, because of its simplicity, flexibility, reproducibility, and large selectivity.[14] This study does not focus on specific compounds in basils, but aims to obtain a general profile of major components. The extracts obtained with a volatile solvent can be readily preconcentrated and subjected to gas chromatography/mass spectrometry (GC/MS) analysis; however, very volatile compounds may be undetected if they elute during the solvent delay period during the GC/MS run.

After GC/MS two-way data sets were obtained, computational data pretreatment should be performed. Baseline correction and retention time alignment are often performed to improve models. Chromatographic baselines are usually not constant during chromatographic runs, due to the use of temperature programs and the thermal degradation or vaporization of the stationary phase.[15] In addition, unavoidable run-to-run retention time variations can wreak havoc with multivariate models. Other undesirable variations are mainly due to the detector nonlinearity,[16] ionization suppression,[17] and changes in instrument parameters (e.g., temperature and gas flow fluctuations and matrix effects). The computational process of classifying data objects is referred to as pattern recognition, in which an algorithm (i.e., a classifier) assigns a class to an object, based on the description of the object. Commonly used classifiers can be crisp or fuzzy. In this study, four classifiers, i.e., the fuzzy rule-building expert system (FuRES),[18] the fuzzy optimal associative memory (FOAM),[19] soft independent modeling by class analogy (SIMCA), and partial least-squares discriminant analysis (PLS-DA), were evaluated. A brief introduction to these methods is provided in the Theory section.

FuRES and FOAM are fuzzy classifiers while SIMCA and PLS-DA are crisp classifiers. SIMCA and FOAM are both examples of supervised modeling methods, while FuRES and PLS-DA are examples of supervised classification methods. Classification methods are better at tweezing out the differences of the features among objects that belong to different classes. Modeling methods exploit the similarities of the features within one class, are softer, and can be used when only one class is known or present. The FOAM and SIMCA modeling methods are constructed using only data from a single class to build each model.

Modeling methods can be implemented in a nonparametric approach by assigning prediction objects to the best fitting model. The alternative approach is to use parametric boundaries so that objects can belong to multiple classes or no class. Although the parametric approach was used to provide a better comparison between modeling methods and classification methods, the parametric approach will always perform worse than the nonparametric approach because of type I errors (i.e., some predetermined fraction of data objects will fall outside the class boundary). However, the key advantage of modeling methods is the ability to reject objects that belong to none of the known classes. The parametric approach with a 95% confidence interval is used in this study so that the classification accuracy should be limited to 95%.

For unbiased evaluations of classification models, the prediction accuracy obtained by using each of the four classifiers (i.e., FuRES, FOAM, SIMCA, and PLS-DA) was validated by using 100 × 3 bootstrapped Latin partitions (BLPs).[20,21] This approach allows a generalized prediction error to be obtained that does not depend on the selection of the prediction set. In no aspect of the data preprocessing or model building were the prediction objects used for adjusting or optimizing any parameters.

Because a random block experimental design was used with each block being a separate day, classifiers were expected to be robust with respect to time. These validated classifiers were used to classify a new collection of basil samples extracted 2.5 months after the initial experiment with no parametric changes to the procedure. The classification methods were effective while the modeling methods were found to perform poorly. After mean-centering the training set (data sets collected 2.5 months earlier) and mean-centering the validation set (data sets collected 2.5 months later), results were significantly improved. Good performance on this study demonstrates the robustness of the established procedure.

Two fuzzy (i.e., FuRES and FOAM) and two crisp (i.e., SIMCA and PLS-DA) classifiers were evaluated with bootstrapped Latin partition for differentiating basil samples from organic and conventional farming practices, using the two-way chromatographic and mass spectral fingerprints obtained from the GC/MS analysis of basil extracts as well as their total ion chromatograms and mass spectra. Classifiers were built with two-way or one-way data representations, and then classification rates obtained from different classifiers were compared for these different representations. These classifiers were readily used to classify a new collection of unknowns with no parametric changes to the procedure, to demonstrate their robustness with respect to time.

THEORY

Normalization. The purpose of normalization is to remove systematic variations of the data due to varying amounts of sample that may arise from different injections or extractions. Each data object is normalized to unit vector length using eq 1

$$xn_{i,j} = \frac{x_{i,j}}{\sqrt{\sum_{j=1}^{n} {x_{i,j}}^2}} \quad (1)$$

for which the measured intensity, $x_{i,j}$, of object i and measurement j comprises an element of a data matrix with m rows of objects and n columns of measurements. The intensity of the normalized data object is xn. This normalization gives equal weight to each object. For two-way objects, the object is unfolded first into a vector. It is important that normalization is always implemented after baseline correction.

Baseline Correction. Previous work had used mass spectra collected at the end of a gas chromatographic run to model the baseline.[22] However, for the basil study small peaks occurred at the end of the gas chromatogram that prevented the use of spectra from this region being useful for modeling baseline variation. Therefore, GC/MS data objects of solvent blanks were used for baseline correction to construct an orthogonal basis set from the mass spectra.

For each blank GC/MS run, a basis set is created using singular value decomposition (SVD) from the entire collection of mass spectra. The number of components selected for the basis is five, because a conservatively small basis set will prevent overfitting of the data and negative peaks after subtracting the background from the sample runs. When a mass spectrum is projected onto the subspace defined by the basis, its projection is used to reconstruct a best fitting background. By subtracting this reconstructed background from a sample mass spectrum, the baseline variances are attenuated and the signal peaks are ideally unaffected.

In this work, a basis of background mass spectra was collected for each blank run that was collected for each block of sample data. The basis that best fitted a sample object was used to correct the baseline for that sample object. The best fitting basis gave the maximum sum of squared projections for a given sample object.

Retention Time Alignment. Peak drift among the chromatograms was detected. Because each block of data sets was collected on a different day, retention time drift was expected. A retention time alignment algorithm was used to align the mass spectra with respect to retention time so that the correlation of a two-way object with the two-way average object was maximized. A third-order polynomial mapping of retention times was fit and a cubic spline was used for interpolation of intensities to accomplish the alignment. The retention time alignment was applied to each object and the two-way average of the training set. The prediction set was then aligned to the two-way average of the training set.

The FuRES Classifier. FuRES builds a classification tree composed of minimal neutral networks.[18] The algorithm initiates by projecting data from a multidimensional space onto a normalized weight vector to yield scalar scores. These scores are used to calculate the fuzzy entropy of classification. FuRES generates classification rules by using the MATLAB function fminunc to find the weight vector with the lowest fuzzy entropy.[18] This method continues to partition the data by fuzzy membership functions. The fuzzy logistic values are the consequents of each rule. The multivariate rules comprise the branches of the classification tree. The divide and conquer algorithm continues until all the data of each leaf node consist of a single class,[18] and the final classification tree allows the visualization of the inductive structure of the rules.

FuRES[18] has been successfully applied to classify a variety of chemical data from different sources, including matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS),[23,24] high-performance liquid chromatography/mass spectrometry (HPLC/MS),[22] and gas chromatography/differential mobility spectrometry (GC/DMS).[25] The fuzzy aspect of FuRES accommodates both overlapped and underlapped clusters of data.[18] Another key advantage of FuRES over almost all other multivariate classifiers, such as support vector machines, neural networks, PLS-DA, etc., is that there are no adjustable parameters to be optimized.[16] Lastly, the features of the FuRES weight vector characterize general features between groups of classes at the root of the tree and precise differences between two classes from branches close to the leaves of the tree. The features of the weight vector are not susceptible to leverage effects because of the nonlinear fuzzy logistic function. PLS-DA which is a linear classifier may form ill-conditioned models by trying to fit its estimates to the binary values in the dependent block or target matrix used to define class membership values. Thus, FuRES features of the weight vector are more amenable to interpretation than regression coefficients from linear classifiers.

The FOAM Classifier. The FOAM method is an enhanced optimal associative memory (OAM) that encodes multivariate data as a two-way binary image instead of a one-way vector. A fuzzy function is applied to the image, and SVD is used to construct a basis from the fuzzified image data.[19] The FOAM was demonstrated for the background correction of single-scan near-infrared (NIR) spectra.[19] The FOAM was successfully applied to build classification models of chemical data for the first time in this paper.

Characteristically, the FOAM method requires encoding and decoding steps. When a data set is presented to the memory, it is binary-encoded using uniformly sized grids. This model derives from the processing that occurs when a spectrum or chromatogram is observed by an analyst and the optical cells of the analyst's retina are activated. Grids containing signals are set to unity and otherwise to zero. The fuzzy membership function is applied to this grid to render a fuzzy representation. An orthogonal basis is constructed that spans the spaces defined by the fuzzy images of the data. Lastly, the object is decoded by reversing the grid procedure. Classification can be achieved by building a basis for each class and assigning an object to the class with the minimum reconstruction error. The evolution of the FOAM algorithm can be found in previous work.[18,19,26] One FOAM is calculated for each class of data. After reversing the grid process to reconstruct the data, a Q statistic can be calculated to determine the best fitting model and hence the class.

Because a two-way GC/MS object may contain over a million data points and applying a grid function with 100 intensity elements would generate an object with more than 100 million data points, the principal component transform (PCT) was used to compress the data prior to FOAM modeling for the two-way data. Alternatively, mass spectra and chromatograms were modeled directly because of the relative fast speed of the calculation and the practical number of data points for the binary images.

The SIMCA Classifier. The SIMCA method is a classical quadratic discriminant analysis, which focuses on the similarity within a class.[27,28] It models each class separately. In this study, objects were classified using their orthogonal distances to the class model (or Q) and using the fit of their projections T^2 onto single component bases.[27,28] These two statistics were first normalized with respect to their boundary criteria. The model with the minimum root-mean-square of the sum of these two squared statistics was used to assign the class to the prediction object. All SIMCA models were first-order (i.e., used a single component). This design was used to contrast with FOAM, which was full order, in that all the available components were used for the classification models. The strength of SIMCA classification is that it models within-class covariations.

The PLS-DA Classifier. PLS-DA is another classification technique, which is usually used as a reference method for evaluating other algorithms. PLS is a regression method that finds the relationship between a predictor matrix **X** and a response matrix **Y**.[29] In a PLS model, the number of latent variables is selected, which yields the best prediction rates with respect to minimizing the root-mean-square prediction error.[22] PLS-DA is a particular case of PLS, in which **Y** is a set of binary variables describing the categories of categorical variables on **X**.[29] The number of response variables is equal to the number of categories. PLS-DA may outperform SIMCA with respect to classification rate, provided that within-class variability is low.[29]

In this work, the number of latent variables used by PLS was determined by an internal bootstrap that used 10 bootstraps and 2 Latin partitions of the training set. The number of latent variables that yielded the lowest average prediction error was selected for the model which was then used to build a classifier from the entire training set. This approach made the PLS algorithm parameter free and is useful for extensive validation with the bootstrap Latin partition method.

■ EXPERIMENTAL SECTION

Reagents and Supplies. Spectrophotometric grade pentane was purchased from the Baxter Healthcare Corp. (Muskegon, MI). HPLC grade acetone was purchased from the Spectrum Chemical Mfg. Corp. (New Brunswick, NJ). NORM-JECT plastic syringes (1 mL) were purchased from the Restek Corp. (Bellefonte, PA). Syringe filters with nylon membrane (13 mm × 0.22 μm) were provided by the Grand Stable Analysis Technics Co. Ltd. (Shanghai, China). Autosampler vials (2 mL) and polyspring inserts were purchased from the National Scientific Company (Rockwood, TN). National screw thread caps with bonded PTFE septa were purchased from the Thermo Fisher Scientific Inc. (Waltham, MA). Borosilicate glass vials (4 mL) with phenolic screw caps were purchased from the VWR LabShop (West Chester, PA).

Instruments. All the data were collected on a Shimadzu GCMS-QP2010SE gas chromatograph/mass spectrometer equipped with an AOC-20i autoinjector and an AOC-20S autosampler (Shimadzu Scientific Instruments, Columbia, MD). The GC/MS system was controlled by the GCMSsolution software version 2.70 (Shimadzu Scientific Instruments Inc., Columbia, MD).

The MATLAB R2012b (MathWorks Inc., Natick, MA) was used to process data. All the calculations were performed on an Intel Core i7 2.93 GHz personal computer with 12 GB RAM running a Microsoft Windows XP Professional x64 operation system (Microsoft Corp., Redmond, WA).

Materials and Sample Preparation. Five United States Department of Agriculture (USDA) certified organic and five conventional basil leaf samples were gifts from the Frontier Natural Product Co-op (Norway, IA, U.S.A.). All the basils were of the same variety and received as powders. According to the provider, each basil leaf sample was dried and then ground to 20 mesh particle size, following standard drying and grinding procedures in spices industry. Basil leaf powders were stored in a BD Falcon conical tube at 25 °C until analysis.

For sample preparation, 0.1 g of basil leaf powder was mixed with 2 mL of a pentane/acetone (90:10, v:v) solution at ambient temperature and vortexed for 1 min. The mixture was centrifuged at 10 000 rpm for 5 min. The supernatant was collected and filtered through a syringe filter with nylon membrane prior to GC/MS analysis. Eight replicates of each of the 10 basil leaf samples were prepared individually on the same day and stored in borosilicate glass vials at 4 °C until analysis.

Data Collection. The GC separation was accomplished on a 30 m × 0.25 mm × 0.25 μm 5% diphenyl/95% dimethyl polysiloxane cross-linked capillary column (SHRXI-5MS, Shimadzu Scientific Instruments Inc., Columbia, MD). The injector temperature was 260 °C, and the injection volume was 1 μL with a split ratio of 1:10. The temperature program was as follows: 75 °C, hold for 4 min; ramp 10 °C/min; and 280 °C, hold for 10 min. The interface temperature was 260 °C, and the ion source temperature was 200 °C. The carrier gas helium (99.99% purity) was maintained at a flow rate of 1.5 mL/min throughout the experiment.

A random block design was applied to the data collection process. The entire collection of 80 basil extracts (i.e., eight replicates for each of the 10 basil samples) was separated into eight blocks in a way that every block contains 10 different basil extracts (five organic samples and five conventional samples). Each block was analyzed on a separate day, and samples in the block were analyzed in a random order. A pentane blank data object was collected between two runs to reduce the cross contamination.

Another three replicates of each of the 10 basils samples were individually prepared as unknowns 2.5 months after the initial experiment. The data collection process was the same as the initial experiment. A random block design was also applied.

Data Formats. The two-way GC/MS data sets were acquired as computable document format (CDF) files. With an in-house algorithm, CDF files were read into MATLAB. The number of data points in a GC/MS object is the number of retention time measurements multiplied by the number of mass-to-charge ratio (m/z) measurements.

The data sets were binned by retention time from 4.1 to 34.5 min with a 0.01 min increment and binned by mass-to-charge ratio from 50 to 450 Th with a 1 Th increment. Thereafter, for each array, the 3041 rows corresponded to retention times and the 401 columns corresponded to the mass-to-charge ratios. Each two-way object comprised 1 219 441 data points.

Eighty two-way GC/MS data sets were collected. By inspecting their total ion chromatograms (TIC), five apparent outliers were observed. Peaks in the TIC of the outliers had poor signal-to-noise ratios (S/N) and low intensities, indicating weak responses of the detector probably due to low extraction efficiencies. Because basil extracts were prepared individually, the samples that appeared to have low extraction efficiencies were discarded from further analyses. After the outliers were eliminated, the total number of data objects was reduced from 80 to 75.

Data Processing. After data pretreatments (i.e., normalization, baseline correction, and retention time alignment) as explained in the Theory section, the FuRES, FOAM, SIMCA, and PLS-DA classifiers were constructed using identical data sets. The training set and the prediction set were selected using three Latin partitions that were bootstrapped for 100 times. The number of Latin partitions was set to three because the total number of data sets was not large enough to support the modeling methods that would use only half of the objects to construct the model for each class; in addition, the elimination of five outliers that occurred from poor extractions made the number of data sets in each class unequal. The prediction results across the three Latin partitions were pooled and then averaged across the 100 bootstraps. The bootstrap procedure partitioned the data by sample and not by replicate. In this way, no set of replicate measurements were in both the prediction set and the training set at the same time, i.e., the bootstrapped Latin partition was an unbiased evaluation of pattern recognition models.

■ RESULTS AND DISCUSSION

GC/MS Analysis. The first purpose of this work was to qualitatively discriminate organically grown basil samples from

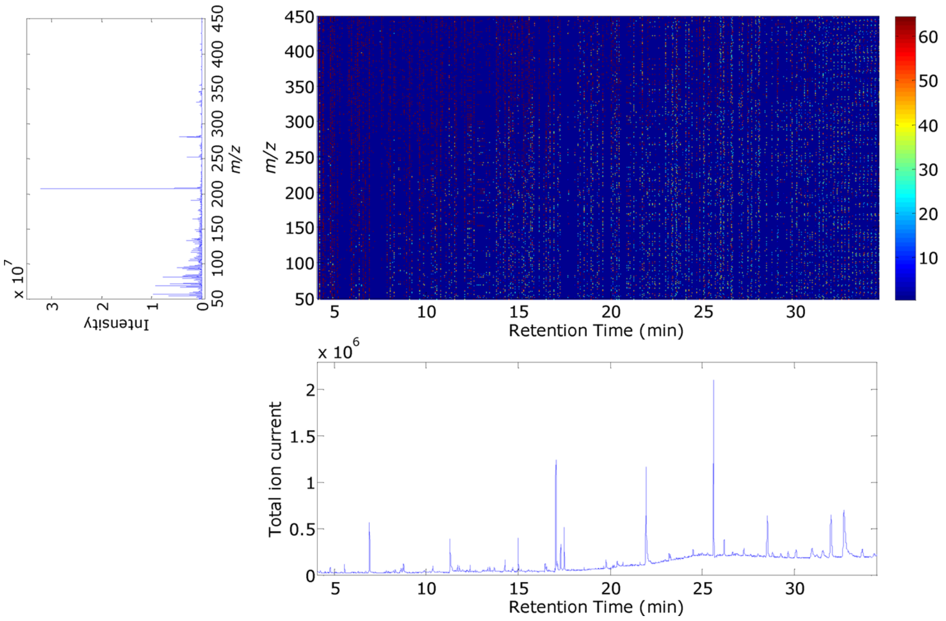


Figure 1. Image of the raw two-way GC/MS data set of a conventional basil sample. The integrated total ion chromatogram and the integrated mass spectrum are provided.

conventionally grown basils samples by the GC/MS analysis. The extraction solvent was optimized to obtain a wide variety of basil components, and the GC/MS conditions were optimized to shorten the GC run time while the separation of components served well for the classification.

Figure 1 provides an image of the raw two-way GC/MS data object of a conventional basil extract, in which pixels at different positions represent different ions. The background is colored dark blue while the larger peaks are red. Integrated peaks are also displayed in the total ion chromatogram (in the GC way) and the total mass spectrum (in the MS way).

After searching the significant peaks against the mass spectral library with retention indices provided by the National Institute of Standards and Technology (NIST) database (as a Shimadzu GCMSsolution software version 2.70 feature), major compounds of basil extracts were putatively identified. For each compound listed in Table 1, the similarity index is greater than 85, although standards must be run to confirm the identities.

Because basil samples in each class were identical with respect to supplier, variety, and geographical location, artificial variation among samples in the same class could be small. It is reasonable to expect that only the cultivation condition, i.e., organic or conventional environment, primarily contributed to the difference among samples. As for major components, organic and conventional basils were similar to a large extent. These putatively identified components are provided in Table 1.

Pretreatment Effects on the Data Sets. The second purpose of this work was to evaluate the effect of data pretreatments. Baseline correction and retention time alignment were applied to the data sets prior to data processing.

Baseline Correction. Baseline drift could adversely affect pattern recognition. To correct the GC baselines of basil extracts, 10 GC/MS data sets of solvent blank were collected. A single orthonormal basis set was constructed for each blank. As described in the Theory section, the basis set that best fit the sample object was chosen for baseline correction. This optimal basis set was used to reconstruct the mass spectra that modeled the baseline. Each reconstructed mass spectrum was subtracted from each corresponding sample mass spectrum, to provide baseline-corrected data objects for classification.

A comparison between the TICs before and after baseline correction was performed. The TICs from 30 to 34 min are plotted as an example in Figure 2, in which the baselines are attenuated significantly after correction. Thereby, baseline

Table 1. Putatively Identified Main Components of Basils

compd name[a]	formula	RT (min)
eucalyptol	$C_{10}H_{18}O$	5.6
3,7-dimethyl-1,6-octadien-3-ol	$C_{10}H_{18}O$	6.9
tetradecane	$C_{14}H_{30}$	10.4
3-alyl-6-methoxyphenol	$C_{10}H_{12}O_2$	11.3
1-tridecene	$C_{13}H_{26}$	11.7
1-pentadecene	$C_{15}H_{30}$	11.7
1-chloro-octadecane	$C_{18}H_{37}Cl$	11.8
2,6-dimethyl-6-(4-methyl-3-pentenyl)-bicyclo[3.1.1]hept-2-ene	$C_{15}H_{24}$	12.4
(1α,4aα,8aα)-1,2.3,4,4a,5,6,8a-octahydro-7-methyl-4-methylene-1-(1-methylethyl)-naphthalene	$C_{15}H_{24}$	13.4
(−)-spathulenol	$C_{15}H_{24}O$	14.3
cubenol	$C_{15}H_{26}O$	14.7
T-cadinol	$C_{15}H_{26}O$	15.0
2-isopropyl-5-methyl-9-methylene	$C_{15}H_{26}O$	15.0
3,7,11,15-tetramethyl-2-bicyclo[4.4.0]dec-1-ene hexadecen-1-ol	$C_{20}H_{40}O$	17.1, 17.3, 17.5
(Z)-9-octadecenamide	$C_{18}H_{35}NO$	21.9
7,7′,8,8′,11,11′,12,12′,15,15′-decahydro-Ψ,Ψ-carotene	$C_{40}H_{66}$	25.6
squalene	$C_{30}H_{50}$	25.7
2,2,4-trimethyl-3-(3,8,12,16-tetramethyl-heptadeca-3,7,11,15-tetraenyl)-cyclohexanol	$C_{30}H_{52}O$	26.8
nonacosane	$C_{29}H_{60}$	28.5, 32.0
vitamin E acetate	$C_{31}H_{52}O_3$	29.3
γ-sitosterol	$C_{29}H_{50}O$	32.7

[a] RT denotes retention time.

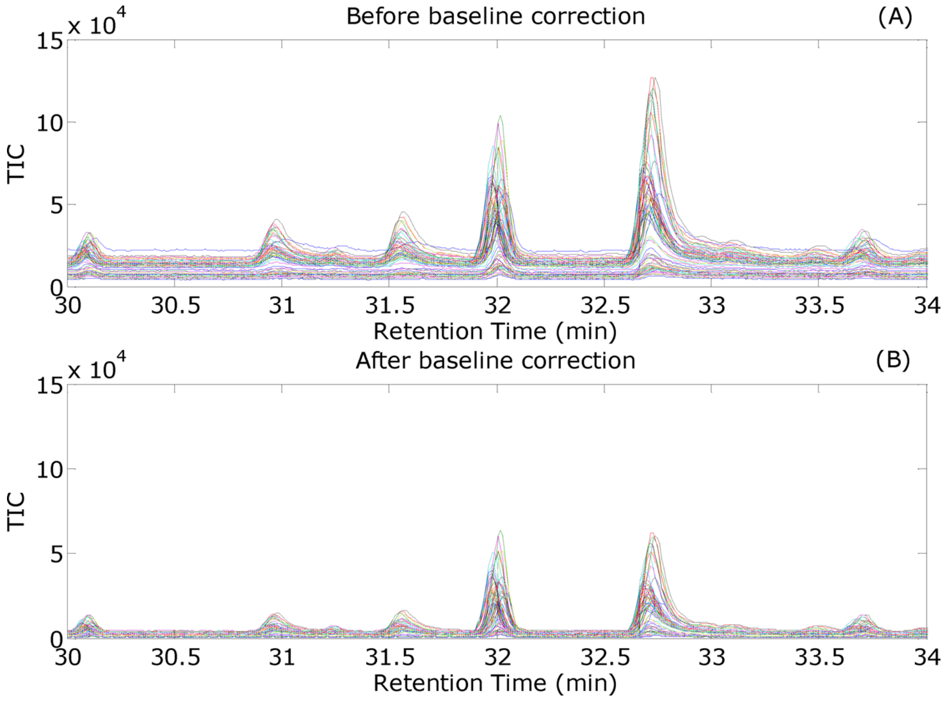


Figure 2. Total ion chromatograms from 30 to 34 min before (A) and after (B) baseline correction, as a demonstration of baseline correction effect.

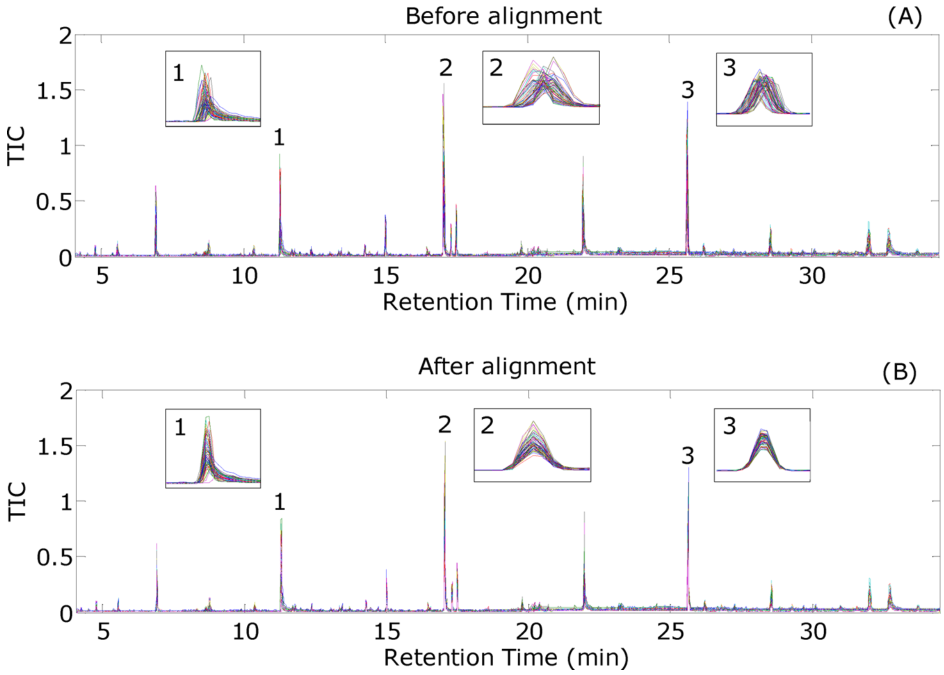


Figure 3. Reconstructed total ion chromatograms with baseline correction before (A) and after (B) retention time alignment. Peak clusters at (1) 11.3, (2) 17.3, and (3) 25.6 min are magnified.

correction improves the quality of the data by removing background features that may be common for all objects. The peaks in the TICs also appear to be reduced by almost a factor of 2 after baseline correction. This attenuation is caused by removal of background components in the mass spectra prior to the integration to yield the points of the TICs.

Retention Time Alignment. The TICs of entire data sets are plotted in Figure 3, in which three major peak clusters at 11.3, 17.3, and 25.6 min are magnified. In Figure 3A, retention time drift among different runs occurs in the entire GC program and becomes more apparent when the temperature is greater. The retention time alignment approach notably increases the similarity among chromatograms of the same class. As illustrated in Figure 3B, the distributions of peak clusters are significantly narrower than those peaks of Figure 3A.

The benefit of data pretreatments can also be visualized in the plot of the PCA scores. Figure 4A indicates that no well-defined separation of data sets can be obtained without pretreatment. After baseline correction, normalization, and retention time alignment, however, the organic basils are

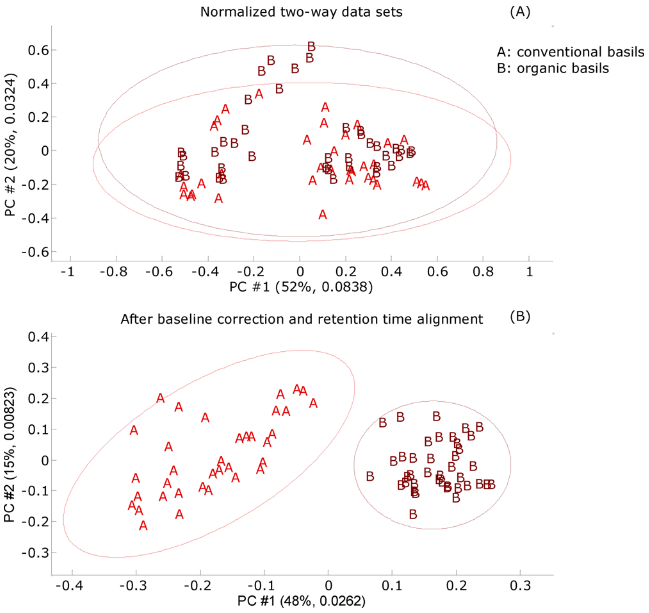


Figure 4. Principal component analysis score plots for two-way GC/MS data sets of conventional and organic basils before (A) and after (B) data pretreatments. The percent variance spanned by the principal components is given in parentheses with the absolute variance. A 95% confidence interval is drawn around the mean of each class.

resolved from the conventional basils, as illustrated in Figure 4B.

Typically, data pretreatment is performed on the entire data set. In practical applications, however, classifiers are established first and then applied to unknown samples, i.e., the classifier-training set and the prediction set are temporally separated. Therefore, in the present work retention time alignment was performed in this practical yet time-consuming procedure: data sets were partitioned randomly into the training set and the prediction set; then, the training set was aligned and the average was returned; afterward, data sets in the prediction set were each aligned to the mean of the training set. Satisfactory classification rates indicate that these established classifiers are robust for practical applications.

Comparison among Classifiers. The third purpose of this work was to evaluate the fuzzy classifiers, i.e., FuRES and FOAM, to the classification of different representations of the GC/MS data sets. Another two widely utilized crisp classifiers, i.e., SIMCA and PLS-DA, were chosen to make a comparison, considering that (1) FuRES and FOAM are fuzzy classifiers while SIMCA and PLS-DA are crisp classifiers and (2) FuRES and PLS-DA are harder classification methods with PLS-DA being the harder of the two, while SIMCA and FOAM are softer modeling methods. Hardness and softness refer to the bias-variance trade-off.[30]

For the classification of organic and conventional basils with SIMCA or FOAM, a model for each class will be established. When a blind unknown is present, it will be modeled and compared with the established model of each class. If the unknown model is rejected by either class, it will be assigned as class 0; if the unknown model is accepted by both classes, it will be assigned to the best fitting class.

The FuRES, FOAM, SIMCA, and PLS-DA classifiers were constructed and evaluated with 100 bootstraps of three Latin partitions. These classifiers were compared with respect to classification accuracy for three different representations of the data. As listed in Table 2, by using two-way data sets, average classification rates of 100 ± 0%, 94.4 ± 0.4%, 93.3 ± 0.4%, and 100 ± 0% are obtained with FuRES, FOAM, SIMCA, and PLS-DA, respectively. The averages are reported with their 95% confidence intervals.

Table 2. Classification Rates with 95% Confidence Intervals Obtained by Using Different Data Representations with Bootstrap Alignment of Training Data

	av classification rates[a]		
classifier	MS (%)	GC (%)	2W (%)
FuRES	100 ± 0	100 ± 0	100 ± 0
FOAM	95.9 ± 0.4	95.4 ± 0.3	94.4 ± 0.4
SIMCA	93.4 ± 0.4	93.9 ± 0.2	93.3 ± 0.4
PLS-DA	100 ± 0	100 ± 0	100 ± 0

[a] MS denotes the total mass spectrum representation, GC denotes the total ion chromatogram representation, and 2W denotes the two-way representation. Averages were calculated from 100 × 3 bootstrapped Latin partitions. For MS no alignment was used.

Moreover, when classifiers were constructed by using one-way representation of data sets, i.e., only the total ion chromatograms (i.e., total ion current with respect to retention time) or only the mass spectra (i.e., intensity with respect to mass-to-charge ratio), classification rates differed from those rates obtained with the two-way GC/MS representation. As listed in Table 2, for the GC representation of the data (i.e., total ion chromatograms) the average classification rates of FuRES, FOAM, SIMCA, and PLS-DA were 100 ± 0%, 95.4 ± 0.3%, 93.9 ± 0.2%, and 100 ± 0%, respectively, for 100 × 3 bootstrapped Latin partitions. For the MS representation of the data (i.e., mass spectra) the average classification rates of FuRES, FOAM, SIMCA, and PLS-DA were 100 ± 0%, 95.9 ± 0.4%, 93.4 ± 0.4%, are 100 ± 0%, respectively, for 100 × 3 bootstrapped Latin partitions (Table 2).

Thereby, the FuRES, FOAM, SIMCA, and PLS-DA classifiers were validated 100 times. These results indicate that (1) FuRES and PLS-DA both performed perfectly, however, FuRES has the key advantage of being a parameter-free classifier while the number of latent variables must be predetermined in the PLS-DA algorithm, and thus the implementation of FuRES is easier than PLS-DA, and (2) between the two modeling methods, FOAM was more robust than SIMCA. Note that the modeling methods were expected to perform at a 95% classification rate, because the parametric boundary was set to a 95% confidence interval. All three representations performed equally well with the FOAM working better for the MS and GC representations than for the two-way representation.

Classification of Blind Unknowns. Ten basil extracts were freshly prepared 2.5 months after initial two-way GC/MS data sets were collected. Three replicates were individually prepared for each sample. Two-way data sets were collected as blind unknowns by using the same GC/MS program as for the initial experiment. These 30 blind unknowns comprised a new validation set and were subjected to the same data processing and pattern recognition processes as the initial 75 basil samples.

Classifiers constructed by using different representations of initial 75 data sets were used on the 30 new data sets. As listed in Table 3, all these 30 blind unknowns are accurately classified by the FuRES and PLS-DA classifiers established by using two-way data sets, yielding a classification rate of 100%. As a

Table 3. Classifier Validation with Data Collected 2.5 Months Later with No Parametric Changes

	classification rates[a]					
classifier	MS (%)	MS.MC (%)	GC (%)	GC.MC (%)	2W (%)	2W.MC (%)
FuRES	90	100	90	100	100	100
FOAM	0	17	7	80	20	93
SIMCA	0	13	7	100	47	97
PLS-DA	67	83	100	100	97	100

[a] MS denotes the total mass spectrum representation, GC denotes the total ion chromatogram representation, and 2W denotes the two-way representation. For MS no alignment was used. The validation set was aligned to the training set. MC refers to individually mean-centering the training set and the validation set.

comparison, classification rates of 93% and 97% are obtained with FOAM and SIMCA, respectively. By using the MS representation, the classification rates with FuRES, FOAM, SIMCA, and PLS-DA are 100%, 17%, 13%, and 83%, respectively. By using the GC representation, the classification rates with FuRES, FOAM, SIMCA, and PLS-DA are 100%, 80%, 100%, and 100%, respectively.

After the initial experiment which was performed 2.5 months ago, small deviations of the instrument status or manual operation may caused inconsistencies between the new and old data sets, although experimental conditions were identical. Additional data pretreatment on new data sets was then necessary. Because FOAM is sensitive to differences in the entire profile and SIMCA is sensitive to changes in the latent structure, classification rates with FOAM were lower than with SIMCA for the long-term evaluation of the established classifiers.

The model-building set (i.e., collected earlier) and the prediction set (i.e., collected 2.5 months later) were each individual mean-centered. As illustrated in Figure 5, no well-defined separation of organic and conventional basils is obtained before mean-centering the data sets. In addition, for each class (organic or conventional) of basils, the scores of the objects in the model-building set are not close to the scores of the objects from the prediction set. After individually mean-centering the training set and the prediction set, however, the organic basils are well-resolved from the conventional basils, and for each class, the score of objects in the prediction set correspond to the scores of the objects in the training set.

Finally, it is worthy to mention that in pattern recognition an individual observation can be analyzed into a set of quantifiable properties or features. Some features are critical to discriminate different classes while others contain noise or irrelevant information. Choosing distinctive and independent features to represent the original two-way data set reduces data dimension and increases the efficiency of classification. This process can be referred to as feature selection. In a follow-up paper, feature selection using FuRES will be used to find characteristic components in the organic and conventional basil extracts that will then be identified.[31]

■ CONCLUSIONS

The GC/MS analysis of basil extracts demonstrated that organically and conventionally grown basils could be distinguished by their chemical profiles. Major components of basils were detected and putatively identified by library search and retention time index. Seventy-five basil samples were

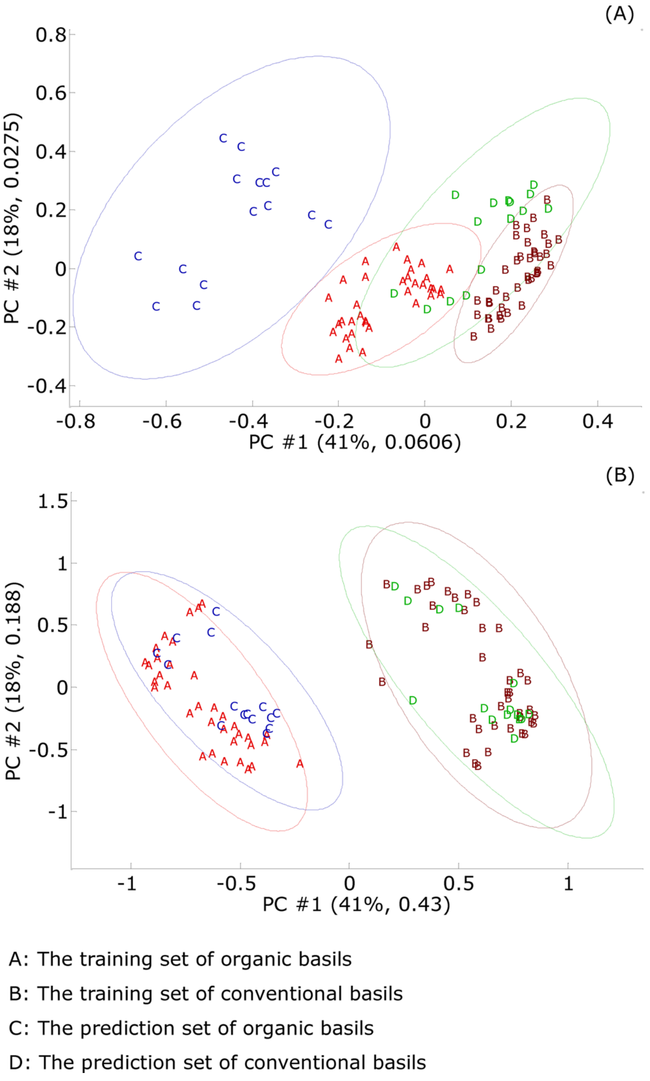


Figure 5. Principal component analysis score plots for two-way GC/MS data sets of conventional and organic basils before (A) and after (B) mean-centering the data sets. The percent variance spanned by the principal components is given in parentheses with the absolute variance. A 95% confidence interval is drawn around the mean of each class. The prediction set was collected 2.5 months after the training set.

classified by four classifiers, i.e., FuRES, FOAM, SIMCA, and PLS-DA. After baseline correction and retention time alignment, the average classification rates for 100 × 3 bootstrapped Latin partitions were 100 ± 0% with FuRES, 94.4 ± 0.4% with FOAM, 93.3 ± 0.4% with SIMCA, and 100 ± 0% with PLS-DA. These established classifiers were used to classify 30 unknown samples collected 2.5 months after the initial experiment, and the classification rates obtained with FuRES, FOAM, SIMCA, and PLS-DA were 100%, 93%, 97%, and 100%, respectively.

The GC/MS analysis coupled with pattern recognition can be used for authentication of plant materials. The results show that robust classifiers obtained from random block designs may build models that are stable for months. All four classifiers performed well, with the modeling methods as expected lagging in performance from the harder classifiers. For the first time a FOAM was applied to the classification of chemical data or profiling. The data representation, MS, GC, or two-way all performed satisfactorily well, except a significant improvement was detected for the FOAM prediction of the two-way data obtained from the 2.5 months study. In addition, a novel baseline correction algorithm was presented. Future papers will present optimization of the FOAM modeling method and further development of baseline correction algorithms. In addition, other botanicals will be examined on a larger scale. Feature selection and validation are presented in a follow-up publication.

AUTHOR INFORMATION

Corresponding Author

*Phone: +01 740-994-0265. Fax: +01 740-593-0148. E-mail: peter.harrington@ohio.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the Agricultural Research Service of the U.S. Department of Agriculture and an Interagency Agreement with the Office of Dietary Supplements of the National Institutes of Health (Grant #Y01-OD001298-01). The Center for Intelligent Chemical Instrumentation and Department of Chemistry and Biochemistry at Ohio University are acknowledged for the financial support. Dr. Xiaobo Sun is thanked for her suggestions.

REFERENCES

(1) Soule, J. A. *Father Kino's Herbs: Growing & Using Them Today*; Tierra del Sol Institute: Tuscon, AZ, 2011.
(2) Chiang, L.-C.; Ng, L.-T.; Cheng, P.-W.; Chiang, W.; Lin, C.-C. *Clin. Exp. Pharmacol. Physiol.* **2005**, *32*, 811−816.
(3) Bozin, B.; Mimica-Dukic, N.; Simin, N.; Anackov, G. *J. Agric. Food Chem.* **2006**, *54*, 1822−1828.
(4) Almeida, I. d.; Alviano, D. S.; Vieira, D. P.; Alves, P. B.; Blank, A. F.; Lopes, A. H. C. S.; Alviano, C. S.; Rosa, M. d. S. S. *Parasitol. Res.* **2007**, *101*, 443−452.
(5) Nunoz-Acevedo, A.; Stashenko, E. E.; Kouznetsov, V. V.; Martinez, J. R. *J. Essent. Oil-Bear. Plants* **2011**, *14*, 387−395.
(6) Dev, N.; Das, A. K.; Hossain, M. A.; Rahman, S. M. M. *J. Sci. Res.* **2011**, *3*, 197−206.
(7) Díaz-Maroto, M. C.; Pérez-Coello, M. S.; Cabezudo, M. D. *Chromatographia* **2002**, *55*, 723−728.
(8) Charles, D. J.; Simon, J. E. *J. Am. Soc. Hortic. Sci.* **1990**, *115*, 458−462.
(9) Viña, A.; Murillo, E. *J. Braz. Chem. Soc.* **2003**, *14*, 744−749.
(10) Allen, G.; Albala, K. *The Business of Food: Encyclopedia of the Food and Drink Industries*; Greenwood Press: Westport, CT, 2007.
(11) Jirovetz, L.; Buchbauer, G.; Shafi, M. P.; Kaniampady, M. M. *Eur. Food Res. Technol.* **2003**, *217*, 120−124.
(12) Klimankova, E.; Holadova, K.; Hajslova, J.; Cajka, T.; Poustka, J.; Koudela, M. *Food Chem.* **2008**, *107*, 464−472.
(13) Nerín, C.; Salafranca, J.; Aznar, M.; Batlle, R. *Anal. Bioanal. Chem.* **2009**, *393*, 809−833.
(14) Hasegawa, Y.; Tajima, K.; Toi, N.; Sugimura, Y. *Flavour Fragrance J.* **1997**, *12*, 195−200.
(15) Xu, Z.; Sun, X.; Harrington, P. d. B. *Anal. Chem.* **2011**, *83*, 7464−7471.
(16) Carr, P. W. *Anal. Chem.* **1980**, *52*, 1746−1750.
(17) Annesley, T. M. *Clin. Chem.* **2003**, *49*, 1041−1044.
(18) Harrington, P. B. *J. Chemom.* **1991**, *5*, 467−486.
(19) Wabuyele, B. W.; Harrington, P. d. B. *Appl. Spectrosc.* **1996**, *50*, 35−42.
(20) Harrington, P. d B. *TrAC, Trends Anal. Chem.* **2006**, *25*, 1112−1124.
(21) Lu, W.; Rankin, J. G.; Bondra, A.; Trader, C.; Heeren, A.; Harrington, P. d. B. *Forensic Sci. Int.* **2012**, *220*, 210−218.
(22) Sun, X.; Chen, P.; Cook, S. L.; Jackson, G. P.; Harnly, J. M.; Harrington, P. d. B. *Anal. Chem.* **2012**, *84*, 3628−3634.
(23) Harrington, P. d. B.; Vieira, N. E.; Chen, P.; Espinoza, J.; Nien, J. K.; Romero, R.; Yergey, A. L. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 283−293.
(24) Laurent, C.; Harrington, P. B.; Levinson, D. F.; Levitt, P.; Markey, S. P. *Anal. Chim. Acta* **2007**, *599*, 219−231.
(25) Rearden, P.; Harrington, P. B.; Karnes, J. J.; Bunker, C. E. *Anal. Chem.* **2007**, *79*, 1485−1491.
(26) Wabuyele, B. W.; Harrington, P. d. B. *Anal. Chem.* **1994**, *66*, 2047−2051.
(27) Frank, I. E. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 247−256.
(28) Tominaga, Y. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 105−115.
(29) Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholoson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20*, 341−351.
(30) Geman, S.; Bienenstock, E.; Doursat, R. *Neural Comput.* **1992**, *4*, 1−58.
(31) Wang, Z.; Harrington, P. d. B. *Anal. Chem.* **2013**, submitted for publication.