Talanta 94 (2012) 320–327

Contents lists available at SciVerse ScienceDirect

Talanta

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

Predicting feedstock and percent composition for blends of biodiesel with conventional diesel using chemometrics and gas chromatography–mass spectrometry

Stephen P. Schale, Trang M. Le, Karisa M. Pierce*

Seattle Pacific University, 3307 Third Avenue West, Seattle, WA 98119-1950, United States

ARTICLE INFO

Article history:
Received 22 February 2012
Received in revised form 13 March 2012
Accepted 20 March 2012
Available online 28 March 2012

Keywords:
Chromatography
Biodiesel
Blend
Chemometrics

ABSTRACT

The two main goals of the analytical method described herein were to (1) use principal component analysis (PCA), hierarchical clustering (HCA) and K-nearest neighbors (KNN) to determine the feedstock source of blends of biodiesel and conventional diesel (feedstocks were two sources of soy, two strains of jatropha, and a local feedstock) and (2) use a partial least squares (PLS) model built specifically for each feedstock to determine the percent composition of the blend. The chemometric models were built using training sets composed of total ion current chromatograms from gas chromatography–quadrupole mass spectrometry (GC–qMS) using a polar column. The models were used to semi-automatically determine feedstock and blend percent composition of independent test set samples. The PLS predictions for jatropha blends had RMSEC = 0.6, RMSECV = 1.2, and RMSEP = 1.4. The PLS predictions for soy blends had RMSEC = 0.5, RMSECV = 0.8, and RMSEP = 1.2. The average relative error in predicted test set sample compositions was 5% for jatropha blends and 4% for soy blends.

1. Introduction

Biofuels are an important alternative to fossil fuels. Biodiesel, in particular, is an important substitute to conventional diesel since blends of biodiesel and conventional diesel can be burned in conventional diesel engines without modification [1]. As alternative fuel sources are developed, analytical methods that provide quality assurance information can be useful.

Biodiesels produced from different feedstocks are composed of unique and characteristic fatty acid methyl esters (FAME) [2]. The biodiesel FAMEs have significantly different properties than the hydrocarbons in conventional diesel, so it is difficult to achieve full chromatographic resolution of all components of a blend of biodiesel and conventional diesel. Without full chromatographic resolution, traditional methods of chromatographic data analysis such as the internal standard method involving integrated peak areas are not convenient. However, even without ideal chromatographic resolution, we can exploit the reproducible gas chromatography–quadrupole mass spectrometry (GC–qMS) chromatographic fingerprints of the samples by using the chemometric pattern recognition techniques principal component analysis (PCA), hierarchical cluster analysis dendrograms (HCA), and K-nearest neighbors (KNN). We will also use the multivariate calibration techniques partial least squares analysis (PLS) and n-way partial least squares analysis (NPLS) to model the total ion current (GC-TIC) and GC–qMS fingerprints. PCA, HCA, and KNN are well known unsupervised pattern recognition techniques that are described in many excellent chemometric textbooks [3–5]. In short, these techniques cluster similar training set chromatograms together in transformed variable space, thus test set chromatograms are semi-automatically determined to be from the same feedstock as the nearest training set cluster. PLS and NPLS are well known multivariate calibration techniques that are also described in chemometric textbooks [3–5]. PLS and NPLS training set models are used to predict the percent composition of test set samples. Biodiesels are commercially available as 5% and 20% by volume of biodiesel blended in conventional diesel (also known as B5 and B20) so we prepared training set and test set blends ranging from 0% through 30%.

Variations in retention time and injection volume can obscure important chemical variations so our analytical method includes procedures to reduce the chemically irrelevant sources of variation [6]. The procedures include baseline correction, sum-normalization, alignment, feature selection, and a temporary scaling step prior to PCA, HCA, and KNN modeling. Prior to PLS and NPLS modeling, preprocessing procedures were baseline correction, sum-normalization, and alignment.

We specifically chose to study biodiesel blends because the ability to predict retail biodiesel blend percent composition is important to regulatory agencies, fuel compliance officers, and

* Corresponding author. Tel.: +1 206 281 2102; fax: +1 206 281 2882.
E-mail address: piercek@spu.edu (K.M. Pierce).

0039-9140/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.03.050

distributors of transport fuel who may be interested in monitoring authenticity, quality, contamination, adulteration, and accuracy of reported blend percent compositions. Instrumental methods for monitoring biodiesel blends have been reported using Fourier-transform infrared spectroscopy [7–9], near-infrared spectroscopy [10,11], mid-infrared spectroscopy [12], nuclear magnetic resonance spectroscopy [11,13–16], radio carbon analysis [17], electrospray ionization mass spectrometry [18], liquid chromatography [19,20], and GC × GC with flame ionization detection [2,21]. Other reports quantified biodiesel blends through GC-FID [22] and GC–MS [23,24]. An ambient mass spectrometry method has been reported to identify plant sources of biodiesel [25]. A PCA method was used to analyze data collected by an electronic nose, with some degree of quantification achieved with an artificial neural network [26]. We previously reported a GC × GC–MS method to determine percent composition of biodiesel and conventional diesel blends using NPLS, and the current work reported herein is an extension of that project [27]. The previous project only had one feedstock of biodiesels, but the project herein involves multiple biodiesel feedstocks. The two main goals of the analytical method described herein are to (1) use chemometric pattern recognition to determine the feedstock source of biodiesel in a blend of biodiesel and conventional diesel (i.e. whether the biodiesel feedstock is from soy, jatropha, or a commercial feedstock), and (2) use a chemometric calibration model built for the particular feedstock to determine the percent composition of the biodiesel in the blend.

2. Materials and methods

2.1. Materials

Commercial biodiesel and commercial conventional diesels were obtained from pumps at commercial stations in the Seattle, WA area. The feedstock for the commercial biodiesel is unknown to the authors, but according to the company's website, it is supposed to be a local feedstock such as canola, switchgrass, soy, mustard, corn, or camelina. The fatty acid profile of the commercial biodiesel does not match reports of soy and jatropha fatty acid profiles [28,29] so herein we assume the commercial biodiesel is not soy nor jatropha; additionally, later in this text (Fig. 3) it is shown that unsupervised pattern recognition clusters the commercial feedstock separate from the soy and jatropha clusters, supporting our assumption that our samples are composed of three different biodiesel feedstocks even though one feedstock is unknown. The pure jatropha and soy biodiesels were provided by a major energy company. The provided samples were from two strains of jatropha and two different sources of soy. The ten samples are designated Jatropha A (Jat A), Jatropha B (Jat B), Soy A, Soy B, Commercial Retailer (Com), Crown Hill retailer conventional diesel (Conv 1), Nickerson retailer conventional diesel (Conv 2), Shoreline retailer conventional diesel (Conv 3), West Seattle retailer conventional diesel (Conv 4), and Interbay retailer conventional diesel (Conv 5). The abbreviations shown in parentheses are used in Tables 1–4. Samples were prepared neat via mixing to create blends of biodiesel and conventional diesel wherein biodiesel concentrations were 0%; 1%; 5%; 10%; 15%; 20%; 25% and 30% (v/v); yielding 66 different mixtures. Tables 1–3 describe which of these 66 samples were used in the training sets and which were used in the test sets.

2.2. Chromatographic method

The chromatograms were obtained by an Agilent Technologies 6890 Network GC System with a 5973 Mass Selective Detector and 7683 Series autoinjector. Samples were separated on a nonpolar column and on a polar wax column. The polar column was an Agilent 19091N-113 HP-INNOWax (polyethylene glycol) with dimensions 30 m × 320 μm × 0.25 μm. For the polar column separations, the injector was held at 250 °C, samples were injected neat with a 120:1 split ratio, and the He flow was constantly 1 mL/min. The oven was heated from 60 °C at 2 °C/min to 228 °C. The transfer line heater was held at 230 °C. The detector scanned ions of m/z 50–300 at 5.46 spectra/s with 150 ion count threshold and no solvent delay. The nonpolar column was an Agilent

Table 1
Samples prepared for PCA training set (left) and PCA test set (right). Percent refers to percent by volume of biodiesel in blends of biodiesel and conventional diesel. See Section 2.1 for definition of abbreviations.

PCA training set			PCA test set		
Biodiesel	Conventional	Percent	Biodiesel	Conventional	Percent
Jat A	Conv 1	20	Soy A	Conv 1	10
Jat A	Conv 2	20	Jat A	Conv 2	10
Jat A	Conv 3	20	Jat B	Conv 3	10
Jat A	Conv 4	20	Com	Conv 4	10
Jat B	Conv 1	20	Soy B	Conv 5	10
Jat B	Conv 2	20	Soy A	Conv 5	15
Jat B	Conv 3	20	Jat A	Conv 5	15
Jat B	Conv 4	20	Jat B	Conv 5	15
Com	Conv 1	20	Com	Conv 5	15
Com	Conv 2	20	Soy B	Conv 5	15
Com	Conv 3	20	Soy A	Conv 5	25
Com	Conv 4	20	Jat A	Conv 5	25
Com	Conv 5	20	Jat B	Conv 5	25
Soy A	Conv 1	20	Com	Conv 5	25
Soy A	Conv 2	20	Soy B	Conv 5	25
Soy A	Conv 3	20	Soy A	Conv 1	30
Soy A	Conv 4	20	Jat A	Conv 2	30
Soy B	Conv 1	20	Jat B	Conv 3	30
Soy B	Conv 2	20	Com	Conv 4	30
Soy B	Conv 3	20	Soy B	Conv 5	30
Soy B	Conv 4	20			
–	Conv 1	0			
–	Conv 2	0			
–	Conv 3	0			
–	Conv 4	0			
–	Conv 5	0			

Table 2
Blends prepared for the jatropha PLS training set and the soy PLS training set. Percent refers to percent by volume of biodiesel in blends of biodiesel and conventional diesel. See Section 2.1 for definition of abbreviations.

Jatropha PLS training set			Soy PLS training set		
Biodiesel	Conventional	Percent	Biodiesel	Conventional	Percent
Jat A	Conv 5	1	Soy A	Conv 5	1
Jat A	Conv 5	5	Soy A	Conv 5	5
Jat A	Conv 5	10	Soy A	Conv 5	10
Jat A	Conv 5	20	Soy A	Conv 5	20
Jat A	Conv 5	30	Soy A	Conv 5	30
Jat B	Conv 5	1	Soy B	Conv 2	1
Jat B	Conv 5	5	Soy B	Conv 2	5
Jat B	Conv 5	10	Soy B	Conv 2	15
Jat B	Conv 5	20	Soy B	Conv 2	20
Jat B	Conv 5	30	Soy B	Conv 2	30

Table 3
Blends prepared for jatropha PLS test set and soy PLS test set, and the blend percent composition predicted by PLS. Percent refers to percent by volume of biodiesel in blends of biodiesel and conventional diesel. See Section 2.1 for definition of abbreviations.

Biodiesel	Conventional	Percent	PLS prediction
Soy A	Conv 1	10	9.9
Soy B	Conv 5	10	10.1
Soy A	Conv 5	15	16.7
Soy B	Conv 5	15	15.5
Soy A	Conv 5	25	24.4
Soy B	Conv 5	25	22.9
Soy A	Conv 1	30	31.3
Soy B	Conv 5	30	28.7
Jat A	Conv 2	10	10.4
Jat B	Conv 3	10	10.8
Jat A	Conv 5	15	15.3
Jat B	Conv 5	15	15.5
Jat A	Conv 5	25	25.4
Jat B	Conv 5	25	25.1
Jat A	Conv 2	30	31.5
Jat B	Conv 3	30	33.5

09091S-433 HP-5MS (5% phenyl methyl siloxane) with dimensions 30 m × 250 μm × 0.25 μm. For the nonpolar column separations, the injector was held at 250 °C, samples were injected neat with a 120:1 split ratio, and the He flow was constantly 0.8 mL/min. The oven was heated from 60 °C at 3 °C/min to 288 °C. The transfer line heater was held at 300 °C. The detector scanned ions of *m*/*z* 50–300 at 5.46 spectra/s with 150 ion count threshold and 4.4 min solvent delay.

In general, the nonpolar column was able to resolve conventional diesel peaks, and the polar column that was able to resolve biodiesel peaks (FAME peaks). Fig. 1 shows the chromatograms of a sample of 20% soy biodiesel blended in Interbay diesel. For both columns, the conventional diesel components elute earlier than the FAMEs. In Fig. 1A, the conventional diesel peaks are relatively well resolved by the nonpolar column, while the FAME peaks at ~48 min are unresolved (the peak at 44 min is the earliest eluting FAME). Conversely, in Fig. 1B, the polar column does not resolve the conventional diesel peaks, while the FAME peaks at ~47–67 min are relatively well resolved. The ChemStation software with the NIST05 mass spectral matching library was used to identify the FAME peaks labeled in Fig. 1. The reported identifications agreed among three chromatograms and had quality match factors equal to or greater than 95 out of 100.

2.3. Data analysis method

The chromatograms were exported as .txt files including all scans and all *m*/*z* using Enhanced ChemStation E.02.00.493. The .txt files were imported into MATLAB 7.9.0 (Mathworks, Natick, MA) for analysis. The pixel level GC–qMS chromatograms were preprocessed by baseline correction, sum-normalization, and piecewise alignment using window size = 300 points, estimated peak size = 50 points and maximum shift = 100 points, target = ninth training set sample (arbitrarily chosen) [30]. The baseline corrected, normalized, and aligned GC–qMS chromatograms were then summed along the *m*/*z* dimension to yield GC-TIC chromatograms. The GC-TIC chromatograms were combined into a single 2D matrix, and this matrix was submitted to PCA, HCA, KNN, and PLS using PLS Toolbox 6.2.1 software (Eigenvector Research, Manson, WA). The baseline corrected, normalized, and aligned GC–qMS chromatograms were also combined into a three-way array, and this array was submitted to NPLS using PLS Toolbox 6.2.1. The three-way array of GC–qMS chromatograms was unfolded along the *m*/*z* dimension for submission to PCA. PC 1 and PC 2 were used for PCA modeling. Three latent variables were used for PLS and NPLS modeling. Datasets were mean centered prior to PCA, PLS, and NPLS. The *f*-ratio threshold was determined by maximizing the percent variance captured by PCA models built for reduced training sets as a function of a wide range of *f*-ratio threshold values. For example, the polar column GC-TIC training set, was reduced by retaining only features with *f*-ratio values above 130. This threshold corresponded to a PCA model that captured a high percentage of variation (98%) and it corresponded to a point where further increasing the threshold and reducing the data set caused little increase in the percent variance captured by PCA.

Table 4
Least squares linear fit of plot of predicted percent composition (ordinate) versus accepted percent composition (abscissa). *m*, slope; SD, standard deviation; *y*-int, *y*-intercept, *y*, predicted percent composition values.

	n	*m*	SD_m	*y*-int	$SD_{y\text{-int}}$	r^2	SD_y	Average relative error (%)	
Jatropha PLS test set	8	1.07	0.05	−0.55	0.98	0.99	1.0	5	RMSEP = 1.4
Soy PLS test set	8	0.95	0.06	0.88	1.20	0.98	1.3	4	RMSEP = 1.2
Jatropha PLS training set cross validation	10	0.97	0.04	0.29	0.66	0.99	1.3	13	RMSECV = 1.2
Soy PLS training set cross validation	10	0.99	0.03	0.00	0.44	0.99	0.9	22	RMSECV = 0.8
Jatropha NPLS test set	8	1.05	0.07	0.40	1.50	0.98	1.5	8	RMSEP = 0.8
Soy NPLS test set	8	0.94	0.06	1.55	1.20	0.98	1.3	7	RMSEP =1.5
Jatropha NPLS training set cross validation	10	0.97	0.04	0.26	0.71	0.99	1.4	28	RMSECV = 1.3
Soy NPLS training set cross validation	10	0.99	0.03	0.15	0.45	0.99	0.9	22	RMSECV = 0.8

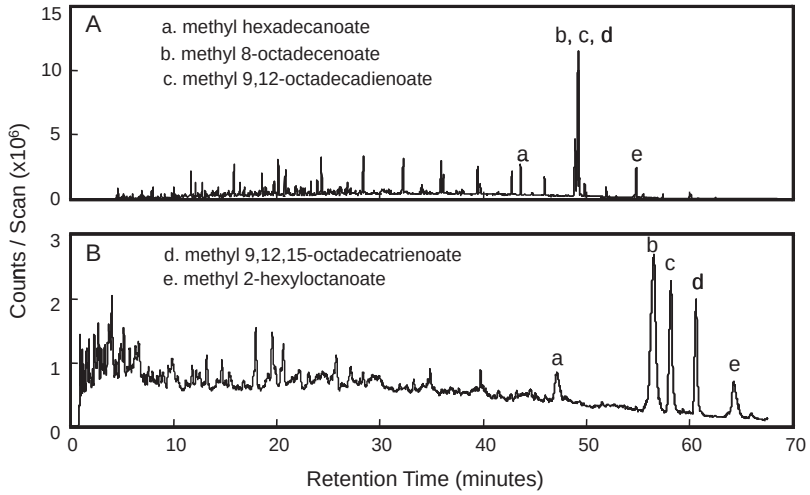

Fig. 1. GC-TIC chromatograms of a 20% blend of soy biodiesel and conventional diesel separated by (A) a nonpolar column and (B) a polar column.

3. Results and discussion

3.1. Preprocessing

The training set chromatograms listed in Table 1 were blends of conventional diesel with jatropha, soy, commercial, or zero biodiesel. We will first completely describe the treatment of the GC-TIC chromatograms before showing results for the same treatment applied to the GC–qMS chromatograms.

The baseline corrected, normalized, and aligned polar column GC-TIC training set chromatograms from Table 1 were submitted to a previously described supervised feature selection method which is based on the *f*-ratio (the ratio of between class variance divided by within class variance for each retention time pixel across all chromatograms in the dataset) [6,31]. Fig. 2 is a plot of the calculated *f*-ratios. The retention times with largest *f*-ratios are the FAMEs. The training set was reduced to retain only retention time pixels with *f*-ratios greater than 130 (the horizontal line in Fig. 2 shows the *f*-ratio = 130 threshold). These selected retention times were recorded so they could later be used to reduce the unsupervised test set down to the exact same subset of features.

Notice that the training set in Table 1 was composed only of 20% and 0% blends. If a variety of percent compositions are included in each class, the concentration variations will obscure the feedstock variations during PCA. Samples will not always be 20% or 0%, and, indeed, the test set listed in Table 1 is composed of blends that range from 10% to 30%. So right after feature selection, the reduced training set of polar column GC-TIC chromatograms was temporarily scaled to remove any percent composition variations. The scaling method was to simply divide all the data points in each reduced chromatogram by its signal at 47 min. For samples that did have biodiesel present, 47 min was the retention time of the earliest eluting FAME. This scaling was applied to every chromatogram, whether or not biodiesel was present. The scaling preserved the relative FAME ratios within each chromatogram, while it removed percent composition information from each chromatogram. This scaling preserved the FAME ratios within each feedstock and removed percent composition variations between the 0% and 30% blends. This scaling was implemented to help PCA capture the relative FAME content characteristic of each feedstock. The nonpolar column GC-TIC chromatograms also underwent the same procedures of baseline correction, sum-normalization, alignment, feature selection, and scaling by the signal at 44 min (the retention time of the earliest eluting FAME, shown in Fig. 1A).

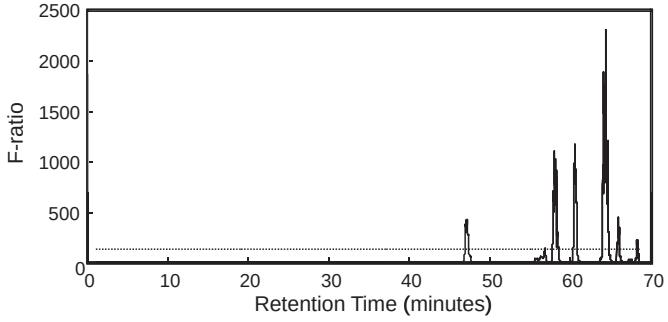

Fig. 2. The *f*-ratio for each data point in the PCA training set (polar column chromatograms) was calculated using given biodiesel class information (jatropha, soy, commercial, or none). The horizontal line marks *f*-ratio = 130. Data set reduction was achieved by retaining only data points with *f*-ratio greater than a certain threshold for further processing (in this case, threshold = 130).

3.2. Determine feedstock using pattern recognition techniques

3.2.1. Build pattern recognition model using training set

The reduced and scaled polar column GC-TIC training set chromatograms listed in Table 1 were then submitted to PCA and modeled with PC 1 and PC 2, capturing 97.97% of the total variance in the data set. The scores on PC 1 and PC 2 are plotted in Fig. 3A for the polar column chromatograms. The scores for the nonpolar column GC-TIC chromatograms are plotted in Fig. 3B. The polar column chromatograms produced a scores plot with the tightest clusters, and upon manual inspection, PCA correctly clustered the biodiesel feedstocks of the polar training set chromatograms. Fig. 3C shows scores from a PCA model built without feature selection and without scaling for the polar column training set. Fig. 3D shows scores from a PCA model built without feature selection and without scaling for the nonpolar column training set. According to manual inspection, the clustering in Fig. 3A is better than the clustering in Fig. 3B–D. Thus, it was decided the feature selection and scaling procedures were helpful, and it was decided that the nonpolar column did not provide sufficient resolution between FAMEs to allow PCA to model the characteristic FAME profiles necessary to correctly cluster biodiesel feedstocks. Therefore, the nonpolar column chromatograms were not analyzed beyond this point and no further discussion of the nonpolar column chromatograms will be included herein.

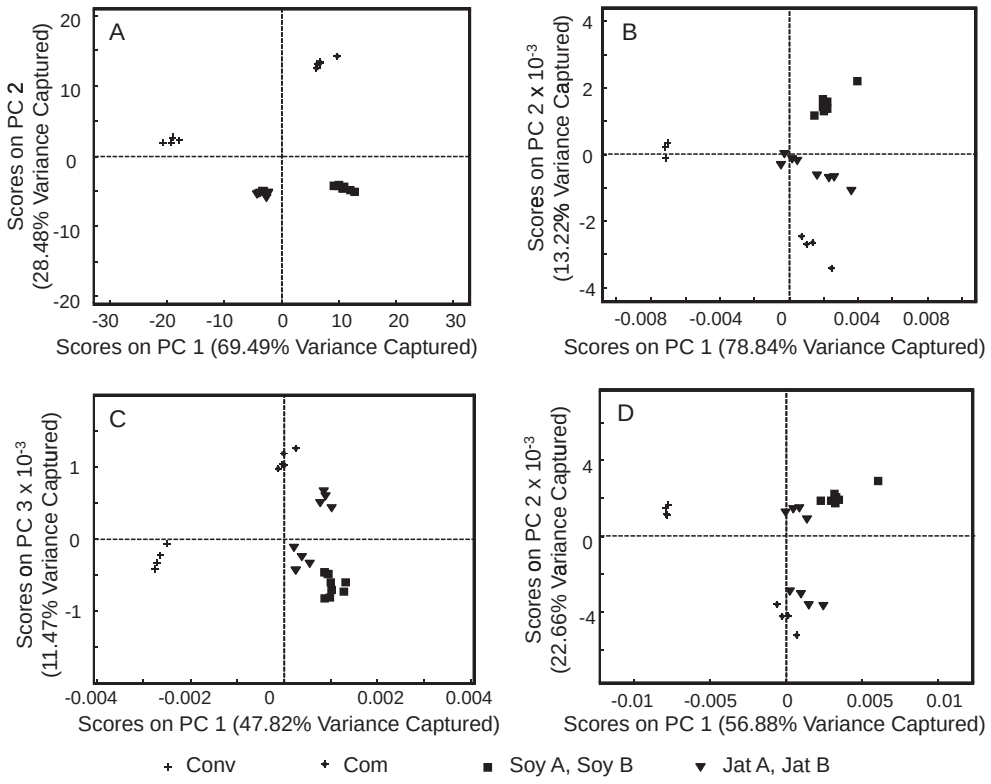

Fig. 3. Scores on PC 1 and PC 2 are plotted for (A) the polar column training set and (B) the nonpolar column training set after both training sets underwent baseline correction, sum-normalization, alignment, feature selection, and scaling. Scores on PC 1 and PC 2 are plotted for (C) the polar column training set and (D) the nonpolar column training set after both training sets underwent baseline correction, sum-normalization, and alignment. Conv = pure conventional diesel, Com = commercial biodiesel mixed with Conv, Soy A = soy biodiesel A mixed with Conv, Soy B = soy biodiesel B mixed with Conv, Jat A = jatropha biodiesel A mixed with Conv, Jat B = jatropha biodiesel B mixed with Conv.

3.2.2. Apply pattern recognition model to independent test set

Finally, the independent test set of polar column GC-TIC chromatograms listed in Table 1 was baseline corrected, sum-normalized, and aligned to the same target used for the training set. Then the exact same features that were previously selected in the training set were selected in the test set. This reduced test set underwent the scaling procedure (divide each chromatogram by its signal at 47 min) and was submitted to PCA. The test set scores are overlaid directly on the training set scores in Fig. 4. According to manual inspection, all the test set samples appear closest to the correct training set cluster, but some deviations are visible. The three test set samples that visibly deviated the most from the correct cluster were the 11th, 14th, and 15th test set polar column GC-TIC chromatograms listed in Table 1. Therefore, if we assume that a test set chromatogram is from the same feedstock as the nearest training set cluster, then we have achieved the first goal of our analytical method, which was to determine the feedstock of truly unknown blends of biodiesel and conventional diesel. The limitations of this assumption are that this only works for the feedstocks we include in the training set to build the PCA model. We were limited in having access to only three pure feedstocks (two strains of jatropha, two sources of soy, and a retail feedstock) but if more pure feedstocks could be acquired, the method described heretofore could be extended to accommodate a wider variety of feedstocks.

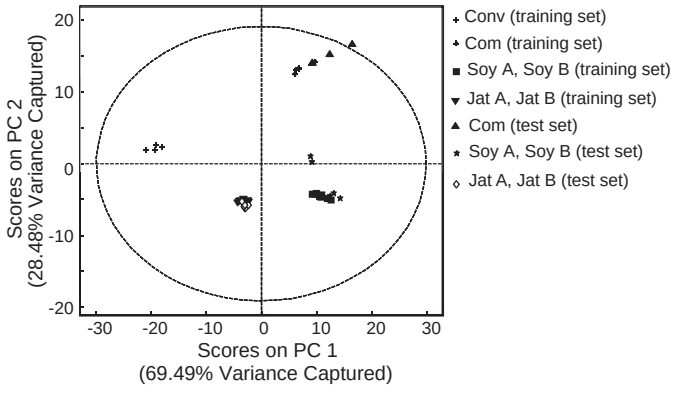

Fig. 4. Plot of test set PCA scores overlaid on training set PCA scores for the polar column GC-TIC chromatograms. The 95% confidence interval is shown.

The procedures and PCA method described heretofore correctly clustered at least 16 of the 20 test set polar column GC-TIC chromatograms. A scores plot is an excellent qualitative tool, and clusters of scores can be easily defined by the human eye, but HCA can also be used to describe clustering in a dataset. The polar column GC-TIC training set and test set scores from PCA (shown in Fig. 4) were exported out of PCA and then submitted to unsupervised Ward's method HCA to build a dendrogram of the distances in PC-space among the test set and training set samples [27]. The dendrogram in Fig. 5 clearly differentiates four clusters based upon the distance between their centers. The dendrogram properly associates each sample of the test set with its corresponding training set cluster. Upon visual inspection, even the samples in Fig. 4 scores plot that appeared to have strayed from their correct cluster are properly classified by the dendrogram.

Since the HCA dendrogram still relies on manual interpretation to identify groups, the unsupervised KNN method was also used to build a KNN model of the PC 1 and PC 2 scores of the polar column GC-TIC training set data listed in Table 1. When the polar column GC-TIC test set scores were submitted to that KNN model, it correctly predicted the biodiesel feedstock of all 20 of the test

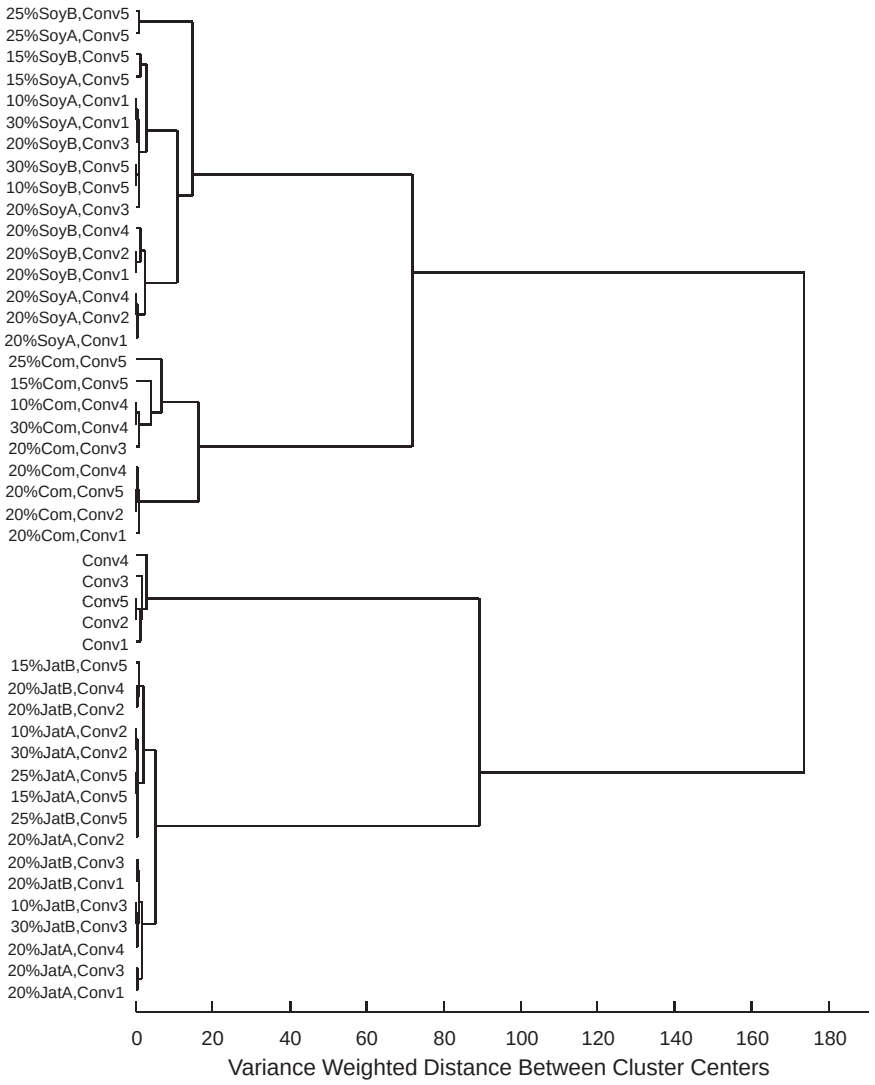

Fig. 5. HCA dendrogram built using the PCA scores shown in Fig. 4. Training set samples are the 20% blends and pure conventional diesels. Test set samples are the 10%, 15%, 25%, and 30% blends.

set samples, with complete agreement between each of the three nearest neighbors. So submitting the reduced test set to the PCA model built using the training set, then submitting the PCA scores to KNN was an accurate method of objectively classifying the test set samples.

A KNN model was also built using the reduced training set chromatograms at the chromatographic data point level, rather than using the PCA scores. When the reduced test set chromatograms were submitted to that KNN model, the biodiesel feedstock of all 20 test set samples was correctly predicted with 100% agreement between each of the three nearest neighbors.

A major limitation of using these PCA models to predict feedstock of biodiesel blends is that these particular models will not be useful for feedstocks that are not included in the model. For example, if a biodiesel blend made from animal fat was part of the test set and if it was submitted to the PCA models in Fig. 3 or Fig. 4, the animal biodiesel's score would probably appear outside the four clusters and the analyst would have to classify it as "other". Likewise, PCA scores for mixtures of feedstocks would not appear to fit into the four clusters seen in Fig. 3 or Fig. 4. So the PCA model is able to predict feedstocks only of types that were used to construct the model; in our case, we were only able to acquire five pure biodiesel feedstock samples, four of which were generously provided by the oil refinery. This is a major limitation of the PCA method described herein, but it is conceivable that analysts with access to more pure feedstocks could extend the PCA method described herein to model more types of feedstocks, and perhaps even feedstock mixtures.

3.3. Determine blend percent composition using multivariate calibration techniques

3.3.1. Build partial least squares model using training set

The PCA, dendrogram, and KNN models successfully predicted the feedstock of independent test set biodiesel blends using GC-TIC, thus fulfilling the first goal of the project. The second goal was to predict the percent composition of independent test set biodiesel blends using a PLS model built specifically for the feedstock of each test set sample.

Two training sets were built that were composed of polar column GC-TIC chromatograms for 1–30% (v/v) soy or jatropha blends, which are listed in Table 2. The chromatograms were baseline corrected, sum-normalized, and aligned (they were not reduced by feature selection, nor scaled to the first eluting FAME signal). Separate PLS models were built for the jatropha training set and the soy training set. The PLS predictions for jatropha blends had RMSEC = 0.6 and RMSECV = 1.2. The PLS predictions for soy blends had RMSEC = 0.5 and RMSECV = 0.8.

3.3.2. Apply partial least squares model to independent test set

Table 3 lists the two independent test sets that were composed of polar column GC-TIC chromatograms for soy or jatropha blends. The test sets were baseline corrected, sum-normalized, and aligned to the same alignment target used for the training set. The preprocessed jatropha test set was submitted to the jatropha PLS model, and the preprocessed soy test set was submitted to the soy PLS model. Table 3 shows the predicted blend percent composition values for the jatropha and soy test sets. The PLS predictions for the test set of jatropha blends had RMSEP (root mean squared error of prediction) = 1.4. The PLS predictions for the test set of soy blends had RMSEP = 1.2.

The predicted values were plotted versus accepted values and least squares linear fits were calculated. Ideal PLS models will produce plots of predicted values versus accepted values with least squares linear fits that have slope = 1, y-intercept = 0, $r^2 = 1$, and average relative error = 0%. Average relative errors are also reported for each test set, where average relative error is defined as the average of the absolute value of the predicted value minus the accepted value divided by the accepted value. The linear fit of the jatropha PLS test set results was $y = 1.07x - 0.55$, $r^2 = 0.99$, standard deviation of predicted composition = 1.0, average relative error = 5%, slope standard deviation = 0.05, intercept standard deviation = 0.98, for $n = 8$. The linear fit of the soy PLS test set results was $y = 0.95x + 0.88$, $r^2 = 0.98$, standard deviation of predicted composition = 1.3, average relative error = 4%, slope standard deviation = 0.06, intercept standard deviation = 1.20, for $n = 8$. These results are listed in Table 4. The PLS models were also evaluated using leave-one-out cross validation, and the results are also listed in Table 4. Some of the average relative errors were quite high, especially for the cross-validation results. This is partly due to small magnitude errors in the 1% blends having a much greater influence on the calculated average relative error than similar magnitude errors in the more concentrated blends. In addition, prediction performance degrades when these models are expected to extrapolate beyond the range of variations modeled in the training set, so if samples quite different from the training set are regressed onto the PCA and PLS models, accurate PCA and PLS results may not be expected.

3.4. Compare GC-TIC and GC–qMS pattern recognition results

A goal of this project was to compare the GC-TIC PCA results to GC–qMS PCA results for the polar column chromatograms. A PCA model was built for the polar column GC–qMS training set data listed in Table 1 after baseline correction, sum-normalization, alignment, feature selection with optimized f-ratio threshold = 500, and scaling to the maximum m/z signal at 47 min. The PCA results are shown in Fig. 6 and they were similar to the results provided by the GC-TIC data that were shown in Fig. 4. A dendrogram built using the PCA scores from the GC–qMS data was similar to Fig. 5 (not shown). The KNN model correctly predicted the biodiesel source of all 20 of the test set samples, with complete agreement between the three nearest neighbors. Therefore it was decided that the pattern recognition methods could accurately classify the independent test set polar column GC–qMS chromatograms. Next, separate NPLS models were built for the polar column GC–qMS jatropha and soy training sets listed in Table 2. The test sets listed in Table 3 were submitted to the corresponding jatropha or soy NPLS model. The predicted values were plotted versus accepted values and least squares linear fits were calculated. The results are listed in Table 4 along with the cross validation results. The NPLS predictions for jatropha blends had RMSEC = 0.8, RMSECV = 1.3, and RMSEP = 0.8. The NPLS predictions for soy blends had RMSEC = 0.5, RMSECV = 0.8, and RMSEP = 1.5. The average relative error in predicted test set sample compositions was 8% for jatropha blends and 7% for soy blends. Most of these error analysis values are slightly higher than the corresponding PLS values. The predicted blend percent composition values provided by PLS and GC-TIC data are slightly more accurate than the values predicted using NPLS and GC–qMS data.

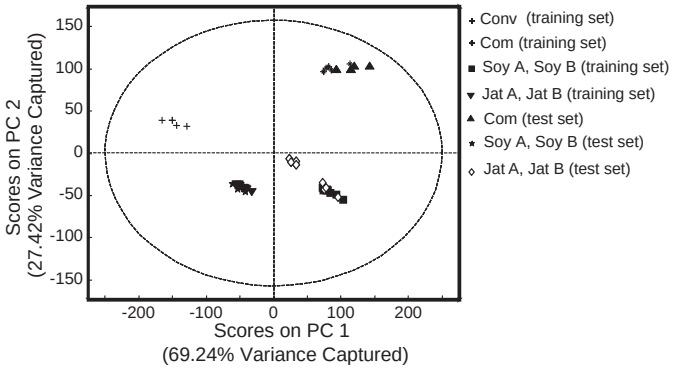

Fig. 6. Plot of test set PCA scores overlaid on training set PCA scores for the polar column GC–qMS chromatograms. The 95% confidence interval is shown.

Using the polar column instrumental method, the GC–qMS chromatograms yielded no additional information compared to the GC-TIC chromatograms. Perhaps summing the qMS dimension to yield the TIC chromatograms reduced some of the random indeterminate noise. Perhaps the greater chemical selectivity one might expect to be provided by the qMS dimension compared to the TIC dimension was obscured by the indeterminate noise. Perhaps there actually is no reason to expect GC–qMS to provide more chemical selectivity information than GC-TIC chromatograms for this system.

4. Conclusions

Blends of biodiesel and conventional diesel were prepared from five conventional diesels, two different strains of jatropha, two different soy sources, and a biodiesel from a third feedstock that was acquired from a retailer's pump. The variety of conventional diesels and biodiesels were acquired in order to construct challenging data sets containing variations expected of truly unknown samples that could be collected for sample authentication purposes. For test set blends of biodiesel and conventional diesel, the biodiesel feedstock was determined using PCA, HCA, and KNN combined with the preprocessing procedures baseline correction, sum-normalization, alignment, feature selection, and scaling to the earliest eluting FAME signal. Then the blend percent compositions of the test set samples were determined using a PLS model built specifically for the test set sample's feedstock. PLS required the chromatograms to be baseline corrected, sum-normalized, and aligned.

The polar chromatography column method provided better PCA classification results than the nonpolar column method, so we only built PLS models for the polar column data in this work. However, it is worth noting that the nonpolar column method was more thoroughly studied for a PLS application in our previous work [27]. Our previous work involved using PLS to quantify only one feedstock (the commercial feedstock) of biodiesel samples using GC–qMS with a nonpolar column, and GC × GC-TOFMS with a nonpolar primary column and polar secondary column [27]. In that previous work, the nonpolar column alone did not provide an accurate PLS model from GC–qMS data, but the GC × GC-TOFMS method did provide an accurate model precisely because of the resolution achieved among the FAMEs due to the polar secondary column. This is why we purchased a long polar column and developed the GC–qMS method described herein to build accurate PLS models using GC–qMS data.

Acknowledgments

The authors thank Chevron ETC for helpful discussions and for providing soy and jatropha biodiesel samples. This work was funded by Seattle Pacific University, the M.J. Murdock Charitable Trust, and the Montana Family Endowment.

References

[1] M.P. Dorado, E. Ballesteros, J.M. Arnal, J. Gómez, F.J.L. Giménez, Energy Fuels 17 (2003) 1560–1565.
[2] F. Adam, F. Bertoncini, V. Coupard, N. Charon, D. Thiébaut, D. Espinat, M.-C. Hennion, J. Chromatogr. A 1186 (2008) 236–244.
[3] R.G. Brereton, Chemometrics, Data Analysis for the Laboratory and Chemical Plant, Wiley, New York, 2003.
[4] D.L. Massart, Chemometrics: A Textbook, Elsevier Sciences Ltd., New York, 1988.
[5] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley-Interscience, New York, 1998.
[6] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101–110.
[7] M. Fernanda Pimentel, G.M.G.S. Ribeiro, R.S. Da Cruz, L. Stragevitch, J.G.A. Pacheco Filho, L.S.G. Teixeira, Microchem. J. 82 (2006) 201–206.
[8] F.C.C. Oliveira, C.R.R. Brandão, H.F. Ramalho, L.A.F.d. Costa, P.A.Z. Suarez, J.C. Rubim, Anal. Chim. Acta 587 (2007) 194–199.
[9] J.S. Oliveira, R. Montalvão, L. Daher, P.A.Z. Suarez, J.C. Rubim, Talanta 69 (2006) 1278–1284.
[10] R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, S.L. Rose-Pehrsson, Energy Fuels 23 (2009) 1610–1618.
[11] G. Knothe, JAOCS J. Am. Oil Chem. Soc. 78 (2001) 1025–1028.
[12] J.A. Cramer, R.E. Morris, B. Giordano, S.L. Rose-Pehrsson, Energy Fuels 23 (2009) 894–902.
[13] M.A. Aliske, G.F. Zagonel, B.J. Costa, W. Veiga, C.K. Saul, Fuel 86 (2007) 1461–1464.
[14] B. Diehl, G. Randel, Lipid Technol. 19 (2007) 258–260.
[15] M.R. Monteiro, A.R.P. Ambrozin, L.M. Lião, A.G. Ferreira, Fuel 88 (2009) 691–696.
[16] M.R. Monteiro, A.R.P. Ambrozin, M. da Silva Santos, E.F. Boffo, E.R. Pereira-Filho, L.M. Lião, A.G. Ferreira, Talanta 78 (2009) 660–664.
[17] C.M. Reddy, J.A. Demello, C.A. Carmichael, E.E. Peacock, L. Xu, J.S. Arey, Environ. Sci. Technol. 42 (2008) 2476–2482.
[18] I. Eide, K. Zahlsen, Energy Fuels 21 (2007) 3702–3708.
[19] T.A. Foglia, K.C. Jones, J.G. Phillips, Chromatographia 62 (2005) 115–119.
[20] M. Kamiński, E. Gilgenast, A. Przyjazny, G. Romanik, J. Chromatogr. A 1122 (2006) 153–160.
[21] J.V. Seeley, S.K. Seeley, E.K. Libby, J.D. McCurry, J. Chromatogr. Sci. 45 (2007) 650–656.
[22] P. Bondioli, L. Della Bella, A. Manglaviti, OCL – Oleagineux Corps Gras Lipides 10 (2003) 150–154.
[23] R.C.M. Faria, M.J.C. Rezende, C.M. Rezende, A.C. Pinto, Quim. Nova 30 (2007) 1900–1905.
[24] K.J. Johnson, S.L. Rose-Pehrsson, R.E. Morris, Petrol. Sci. Technol. 24 (2006) 1175–1186.
[25] L.S. Eberlin, P.V. Abdelnur, A. Passero, G.F. De Sa, R.J. Daroda, V. De Souza, M.N. Eberlin, Analyst 134 (2009) 1652–1657.
[26] D.S. Giordani, A.F. Siqueira, M.L.C.P. Silva, P.C. Oliveira, H.F. de Castro, Energy Fuels 22 (2008) 2743–2747.
[27] K.M. Pierce, S.P. Schale, Talanta 83 (2011) 1254–1259.
[28] H.J. Berchmans, S. Hirata, Bioresour. Technol. 99 (2008) 1716–1721.
[29] G. Graef, B.J. LaVallee, P. Tenopir, M. Tat, B. Schweiger, A.J. Kinney, J.H.V. Gerpen, T.E. Clemente, Plant Biotechnol. J. 7 (2009) 411–421.
[30] J.S. Nadeau, B.W. Wright, R.E. Synovec, Talanta 81 (2010) 120–128.
[31] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Anal. Chem. 78 (2006) 5068–5075.