



Full Length Article

Evaluation of single and multi-feedstock biodiesel – diesel blends using GCMS and chemometric methods



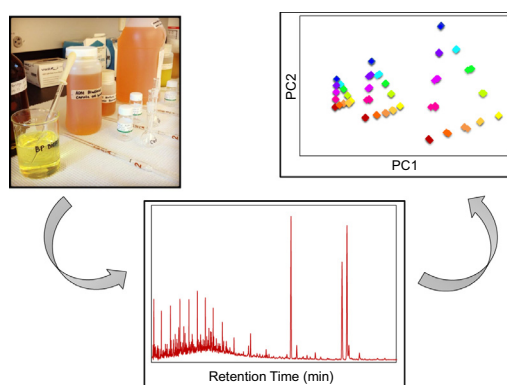
Mariel E. Flood, Mary P. Connolly, Michael C. Comiskey, Amber M. Hupp*

Department of Chemistry, College of the Holy Cross, One College Street, Worcester, MA 01610, USA

HIGHLIGHTS

- GCMS analysis of biodiesel-diesel fuel blends from single & multiple feedstocks.
- PCA categorizes fuels based on concentration and biodiesel feedstock.
- Supervised models validate training set and predict unknowns.
- kNN, SIMCA, and PLS models are compared and a sequence of methods is prescribed.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 5 July 2016

Received in revised form 8 August 2016

Accepted 15 August 2016

Keywords:

Biodiesel fuel

Multifeedstock biodiesel diesel blends

Gas Chromatography Mass Spectrometry (GCMS)

Principal Component Analysis (PCA)

Soft Independent Modeling of Class Analogy (SIMCA)

Partial Least Squares (PLS)

ABSTRACT

Single and multi-feedstock biodiesel-diesel blends were evaluated using gas chromatography mass spectrometry along with several unsupervised and supervised chemometric methods. Peak areas of diesel alkane components and/or biodiesel fatty acid methyl ester (FAME) components were evaluated using Principal Component Analysis (PCA), k nearest neighbors (kNN), soft independent modeling of class analogy (SIMCA), and partial least squares (PLS) analysis. Using PCA (an unsupervised method), samples clustered based on feedstock type (soybean, waste grease, canola, tallow) and concentration (diesel, B2-B100). Using the supervised chemometric methods, feedstock type and concentration were validated for the training set and predicted for several unknown test samples. Concentration and feedstock were predicted using kNN, while concentration alone was predicted using SIMCA. PLS also allowed prediction of concentration but the success of the prediction heavily depended on the training model used. In addition, multi-feedstock fuel blends created from 2 and 3 feedstock components (soybean, canola, tallow) were evaluated with PCA, kNN, and SIMCA. Samples clustered based on concentration and feedstock makeup/ratio in PCA. Using kNN and SIMCA, multifeedstock blends were predicted based on concentration, while kNN could be used to predict relative ratio of multiple feedstocks. The results demonstrate the utility of chemometric analysis on a complex fuel-based data set, using methods that could be performed in a variety of laboratories and fields without the need for complex data preprocessing.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Biodiesel is an accessible and renewable international fuel source. With increasing energy and environmental awareness,

* Corresponding author.

E-mail address: ahupp@holycross.edu (A.M. Hupp).

biodiesel has become a more attractive energy source and is sold commercially across the nation. According to the U.S. Energy Information Administration, biodiesel production has steadily been increasing over the last eight years, almost doubling in production capacity during this time (from 2008 to 2015: 678–1268 million gallons) [1]. Biodiesel can be produced from both plant and animal sources, both of which are miscible in petro-diesel. As such, it is sold as both 100% biodiesel and as biodiesel-diesel blends in varying concentrations as low as 2% (B2). It is most commonly sold in blended form, ranging from B2 to B20. In the United States, most biodiesel is produced from soybean, canola, and corn plants, with poultry and tallow contributing as well [1]. Biodiesel is composed of fatty acid methyl esters (FAMES) produced via transesterification reaction of raw plant oils or animal tallow in the presence of a catalyst and methanol. The FAMES produced in the biodiesel are preserved from the triglycerides in the raw plant or animal material, yielding unique structural differences based on feedstock type. For example, a biodiesel produced from coconut oil may contain C6:0 to C18:2 FAMES with a majority from C12:0, while a biodiesel produced from soybean may contain C16:0 to C18:3 FAMES with a majority from C18:2 [2]. The fuel quality and efficiency (cetane number, etc) are dependent on FAME content, which is thus dependent on the feedstock type [2]. Biodiesel-diesel blends are quickly becoming common in commercial manufacturing due to financial and environmental issues and a need exists for a method to quickly and efficiently identify the type of feedstock/s and concentration of the blend [1].

The concentration of biodiesel in blended fuels has been measured using various spectroscopic techniques, including infrared [3–7] and fluorescence [8,9] spectroscopy combined with multivariate analysis methods, such as Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA), or Partial Least Squares (PLS). Electrospray ionization mass spectrometry has been combined with chemometric methods to analyze different feedstocks and blend compositions of biodiesel [10,11]. Each of these methods provides a summative analysis of the fuel. That is, the individual components of the fuel sample contribute to an overall signal that is acquired and used to describe the fuel. In order to investigate the individual components that make up a fuel, a chromatographic method is warranted. Gas chromatography-mass spectrometry (GCMS) is the most widely used technique for the analysis of biodiesel blends because it requires limited sample preparation and allows for the separation and detection of individual FAMES [12–15]. These GC studies have utilized a variety of column chemistries, yet typically include biodiesels from a limited number (one to three) of single component feedstocks. Petroleum products (e.g. diesel, gasoline) have been investigated and successfully fingerprinted using various instrumental and analytical techniques with various chemometric methods [16–20].

Furthermore, the analysis of multi-feedstock biodiesel blends is an extremely relevant avenue of biodiesel-diesel fuel blends research. Mixing multiple biodiesel feedstocks, either as pure biodiesel or biodiesel-diesel blends, could lead to the discovery of more efficient fuels with desirable properties. Moreover, during the production of blends, diesel companies may not ensure that all biodiesels used are from the same feedstock. Since biodiesel is commercially available in a variety of feedstocks, it is likely to be blended either at the station or inside an automobile fuel tank as a result of fueling up at different gas stations. In addition, biodiesels could be mixed deliberately to create a “designer” biodiesel that takes into account the cetane number and other properties, potentially optimizing the fuel’s FAME content to produce the most energy efficient biodiesel fuel [21]. As such, work has begun in this area of research. Multidimensional GC (GCxGC) has been employed to successfully characterize a limited number of biodiesel and biodiesel-diesel blends (B5) from multiple feedstocks [22] and

chemometric methods have been utilized to examine various fuel blends [23]. The effects of using multifeedstock mixtures in diesel engines in terms of fuel properties, engine performance, and emissions has also been investigated [24–26], thus showing an increased interest in multifeedstock biodiesel mixtures. However, these techniques are complicated and expensive. Thus, a need exists for simple and direct classification and prediction of feedstock type in these complex and interesting biodiesel diesel fuel blends using standard laboratory instrumentation.

Previous work in our research laboratory has determined optimal separation conditions of several biodiesel feedstocks on three different GC column chemistries and utilized unsupervised chemometric methods to characterize feedstock type [27–29]. The current research expands previous research by employing separation and chemometric methods for biodiesel-diesel blends made from these same feedstocks and several diesel sources. Specifically, four different concentrations of biodiesel (B2, B5, B10, B20) made from five different feedstocks (soybean, canola, waste grease, and two types of animal fat) were blended with three different diesels and analyzed via GCMS. Several chemometric techniques, including PCA, k nearest neighbors (kNN), soft independent modeling of class analogy (SIMCA), and partial least squares (PLS), were used to cluster the blended fuels by feedstock and concentration. Several unknown test samples were utilized to determine the performance of the predictive models. This work has direct application to the field of fuels analysis, where unknown fuel samples are routinely analyzed in an analytical, forensic, or environmental application to determine feedstock type, origin, and biodiesel content. As more biodiesel-diesel blends are manufactured and commercially used, it is imperative that a method to analyze these blends be in practice. The method prescribed herein could be used with simple laboratory instrumentation (GCMS) and computer software (chemometric software) to provide classification of biodiesel feedstock type and concentration in a biodiesel-diesel blended fuel.

2. Material and methods

2.1. Chemicals

Biodiesel fuel samples were obtained from various manufacturers across the United States (Minnesota Soybean Processors (Brewster, MN, soybean), ADM Company (Decatur, IL, canola), TMT Biofuels (Port Leyden, NY, waste grease), Iowa Renewable Energy (Washington, IA, tallow) and Texas Green Manufacturing (Littlefield, TX, tallow)) and stored in their original containers at 4 °C. Diesel fuel samples were obtained from Shell (Worcester, MA), Sunoco (Germantown, MD) and Flynn’s (Shrewsbury, MA), and transferred to amber bottles stored at 4 °C. Prior to sample preparation, the biodiesels and diesels were allowed to warm to room temperature and inverted to ensure homogeneity.

Biodiesel-diesel blends (10 mL total volume) were prepared at 2, 5, 10, and 20% biodiesel by volume (B2, B5, B10, B20). 1 mL of each blend was diluted to 50 mL total volume in hexane (Fisher Scientific). Two laboratory-produced test blends were created using soybean (B3) and waste grease (B12) feedstocks (labeled Blends A and B, respectively), while two blended fuels with unknown feedstock type and concentration were obtained from commercial sources (BP Diesel, Worcester, MA; Hess Diesel, Kinnelon, NJ; labeled Blends C and D, respectively). Each test and unknown sample was diluted 1:50 in hexane. Analysis of the unknown blends was conducted in a “blind” manner. Multifeedstock blends were prepared at 5, 10, and 20% biodiesel by volume using pairs of biodiesel feedstocks (soybean, canola, tallow) in ratios of 1:1, 1:3, and 3:1. Additional multifeedstock blends using all three feedstocks were prepared at 20% biodiesel by volume in

ratios of 4:1:1, 3:2:1, and 2:2:2. All prepared samples were stored in brown bottles at 4 °C and were allowed to warm to room temperature before analysis.

2.2. Instrumentation

Separations were performed using an Agilent 6890 gas chromatograph coupled with an Agilent 5973 mass spectrometer (Agilent Technologies, Santa Clara, TX) and have been described previously [27]. The GC was equipped with a polar ZB-Waxplus column (Phenomenex, 30 m × 0.25 mm × 0.25 μm). The oven temperature was optimized for separation of FAME components in the biodiesel as follows: 60 °C (hold 2 min) to 150 °C at 13 °C/min to 230 °C at 2 °C/min. High purity helium was used as a carrier gas at a flow rate of 1.5 mL/min. Each sample was warmed to room temperature and manually injected in triplicate (1 μL from 10 μL syringe, Hamilton Company) with a split ratio of 50:1. The inlet and transfer line temperatures were held at 250 °C and 280 °C, respectively. An electron-impact ionization source was utilized with a quadrupole mass analyzer operated in full-scan mode (m/z 20–600) with a sampling rate of 4.94 scans/s. The mass spectrometer source and quadrupole were held at 230 °C and 150 °C, respectively.

2.3. Data analysis

FAME identification was performed by searching a mass spectra library (NIST mass spectral search program version 2.0a, Gaithersburg, MD) and by comparison to FAME standards (Supelco, Bellefonte, PA). Peak areas were identified using integration at a common threshold (Enhanced Chemstation D.03.00.6aa, Agilent). In these samples, six to nine FAME peaks from each biodiesel could routinely be identified in all samples while hundreds of peaks were visible for the diesel. From each biodiesel feedstock, we chose the six most abundant FAMES, and in each diesel, the six most abundant alkanes (C12–C17) to describe the chromatogram. The most abundant FAMES were not the same in each biodiesel, however, the most abundant alkane peaks were the same between the diesel samples used. If a FAME was not present or did not meet the abundance criteria, a value of zero was used for the peak area for that FAME; this technique assured alignment of peak retention times throughout the data set. In our data set, based on the biodiesel feedstocks we used, the most prevalent FAMES in each sample are similar to one another. However, FAME components do tend to vary in other feedstocks and the importance to use the most abundant FAMES from each sample type is paramount.

The purpose of using peak areas rather than using every data point (the entire chromatogram) was to create a routine and user-friendly approach that any laboratory could potentially use. The use of an entire chromatogram poses many additional problems (alignment, additional computer power, specialized software/programming knowledge) that laboratories without chemometric experience may find daunting. This analysis method can be used with little knowledge of chemometric and preprocessing methods and simple, user friendly chemometric software. Previous research in this area has shown that using peak areas rather than the entire chromatogram does not bias simple data sets [28]. Multifeedstock samples were treated in a similar way, with the six most abundant FAMES in the mixture used.

The area of each peak was normalized using the total area under all selected peaks (Microsoft Excel 2007), and mean-centered (Pirouette 4.5, Infometrix, Bothell, WA) prior to subsequent chemometric analysis (Pirouette 4.5). Normalization is performed to ensure limited variation over the sample set (minimizes run to run injection variation) while mean centering subtracts the variable mean creating a sample set centered at the origin. These pre-

processing techniques are vital for chemometric methods and ensure that the major variations identified are from real chemical sources rather than chemical or instrumental noise.

Unsupervised pattern recognition methods identify groupings or clusters in the data without knowledge of class information (feedstock type, concentration, etc) [30]. Principal Component Analysis (PCA) allows for simplification of the original data set by identifying the variables that contribute to the maximum variation in the data provided. Typically two to three Principal Components (PCs) represent 80–90% of the variation and can be used to describe trends in the data set. The remaining PCs represent variables that do not describe notable variation in the data set and thus are not necessary. The scores for the principal component vectors were plotted in Excel. PCA has the ability to associate samples of similar origin and chemical composition and discriminate samples of different origin or chemical composition.

Supervised pattern recognition methods use class information (feedstock type, concentration, etc) to derive a training model based on unique properties of the different classes. The model is used to verify known materials and to predict classifications of new or unknown materials. Methods of this type include k-Nearest Neighbor (kNN), Soft Independent Modeling of Class Analogies (SIMCA), and Partial Least Squares (PLS) [31]. For kNN, k number of nearest neighbors are compared to a sample. A sample is grouped to a class based on the physical (Euclidean) distance between it and the closest group in the training set, where the majority of the k samples reside [30,31]. The k value is determined based on optimization of the training set and the number of samples within each class. For SIMCA, PCA is performed on each class separately in the training model. The model exists as a multi-dimensional box around each class type. A sample is classed by projecting it into the model space and comparing it to each of the class boxes created in the training set. If the distance from the sample to the box is smaller than the spread of the box, or the sample falls within the box, it is categorized as belonging to that class [30,31]. If the sample is not close to a training model box, it will remain uncategorized. The shape of the categories has a very real impact on whether an unknown will be classed correctly or at all in SIMCA. For PLS, the goal is to model a relationship between the data matrix and a single response variable (concentration or other physical property). Instead of looking for clusters within the data, the model searches for latent variables (chemical components) to describe the data and to model the variable [31]. Each of these models were used in this research and their use and limitations will be described in more detail in the following section.

3. Results and discussion

3.1. Separation of components in biodiesel-diesel blends

Chromatograms representing the separation of components in a pure diesel sample, B2 to B20 tallow biodiesel-diesel blended samples, and a pure biodiesel sample are shown in Fig. 1. The majority of components from the diesel elute in the first 15 min of the run while the FAME components from the biodiesel elute between 18 and 30 min. The alkanes in the diesel are easy to identify as they are the evenly spaced, most abundant peaks (labeled in Fig. 1a). As the concentration of biodiesel increases in blended samples, so do the peak heights/areas of the FAME components (observed in sequence from Fig. 1b–e). Typically, the most abundant FAMES can be observed in the low concentration blends, while the minor FAMES that can easily be observed in the B100 (Fig. 1f) are more difficult to identify in lower concentration blends. In fact, the minor FAMES can be seen in the B20 yet are missing in the B10.

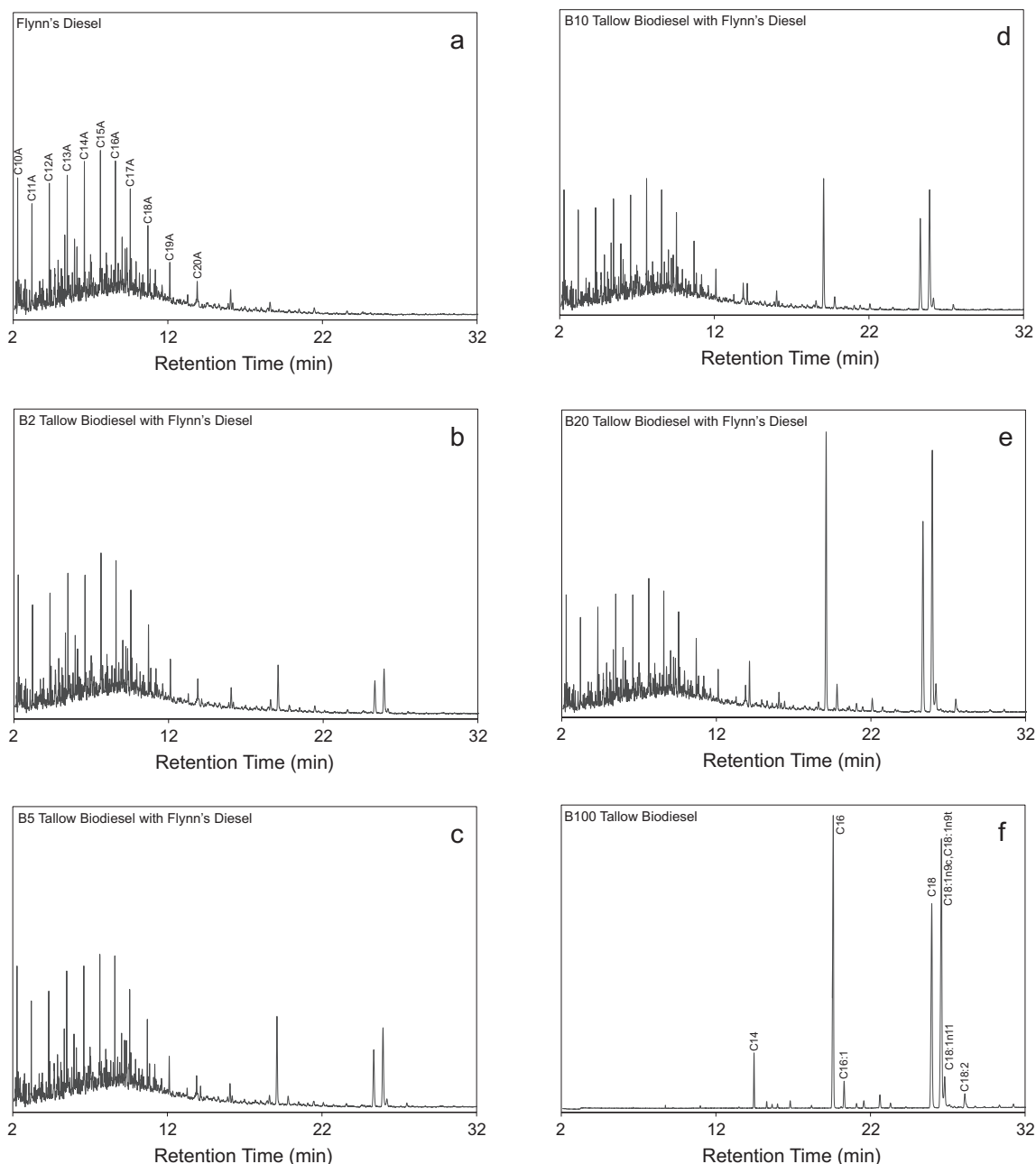


Fig. 1. Chromatograms of (a) Flynn's diesel, (b) B2 tallow biodiesel with Flynn's diesel, (c) B5 tallow biodiesel with Flynn's diesel, (d) B10 tallow biodiesel with Flynn's diesel, (e) B20 tallow biodiesel with Flynn's diesel, (f) B100 tallow biodiesel.

The peak heights of the diesel components do not change much over the range of blends observed here. These chromatograms show typical separation of components for all the biodiesel blends studied. However, the tallow and canola biodiesels typically have more minor components as compared to soybean and waste grease biodiesels [27]. As such, peaks are not equally observed across feedstocks and concentrations. Thus, the peak areas of only the most abundant biodiesel FAME peaks and most abundant diesel alkane peaks are analyzed in this study. In this way, the same number of variables are used regardless of feedstock and concentration.

3.2. Unsupervised chemometric analysis of biodiesel-diesel blends

PCA was performed on each set of biodiesel blends (B2, B5, B10, B20); the scores plot for B20 is shown in Fig. 2a. Looking at any one

concentration, distinct clustering based on feedstock type is observed. Replicate runs of the same sample are neatly clustered together, showing little run to run variation. Tallow samples from two different animal sources are clustered similarly to one another, while soybean and waste grease samples are also clustered similarly. The canola sample is distinct from the other feedstocks. Most of this distinction occurs in the first PC (65.0%), but further differentiation of the canola sample occurs in the second PC (34.3%). This clustering was observed when analyzing the pure biodiesels [28], yet continues to hold for these blended samples even at very low concentrations (e.g. B2, not shown).

PCA was further performed on the entire set of samples, including all pure biodiesels, all blend concentrations, and all pure diesels for all feedstocks (5 types) and all diesels (3 types); the resulting PC scores plot is shown in Fig. 2b. Here distinct clustering is observed

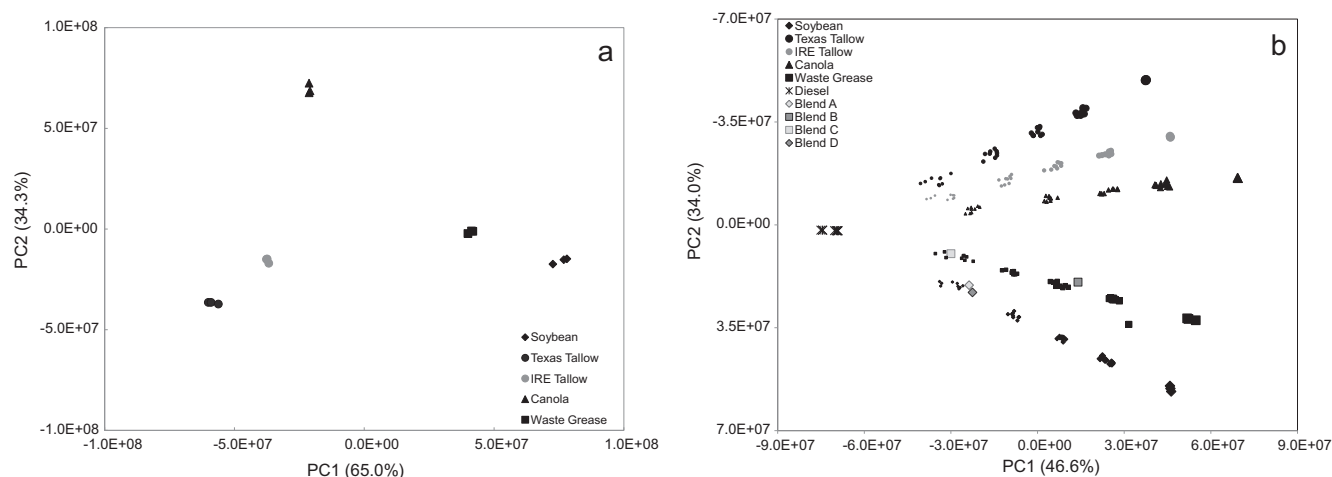


Fig. 2. PCA of (a) B20 biodiesel blends (soybean, tallow, canola, waste grease feedstocks) and (b) pure diesel and biodiesel blends (soybean, tallow, canola, waste grease feedstocks) at B2 (smallest symbols), B5, B10, B20, and B100 (largest symbols) concentrations. Test blends A–D are shown as projected into this PC space.

based on concentration and feedstock type, but not based on the diesel sample. The first PC (46.6%) shows clustering based on concentration, with diesel samples clustered on the far left of the plot in negative PC1 space, B2 samples to the right of those, followed by B5, B10, B20, and B100 (pure biodiesels) on the far right in positive PC1 space. The second PC (34.0%) shows distinction based on feedstock type, with tallow and canola samples loading negatively and waste grease and soybean samples loading positively. The third PC (14.8%, not shown) further differentiates the canola and tallow samples, with canola loading negatively and both tallow samples loading positively. In total, the first three principal component axes describe 95% of the variation in the data set.

The loadings indicate that these clusters arise based on differences in both FAME and alkane concentration. In PC1, the C18:1n9c,t and C18:2 FAME concentrations and the C12, C13, C15, and C16 alkane concentrations distinguish the pure biodiesel from the diesel. The biodiesel FAME components load positively while the diesel alkane components load negatively. This closely matches the clusters in the scores plot that transition from pure diesel on the far left (negative values) to pure biodiesel on the far right (positive values) of the plot. In PC2, differences in the C16:0, C18:0, C18:1n9c,t, and C18:2 concentrations force clustering to occur based on biodiesel feedstock. In fact, biodiesels with high levels of C18:1n9c,t (canola and tallow) load positively while biodiesels with high levels of C18:2 (waste grease and soybean) load negatively in PC2. There is not a significant contribution from the diesel alkane peaks in PC2. In PC3, biodiesels with high levels of C18:0 and C16:0 (both tallows) load positively, while biodiesels with low levels of C18:0 and C16:0 (canola) load negatively. The soybean samples have low to moderate levels of C18:0 and C16:0 and load close to the axis. This leads to the differentiation of canola and tallow samples. In the first three PCs, these biodiesel FAME peaks all load more strongly than any of the diesel alkane peaks. In fact, it takes the fourth principal component (3.4%) to differentiate one of the diesel samples away from the other diesel samples. Thus, it is the FAMES (C16:0, C18:0, C18:1n9c,t, C18:2) that are most descriptive of the differences in this set of biodiesel blended fuels.

It is interesting that there is little distinction based on diesel type observed for the B5, B10, or B20 samples using the first three principal components. Upon close examination of the first two PCs, there is some trend in the pure diesel and B2 samples, which have replicate samples of the same diesel (Sunoco) more tightly clustered together as compared to the rest of the samples within that

specific set. This trend can be seen for the pure diesels and almost all the B2s. The goal of this study was not to investigate differences in diesel type and as such, only six of the numerous alkane and aromatic diesel components were chosen. Thus, it is not surprising that diesel type was not able to be differentiated. However, we were interested to see if diesel could be distinguished if concentration variance was removed. Thus, a sample set containing B10 tallow blends (2 types) made with each diesel (3 types) was analyzed (Fig. 3a). Diesel type is marked in the main plot, while biodiesel feedstock type is marked in the inset. Interestingly, PC1 is able to cluster samples based on diesel type, with loadings indicating a difference in C14 and C18 alkanes. PC2 then clusters based on biodiesel type, with loadings indicating a difference in C18:0 and C18:2 FAMES. Thus, four distinct clusters result. To push the data set a bit, the B20 tallow blends were added to the B10 set and analyzed (Fig. 3b). Here PC1 and PC2 allow distinction between concentration and feedstock type, respectively, but not diesel brand. PC3 (inset of Fig. 3b) allows some differentiation between diesel type, as Sunoco samples are pulled apart from the Shell and Flynn's samples. Thus, when concentration is a factor, diesel type/brand becomes more difficult to distinguish. Yet, when concentration is the same throughout the sample set, diesel type/brand is more easily recognized as a variation in the data. It is noteworthy to find a difference in diesel type with such a small selection of diesel peaks. Further study of blends made with a wide range of diesel types utilizing many diesel components in the PCA would need to be included before conclusions could be drawn in regards to diesel type identification in biodiesel-diesel blends.

To further evaluate the clustering ability of PCA on a biodiesel-diesel data set, four test samples were created and evaluated using PCA along with the main data set just described (Fig. 2b). Blends A and B (lab prepared as B3 soybean and B12 waste grease, respectively) clustered with the B2 soybean samples (on the far right of the B2 cluster, closer to the B5 cluster than to the pure diesel cluster) and the B10 waste grease samples (on the far right of the B10 cluster, closer to the B20 cluster than to the B5 cluster), respectively. Blends C and D (commercial samples) clustered with the B2 waste grease samples and the B2 soybean samples, respectively. While the feedstock and concentration were not known at the time of collection, we can surmise from the scores plot that the diesel was mixed with a low concentration (less than 5%) of either soybean or waste grease biodiesel. In fact, at concentrations less than 5% (B5), no separate labeling is required at the pump [32]. This evaluation is performed as an unsupervised method, without the

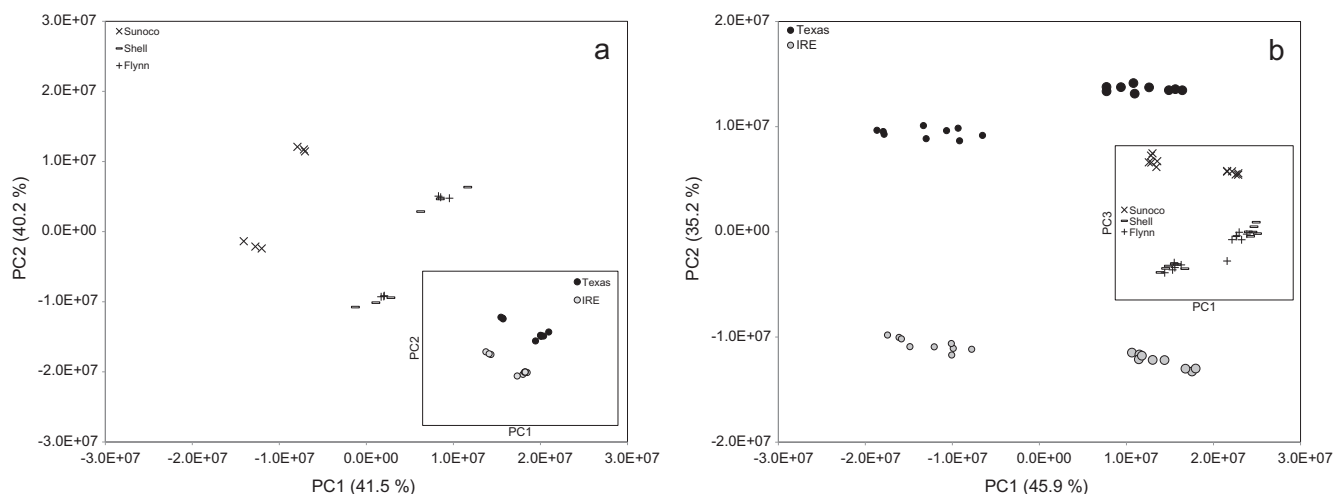


Fig. 3. PCA of two tallow biodiesel fuel types mixed with three diesel fuel types at (a) B10 concentration and (b) B10 (smallest symbols) and B20 concentration (largest symbols). In (a), diesel type is marked in the main plot, while biodiesel feedstock type is marked in the inset. In (b), biodiesel feedstock is marked in the main plot, while diesel type is marked in the inset.

use of a training set or a defined, prescribed method, making PCA an evaluative method. A supervised method is required to provide further certainty of unknown cluster determination.

Overall, PCA was used to reduce the dimensionality of the data set from 13 to 4 variables (chemical components). This process could likely be done manually, but would be a subjective process prone to more error than PCA. This method could be applied to other data sets where the sources of variation are more difficult to analyze (where more peaks are present) and a less subjective, more automated approach is warranted. Such a case could be used for fuel samples taken from various locations in a forensic arson investigation, where the fuel has been burned on different materials for different lengths of time, for example.

3.3. Supervised chemometric analysis of biodiesel-diesel blends

kNN was performed on the entire set of samples including all pure biodiesels, all blend concentrations (B2 to B20), and all pure diesels for all feedstocks (5 types) and all diesels (3 types). Predictive models were created based on: (1) feedstock type, (2) concentration, and (3) feedstock and concentration together. Using 4 nearest neighbors ($k = 4$), each sample in the training set was assigned to the correct feedstock, concentration, or both, respective of the training model used, validating the training model.

Each of the test and unknown samples was tested with the training set categorized to a feedstock and concentration (Table 1). Both laboratory created test blends (A and B) were classified correctly for feedstock. These unknowns were created outside of the concentration categories used for the training set, yet were classified according to their closest neighboring set. The commercial unknown blends (C and D) were categorized based on their nearest neighbors as less than 5% biodiesel (both at B2 level) and as feed-

stocks that are likely used in commercial sources in the United States (waste grease and soybean). When using feedstock and concentration together, test samples were predicted into the same categories as when determined independently.

SIMCA was performed on the entire set of samples. Predictive models were created for groups based on: (1) feedstock type, (2) concentration, and (3) feedstock and concentration together. Results from the SIMCA analysis are shown in Table 1. The SIMCA models were able to classify tests and unknowns based on concentration, but were less successful classifying based on feedstock or both concentration and feedstock. The shapes of the models and the position of the unknowns relative to the model directly impacts whether classification will be successful. That is, if the unknowns show enough variation from the predictive model, they will not be classified with that group. The models for concentration resemble flattened spheres and cover a wider range of the graphed space. This model allows for more variation in the placement of the unknowns and thus allow for a greater predictive quality. SIMCA predicted concentrations correlate with kNN predicted concentrations. The models for feedstock, however, resemble long skinny sticks. They cover a long, but narrow range of graphed space. More of the unknowns vary from the models for feedstock type and thus have less predictive quality. Only one test sample (B) was classified correctly according to feedstock. When both feedstock and concentration are used to create the training set, the models are tightly centered around the replicate samples. The unknowns do not align well with the models and no predictions were made. SIMCA proves to be less powerful for classification for this data set as compared to kNN.

PLS is a predictive method based on a measured variable. As such, feedstock cannot be considered a variable when using PLS and thus only concentration was utilized to create a training set.

Table 1

kNN ($k = 4$) and SIMCA analysis of test and unknown biodiesel blends. – indicates the concentration was not predicted by the training model indicated.

Test sample	Actual	kNN predicted conc	kNN predicted feedstock	kNN predicted feedstock & conc	SIMCA predicted conc	SIMCA predicted feedstock	SIMCA predicted feedstock & conc
A (lab)	B3 soy	B2	Soy	B2 soy	B2	–	–
B (lab)	B12 waste grease	B10	Waste grease	B10 waste grease	B10	Waste grease	–
C (BP)	<B5 ?	B2	Waste grease	B2 waste grease	B2	–	–
D (Hess)	<B5 ?	B2	Soy	B2 soy	B2	–	–

Two training sets (using data that were not normalized) were considered. In the first set, all feedstocks with concentrations spanning from 0% (pure diesel) to 100% (pure biodiesel) were included (204 samples in training set). In the second set, all feedstocks with concentrations spanning from 0% to 20% were included (189 samples in training set). A plot of predicted concentration versus measured concentration is shown in Fig. 5, and several validations and predictions are shown in Table 2.

There are two items of note regarding the PLS model. First, the overall fit of the B0 to B20 data is best for the B0 to B20 training set, while the overall fit of the B0 to B100 data is best for the B0 to B100 training set. In other words, including the B100 data points allows for more accurate determination of biodiesels with higher concentrations, while removing the B100 data points from the training set allows for better determination of biodiesels with lower concentrations. Second, for the B0 to B100 training set the spread in the predictions of the concentration extremes (B0, B100) was much greater (stdev = 8.1, 7.9, respectively) than the spread in the predictions of the other concentrations (B2, B5, B10, B20: stdev = 5.9, 4.5, 3.5, 3.1). For the B0 to B20 training set the spread in predictions for most concentrations are more similar to one another and much smaller than in the other training set (B0, B2, B5, B10, B20: stdev = 1.8, 1.1, 0.6, 0.9, 2.1), the only exception being the B100 s where the predictions are very poor as well (stdev = 7.0). Overall, the B0 to B100 training set was well suited for larger concentrations of biodiesel (e.g. the B20 s and B100 s) but poorly validated the training set at low concentrations and poorly predicted the unknowns (which have low concentrations). Alternatively, the B0 to B20 training set was poorly suited for predicting the concentration of the B100 s, but performed well for validating the training set at low concentrations and predicting the unknowns. It is important to note again that the data used in PLS is not normalized. With chromatographic data, normal fluctuations from injection will occur that are usually corrected with normalization. However, to keep concentration inherent in the data, normalization should not be performed with PLS methods. Thus, fluctuation in the input data will likely lead to fluctuation in the output model.

The data set used is rich and complicated. There are many samples with varying concentration, feedstock, and diesel type. To ensure that the predictions were not suffering from the fullness of the data, PLS was performed using one feedstock at a time. Thus, all concentrations of only one feedstock (e.g. soybean) with various diesel types was used to create the training set (45 samples in each training set). With this training model, similar problems with the fits occur. That is, the B0 to B100 data set fits well for the higher concentrations and performs worse for the lower concentrations. The B0 to B20 data set suits the lowest concentrations well and predicts the unknowns of that feedstock type better (Table 2). The spread in the data is greater for the B0 to B100 predicted

concentrations (stdev = 1.9–11.5) than in the B0 to B20 predictions (stdev = 0.5–5.1). However, the spread in the data for the B0 to B100 set is lower when only one feedstock is considered in the training model than when all feedstocks are utilized.

Overall, PLS was able to predict concentration values regardless of which concentrations were included in the model. The predictions differ based on the training set and care should still be taken when creating the model. That is, many concentrations near the expected concentrations should be utilized in the training model so that predictions are most accurate. In this way, the PLS model offers advantages to the other supervised chemometric methods, as a prediction will be made regardless of the classes used. However, analysts must be wary to use a predictive model that is appropriate and also to utilize multiple methods to ensure consistency in prediction.

3.4. Unsupervised chemometric analysis of multifeedstock biodiesel-diesel blends

PCA was next utilized to analyze a set of samples made from 2 or more biodiesel feedstock types blended with the same diesel, shown in Fig. 4. As only one diesel type was used in this section of the study, only the most abundant FAME peak areas were utilized in the PCA for multifeedstock blends (e.g. soy/tallow mix). Clustering and differentiation of 2 component multifeedstocks in B5, B10, and B20 samples can be seen in Fig. 4a while of 2 and 3 component multifeedstocks in B20 samples can be seen in Fig. 4b. The PCA plots resemble triangles where the vertices correspond to the single feedstock of that concentration (e.g. B20 soy) and the outer edges correspond to the 2 component mixtures (3:1, 1:1, and 1:3) of that concentration (e.g. B20 3:1 soy:tallow). The samples that more chemically resemble the single feedstock blends are closer in space to one another (i.e. 3:1 soy:tallow is closest to the soy vertex, next closest in space to the tallow vertex, and furthest from the canola vertex). In Fig. 4a, three different triangles exist, one for each concentration (B5, B10, B20). The trends are the same throughout each triangle regardless of biodiesel concentration.

The first PC (56.9%) in Fig. 4a allows for distinct separation of overall biodiesel concentration and makeup of the tallow and canola blends, while the second PC (28.6%) and third PC (14.4%, not shown) allows for further separation of the soybean blends. These three PCs make up 99.9% of the variation in the data set. The loadings indicate that these clusters arise mainly based on differences in the C18:1n9c,t in PC1, C18:2 in PC2, and C16:0 and C18:0 in PC3, in agreement with the FAMES identified in the single component blends.

In Fig. 4b, one triangle is present, representing the 2 and 3 component mixtures of the B20 concentration. The vertices and edges

Table 2

PLS analysis of test and unknown biodiesel blends. – indicates the concentration was not predicted by the training model indicated.

Test sample/training sample	Actual	Predicted concentration (%) using B0 to B100 training model (all feedstocks)	Predicted concentration (%) using B0 to B20 training model (all feedstocks)	Predicted concentration (%) using B0 to B100 training model (single feedstock)	Predicted concentration (%) using B0 to B20 training model (single feedstock)
A (lab)	B3 soy	9.7	3.8	9.7 (soy)	3.9 (soy)
B (lab)	B12 waste grease	12.8	11.3	22.2 (waste grease)	14.5 (waste grease)
C (BP)	<B5 ?	0	2.7	0 (waste grease)	2.3 (waste grease)
D (Hess)	<B5 ?	(–)9.5	1.4	0 (soy)	1.7 (soy)
B2 tallow Flynnns	2	(–)3.6	2.0	–	–
B10 tallow Sunoco	10	15.1	10.2	–	–
B20 canola shell	20	28.7	23.4	–	–
B100 canola	100	89.4	51.6	–	–

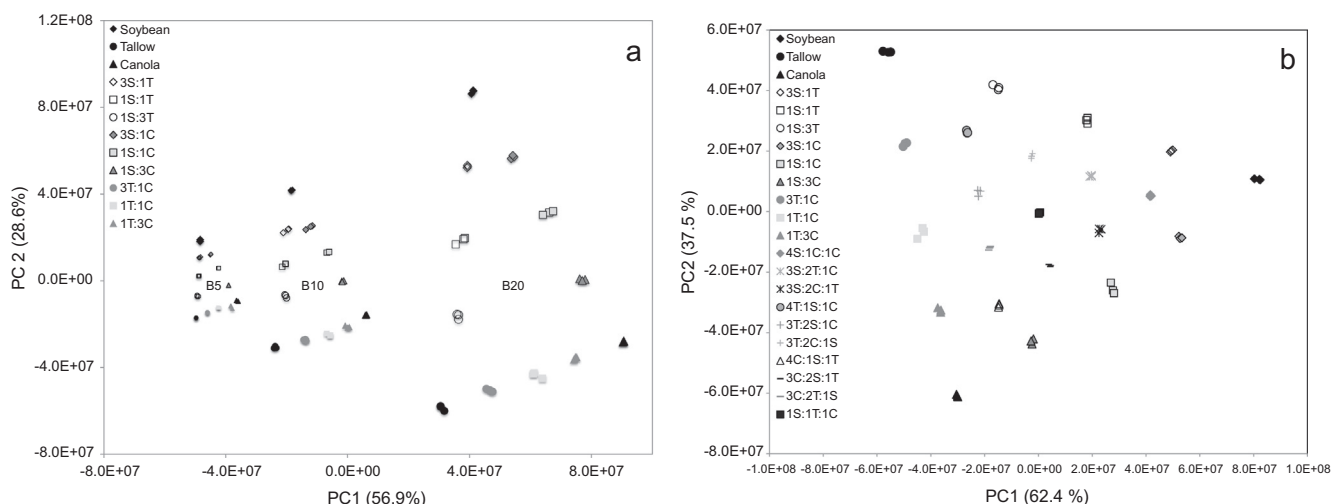


Fig. 4. PCA of mixed feedstock biodiesel blends (a) 2 component mixtures (soybean, canola, tallow) at B5 (smallest symbols), B10, and B20 (largest symbols) concentrations and (b) 2 and 3 component mixtures (soybean, canola, and tallow) at B20 concentration. The PCA plots resemble triangles where the vertices correspond to the single feedstock of that concentration (e.g. B20 soy) and the outer edges correspond to the 2 component mixtures (3:1, 1:1, and 1:3) of that concentration (e.g. B20 3:1 soy:tallow).

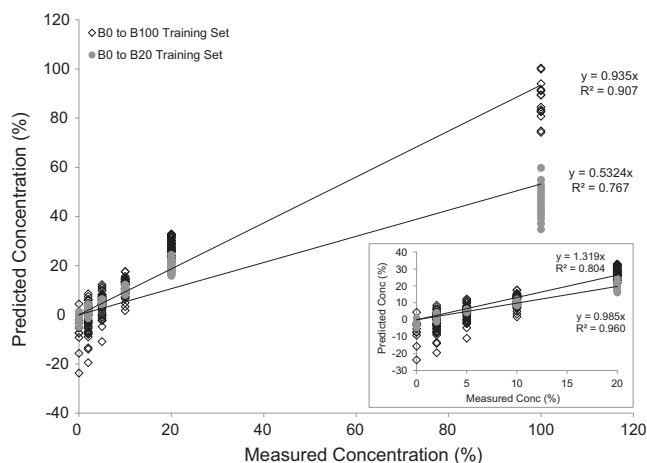


Fig. 5. Predicted concentration via PLS versus measured concentration for biodiesel blends using a training set composed of all feedstocks with either B0 to B100 concentrations or B0 to B20 concentrations. Inset shows regression for B0 to B20 region.

of the triangle contain the same data points (2 component mixtures) as the B20 triangle in Fig. 3a. Some data points load positively in Fig. 4b, but load negatively in Fig. 4a, causing the triangle to appear transposed between figures. Even though the sign flips, the correlation of the loadings to the scores does not and that is the meaningful result here. The inner data points in Fig. 4b represent the 3 component multifeedstocks. They form another triangle inside the first and represent 4:1:1 and 3:2:1 mixtures. The data point in the center of the triangle represents the 1:1:1 mixture. Thus, all ratios of the multifeedstock samples are clearly distinguishable from one another. In fact, the samples that more chemically resemble the single feedstock blends are closer in space to one another (i.e. 4:1:1 soy:tallow:canola is closest in space to the soy vertex and equally spaced between the tallow and canola vertices). The sample that contains equal amounts of each feedstock (1:1:1) is spaced equidistant from each vertex.

The first and second PCs in Fig. 4b allow for distinct separation of feedstock makeup. These PCs represent 99.9% of the variation in the data set. When concentration is not a factor, a third principal component is not needed to differentiate between sample. The

loadings indicate that these clusters arise based on differences in the C18:1n9c,t and C18:2 concentrations for PC1 and differences in the C16:0 and C18:0 concentrations for PC2, in agreement with the FAMES identified in the single component blends. The four FAME peaks that throughout these analyses have indicated differences in the feedstocks are specific to this data set. If another feedstock was utilized that varied in other FAME components, the loadings would likely indicate those FAMES in addition or instead. The scores plots and loadings that result from PCA are dependent on the data set that is used. Thus, if the data set changes (i.e. the variables and/or sample types), the scores and loadings could change accordingly. These methods could thus be applied to other feedstock types, other fuel types, or other samples entirely to identify similarities and differences that arise in sample type.

The plots in Fig. 4a and b showcase the capability of pairing an efficient GC separation with a dynamic analysis program like PCA. These samples were easily and clearly clustered and categorized based on both concentration and feedstock makeup. Only one diesel type/brand was utilized throughout this multifeedstock study and it is hypothesized that diesel type/brand would not be a major source of variation when multiple feedstocks are mixed at several different concentrations. These later factors (concentration and feedstock) likely would contribute more to the variation than diesel type. Additional analyses, ideally using the entire data set (all diesel peaks), could determine if diesel type plays a role.

3.5. Supervised chemometric analysis of multifeedstock biodiesel-diesel blends

kNN was utilized to analyze a set of samples made from 2 or more biodiesel feedstock types blended with the same diesel. Predictive models were created based on: (1) feedstock ratio and (2) concentration. Using 4 nearest neighbors ($k = 4$), each sample in the training set was assigned to the correct feedstock ratio and concentration. Six known test samples were tested on the training sets with results shown in Table 3. Each of the unknowns was categorized correctly according to feedstock ratio and concentration.

SIMCA was performed on the entire set of samples. Predictive models were created for groups based on: (1) feedstock ratio and (2) concentration. The SIMCA models were able to classify test samples based on concentration, but were not successful classifying based on feedstock ratio. All test samples were classified

Table 3

kNN ($k = 4$) and SIMCA analysis of test multifeedstock biodiesel blends. – indicates the concentration was not predicted by the training model indicated.

Test sample	Actual	kNN predicted concentration	kNN Predicted feedstock Ratio	SIMCA predicted concentration	SIMCA predicted feedstock ratio
E	B10 1S:1T	B10	1S:1T	B10	–
F	B10 1S:3T	B10	1S:3T	B10	–
G	B10 3S:1T	B10	3S:1T	B10	–
H	B20 1S:1T	B20	1S:1T	B20	–
I	B20 1S:3T	B20	1S:3T	B20	–
J	B20 3S:1T	B20	3S:1T	B20	–

according to correct concentration, but none of the samples were classified as any class for feedstock ratio (Table 3). The concentration box shapes were broader and more forgiving, while the feedstock ratio box shapes were long and skinny. Again, kNN proves to be a powerful method of prediction for both concentration and feedstock, while SIMCA was more challenged for classification of the multifeedstock blends.

4. Conclusion

In this study, biodiesel-diesel blended samples from varying feedstocks and of varying concentrations were evaluated using GCMS and several chemometric methods. Blended samples clustered based on feedstock type and concentration. Several test blends, including two lab prepared samples and two commercial samples, were predicted using training models based on kNN, SIMCA, and PLS methods. The two lab prepared blends were classified with the correct feedstock and concentration groups. The two commercial blends sold as pure diesel were classified as B2 samples, well below the B5 limit imposed by federal regulation to be sold without biodiesel signage. Multifeedstock blends created from 2 and 3 component feedstocks clustered based on concentration and feedstock makeup. kNN was able to provide correct classification and prediction for the test samples, however, SIMCA was only able to provide correct classification for concentration. PLS was able to predict concentration, but the predictive success depended greatly on the training set that was utilized.

For our data, it is actually not surprising that kNN works as well as it does for the training set that was provided, as the test compounds fall within the same categories as the training set. It is likely to fail when the unknown or test samples do not line up so well with the training set. We see that in our data for the B13 as it actually classifies as a B10, since there is no B13 category/class. If there are unknowns that do not fall within the training categories, the method is likely to fail, which could mean gross misclassification of samples depending on the number and type of categories in the training set. The same type of misclassification could happen with feedstock type if the unknown/test set are of a different chemical nature or from a different source than that in the training set (i.e. including sunflower biodiesel with the training set provided here). Overall, a user will always end up with a result from kNN, it may be that the classification is inaccurate if the training set is not robust. This is a bit different than the SIMCA approach where a sample can be either misclassified or not classified at all if the unknown/test falls outside of the prediction box. Any slight variation from the model, regardless of the shape, can cause misclassification. We see that more clearly in our data, as

the SIMCA model fails for feedstock type and feedstock ratio. With differences in concentration, the SIMCA model would likely fail as well. It actually is surprising that the B13 sample is predicted as a B10, but it shows that the model can have some variability depending on the inputs in the training set. Both kNN and SIMCA are inherently limited by the training set and must be very robust and broad in order to categorize and predict both feedstock and concentration. Interestingly, PLS is not as limited by the training set and predictions can occur for unknowns/tests outside of the categories provided in the training set. PLS is thus a technique that can be used with a smaller training set, where you do not have a class for every concentration that could be possible in an unknown sample. PLS does not work for classification of feedstock, since feedstock is not a numerical variable, yet it is a powerful tool for predicting a numerical property, such as concentration. However, in order to gain the most accurate predictions, the training set provided for PLS must be appropriate for the unknowns being tested.

Overall, the methods discussed within demonstrate the utility of chemometric analysis on a complex data set, using methods that could be performed in any lab without the need for complex data preprocessing. While these methods have been applied to biodiesel, diesel, and biodiesel-blended samples from single or multifeedstocks, a variety of other complex fuel samples could be analyzed in this way. From this research, we propose a multistep approach where feedstock is first classified using kNN with a diverse training set of many different types of feedstock at various concentrations. As a secondary step, concentration is predicted using a PLS training set based on several concentrations of that specific feedstock. In particular, these methods could be used to classify truly unknown fuel samples, analyze changes in fuels with time, or monitor industrial conditions. As more biodiesel-diesel blends are manufactured and commercially used, it is imperative that a method to analyze these blends be in practice. The method defined here is simple enough that it could be used in any number of labs to provide initial differentiation of feedstock type and concentration of biodiesel in a biodiesel-diesel blend.

Funding

This research was supported by Gerard P. and Clare S. Richer (MEF), Jacqueline H. and George A. Paletta, Jr. (MCC), and Alumni/Parents Summer Research Fellowships (MPC), the University Syringe Program Grant from Hamilton Company, and the College of the Holy Cross.

Acknowledgements

The following companies are acknowledged for generously providing biodiesel samples: Minnesota Soybean Processors, ADM Company, TMT Biofuels, Texas Green Manufacturing, and Iowa Renewable Energy. Additional thanks to Scott Ramos (Infometrix) for several insightful conversations, Julian Goding, Ryan Dean, and Carolyn Brown (Holy Cross) for acquiring several chromatograms, and to Dr. Jacolin Murray (NIST) for acquisition of a diesel sample.

References

- [1] <<http://www.eia.gov/biofuels/biodiesel/production/archive/#2016-2017>> [accessed August 4, 2016].
- [2] Moser BR. *Vitro Cell Dev Biol* 2009;45:229–66.
- [3] Veras G, de Araujo Gomes A, da Silva AC, de Brito ALB, de Almeida PBA, de Medeiros EP. *Talanta* 2010;83:565–8.
- [4] Rocha WFC, Vaz BG, Sarmanho GF, Leal LHC, Nogueira R, Silva VF, et al. *Anal Lett* 2012;45:2398–411.
- [5] Pimentel MF, Ribeiro GMGS, da Cruz RS, Stragevitch L, Pacheco Filho JGA, Teixeira LSG. *Microchem J* 2006;82:201–6.
- [6] Alves JCL, Poppi RJ. *Analyst* 2013;138:6477–87.

- [7] Gontigo LC, Guimaraes E, Mitsutake H, de Santana FB, Santos DQ, Neto WB. Fuel 2014;117:1111–4.
- [8] Tomazzoni G, Meira M, Quintella CM, Zagonel GF, Costa BJ, de Oliveira PR, et al. J Am Oil Chem Soc 2014;91:215–27.
- [9] Prakash J, Mishra AK. Fuel 2013;108:351–5.
- [10] Eide I, Zahlsten K. Energy Fuels 2007;21:3702–8.
- [11] Prates RGD, Augusti R, Fortes ICP 2010;24:3183–8.
- [12] Pinto AC, Guarieiro LLN, Rezende MJC, Ribeiro NM, Torres EA, Lopes WA, et al. J Braz Chem Soc 2005;16:1313–30.
- [13] Ragonese C, Tranchida PQ, Sciarone D, Mondello L. J Chromatogr A 2009;1216:1992–8997.
- [14] Schale SP, Le TM, Pierce KM. Talanta 2012;94:320–7.
- [15] Pierce KM, Schale SP. Talanta 2011;83:1254–9.
- [16] Eide I, Zahlsten K. Energy Fuels 2005;19:964–7.
- [17] Zahlsten K, Eide I. Energy Fuels 2006;20:265–70.
- [18] Hupp AM, Marshall LJ, Campbell DI, Smith RW, McGuffin VL. Anal Chim Acta 2008;606:159–71.
- [19] Gaines RB, Hall GJ, Frysinger GS, Gronlund WR, Juare KL. Environ Forensics 2006;7:77–87.
- [20] Johnson KJ, Rose-Pehrsson SL, Morris RE. Energy Fuels 2004;18:844–50.
- [21] Knothe G. Energy Fuels 2009;22:1358–64.
- [22] Tiyaopongpattana W, Wilairat P, Marriott PJ. J Sep Sci 2008;31:2640–9.
- [23] Johnson KJ, Rose-Pehrsson SL, Morris RE. Pet Sci Technol 2006;24:1175–86.
- [24] Habibullah M, Rizwanul Fattah IM, Masjuki HH, Kalam MA. Energy Fuels 2015;29:734–43.
- [25] Iqbal MA, Varman M, Hassan MH, Kalam MA, Hossain S, Sayeed U. J Cleaner Prod 2015;101:262–70.
- [26] Martinez G, Sanchez N, Encinar JM, Gonzalez JF. Biomass Bioenergy 2014;63:22–32.
- [27] Goding JC, Ragon DY, O'Connor JB, Boehm SJ, Hupp AM. Anal Bioanal Chem 2013;405:6087–94.
- [28] Flood ME, Goding JC, O'Connor JB, Ragon DY, Hupp AM. J Am Oil Chem Soc 2014;91:1443–52.
- [29] Soares EJ, Yalla GP, O'Connor JB, Walsh KA, Hupp AM. J Chemometr 2015;29:200–12.
- [30] Massart DL, Vandeginste BGM, Deming SM, Michotte T, Kaufman L. Chemometrics: a textbook. Elsevier; 2003.
- [31] Sharaf MA, Illman DL, Kowalski BR. Chemical analysis. Chemometrics, vol 82. John Wiley & Sons; 1986.
- [32] <http://www.afdc.energy.gov/fuels/biodiesel_blends.html> [accessed August 4, 2016].