EFFECTS OF DATA PRETREATMENT ON THE MULTIVARIATE STATISTICAL ANALYSIS OF CHEMICALLY COMPLEX SAMPLES

By

John William McIlroy

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Criminal Justice—Master of Science

2014

ABSTRACT

EFFECTS OF DATA PRETREATMENT ON THE MULTIVARIATE STATISTICAL ANALYSIS OF CHEMICALLY COMPLEX SAMPLES

By

John William McIlroy

Multivariate statistical procedures, such as principal component analysis (PCA), are often utilized to differentiate and associate a large number of complex samples consisting of thousands of variables. When samples with similar chemical compositions are compared, chemical differences between samples are often overshadowed by non-chemical variation. Therefore, in order to provide meaningful statistical comparisons and differentiate complex and highly similar samples, these non-chemical sources of variation must be minimized, often accomplished by implementing data pretreatment procedures.

In this work, ten diesel samples from different service stations were analyzed in triplicate by gas chromatography-mass spectrometry. The resulting chromatograms were processed with data pretreatment procedures, including baseline correction, smoothing, retention time alignment, and normalization, to evaluate the enhanced discrimination in PCA achieved by minimizing non-chemical variation. For each pretreatment procedure, metrics were developed to evaluate the effect on the chromatogram as well as the PCA results. Normalization and alignment resulted in the greatest enhancement in association of replicate samples, while smoothing and baseline correction were shown to have minimal effect. By applying data pretreatment procedures, replicate samples were closely associated with one another and differentiated from the other diesel samples, allowing for differentiation of complex and similar samples.

ACKNOWLEDGEMENTS

I would like to thank everyone that has helped me throughout my years at Michigan State, while completing my Master's degree. I owe a huge thank you to Ruth Smith for all of her help, guidance, and training during the years as my Master's degree advisor. She has also provided me with numerous opportunities for learning and experience. I would also like to thank Vicki McGuffin and Dan Jones for their guidance on completion of this project. As my Ph.D. advisors, they provided many of the suggestions and guidance throughout the development of this research. They have greatly helped to develop my skills as a scientist. Thank you to Ruth Smith, Vicki McGuffin, and Steven Dow for also taking time to serve on my committee.

I owe a debt of gratitude to all of the McGuffin group students, Forensic Master's students and Chemistry Graduate students that have helped me while at Michigan State. In particular, I would like to thank Lucas Marshall, for analyzing the diesel samples that were utilized in this work and Steve Halpin for helping to write several of the Matlab algorithms that were utilized for this work.

I would also like to thank my friends and family who have supported me throughout my education. I would especially like to thank my wife Katie for all of her support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1. Chemical Analysis of Forensically Significant Samples

Evidence found at a crime scene is a crucial aspect of many police investigations and is often required in court to establish guilt or innocence. Forensics scientists examine the evidence, draw a conclusion from the results obtained, and provide their conclusion in court as an expert's opinion. The examination of the evidence often consists of an identification or a comparison between a questioned and known sample [1-3]. Often the comparisons are facilitated by instrumental analyses, which generate a chemical fingerprint of the samples for comparison [2, 4]. The forensic scientist must then make the determination whether the question and known samples are consistent with one another (*i.e.* a "match").

Even when the conclusions that forensic scientists draw are based on scientific tests, all testing has errors and uncertainties associated with the measurement. These must be taken into account by the forensic scientist when forming their expert opinion. In addition, the opinions of the forensic scientists are susceptible to outside influence and human bias [3]. A 2009 report from the National Academies of Science (NAS) identified the need to address the "accuracy, reliability, and validity" of forensic testing to help reduce testing error and bias [3]. In order to address this concern, forensic research began to focus on the use of statistical procedures to aid in comparison of chemical fingerprints, assign a statistical confidence to forensic tests, and to help minimize errors and human bias [5-18].

1.1.1. Forensic Chemistry

The area of forensic science that is of most interest to analytical chemists is the application of instrumental analysis to forensically relevant samples, either in forensic toxicology or forensic chemistry [2]. Forensic toxicologists generally examine biological fluids for the presence of drugs or poisons and relevant metabolites. Forensic chemists generally utilize analytical techniques such as gas chromatography-mass spectrometry (GC-MS) and infrared spectroscopy to analyze physical evidence such as fire debris, explosives, and controlled substances [2].

Due to the expansive range of evidence types that a forensic chemist may analyze, this work will focus on a single, complex example, the GC-MS analysis of ignitable liquids (specifically diesel) for the detection of accelerants in fire debris. GC-MS is one of the most common analytical instruments in a forensic laboratory. It encompasses a separation and identification aspect (discussed in Section 1.1.2) and is used to confirm the identity of a compound [2, 4, 19]. Diesel fuel will provide a forensically relevant sample that is chemically complex, consisting of hundreds of compounds. Further, diesels from different sources vary in chemical composition, based on the refinery at which the fuel was produced and additives from the individual service stations at which the fuel was obtained. The composition and properties of diesel fuel used for this work will be further discussed in Chapter 2.

1.1.2. Gas Chromatography-Mass Spectrometry

GC-MS is a hyphenated technique that combines gas chromatography, which separates compounds in a mixture based on boiling point or polarity, and mass spectrometry, which breaks compounds into fragments that are characteristic of the compound. Each compound has a reproducible retention time and as well as a unique and reproducible fragmentation pattern, under specific and controlled conditions. The retention time and fragmentation pattern are then utilized to determine the identity of the compound [1, 20]. The GC-MS instrument (Figure 1-1) used in this research is similar to those found in forensic laboratories. The output from the GC-MS is a chromatogram, which contains peaks for compounds present in the sample. Example chromatograms of several ignitable liquids including diesel fuel (a), lighter fluid (b), and gasoline (c) are shown in Figure 1-2. Characteristic compounds are labeled in each chromatogram.

1.1.2.1. Gas Chromatography Theory

Chromatography is a broad class of analytical techniques that is used to separate sample mixtures. In all chromatography methods, the mixture is dissolved into a mobile phase which is moved across a stationary phase. Compounds in the mixture interact differentially with the stationary phase [19]. Compounds that interact more with the stationary phase are more retained, while compounds that interact less with the stationary phase move through the chromatography system quickly and are retained to a lesser extent [19]. As a result of the different extents of interaction with the stationary phase, compounds in a sample mixture are separated, resulting in a chromatogram (Figure 1-2). Compounds with similar chemical properties will generally elute close to one another.

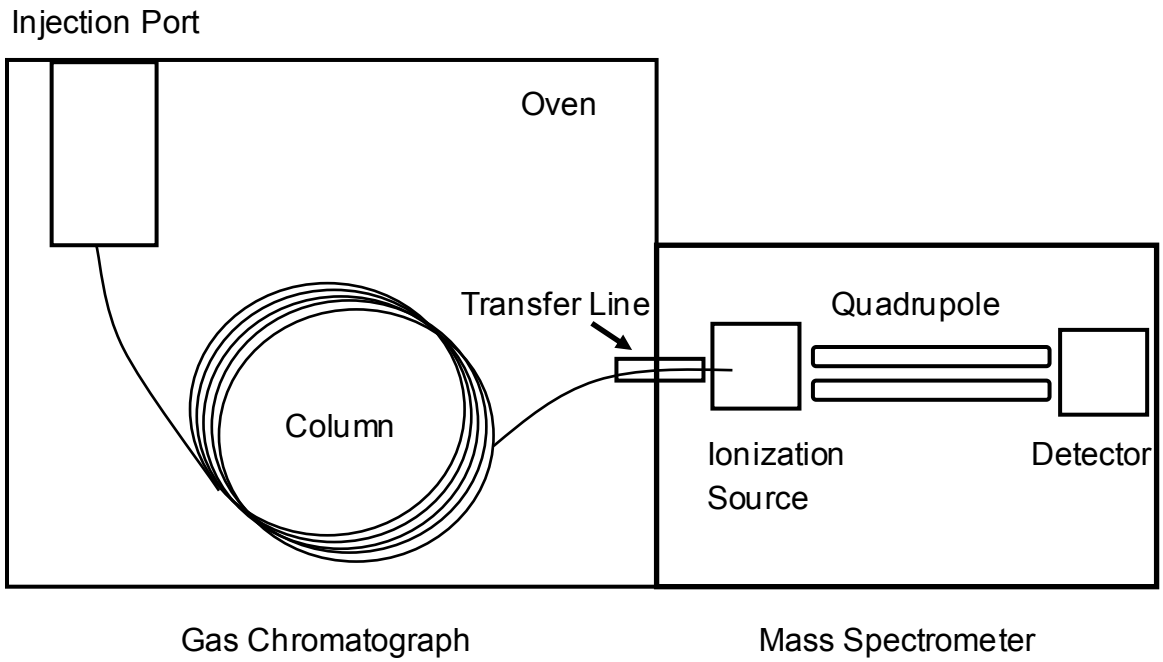


Figure 1-1. Diagram of gas chromatograph-mass spectrometer.

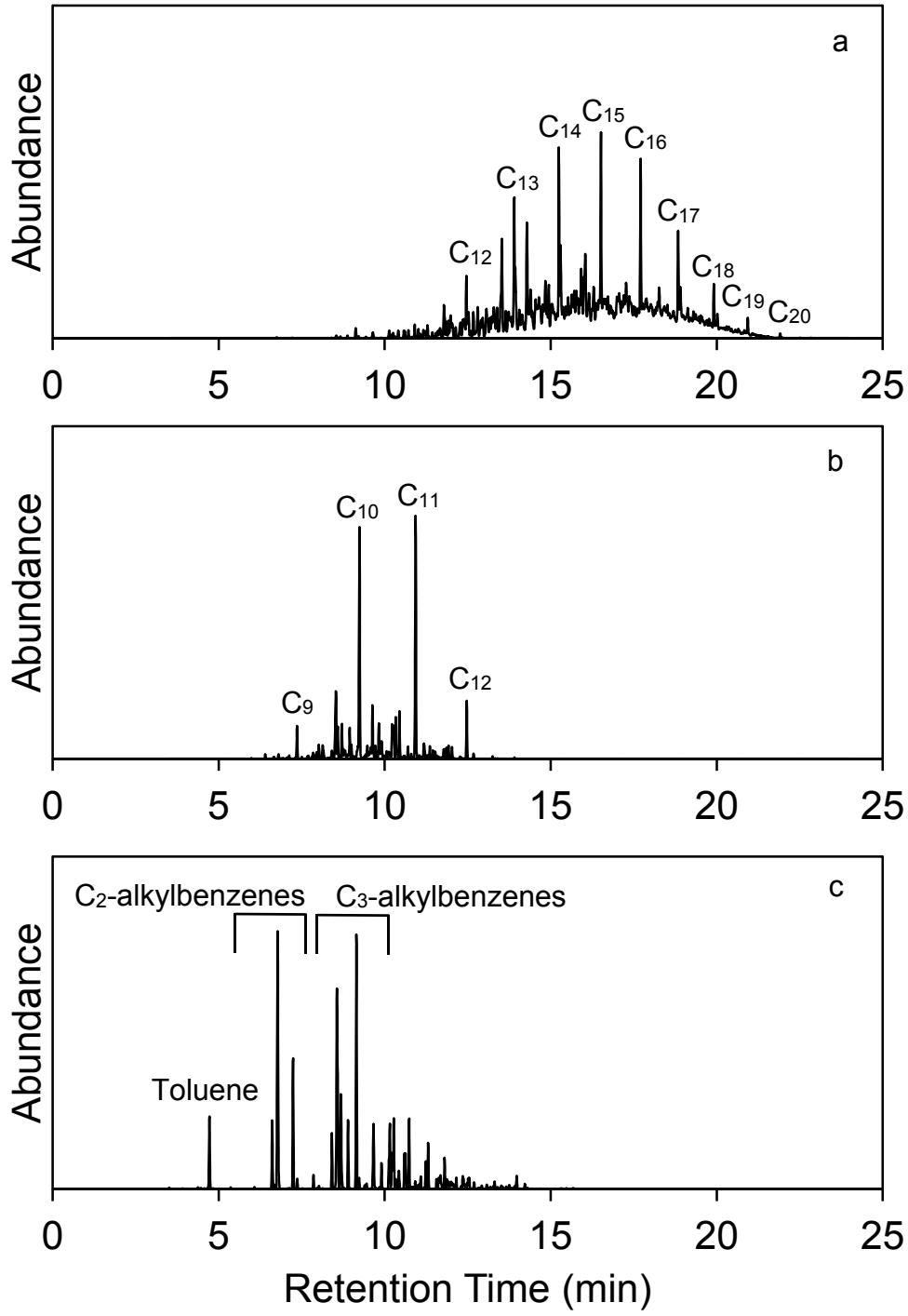


Figure 1-2. Example chromatograms generated for ignitable liquids including diesel fuel (a), lighter fluid (b), and gasoline (c).

One of the most common chromatography systems is the gas chromatograph. In GC analysis, the sample is volatized into the gas phase, using high temperatures, in the injector port of the instrument (Figure 1-1). The gaseous mixture is transferred onto a column using a carrier gas, typically helium, hydrogen, or nitrogen, which is the mobile phase in the separation [19-21]. The stationary phase is contained on the inside of the column. The column is typically inside a temperature-controlled oven, and temperature can be varied during the analysis to change the speed and efficiency of the separation [19]. The compounds from the sample interact with the stationary phase in the column, causing separation. The same compound should have the same retention time from sample to sample on the same instrument and under the same experimental conditions. The effluent from the column then travels into a detector.

1.1.2.2. Mass Spectrometry Theory

One of the most common detectors in forensic science is the mass spectrometer [1, 2]. The MS not only detects the compound as it elutes from the GC, but it can also help to identify the compound [19]. In GC-MS, the end of the column is positioned directly into the ion source of the MS, via a transfer line (Figure 1-1). The MS is under high vacuum, in order to allow molecules from the column to traverse the MS, without colliding with air molecules. While samples are introduced at atmospheric pressures, the low flow rates utilized in capillary GC allow the vacuum pumps in the MS to remove the air and mobile phase molecules, resulting in a high vacuum.

In order to be analyzed by MS, the compounds must first be ionized. In this work, compounds were ionized using electron ionization (EI). In the ion source, a heated

filament produces high energy electrons which are accelerated across the ionization space, towards an anode [19, 22]. The molecules from the sample traverse the ionization space perpendicular to the electron beam. As the sample molecules pass close to the beam, some of the energy is transferred from an electron to a molecule, which causes the molecule to ionize (for a positive ion by removing an electron). However, often more energy is imparted to the molecule than is required for ionization. The additional energy causes the molecule to fragment. Each compound fragments in a unique and reproducible manner under these conditions, allowing for identification of the compound using a known standard or a reference database [19, 22].

After the molecules have been ionized, the mass of each ion is determined using a mass analyzer. In this work, a quadrupole mass analyzer was used. The quadrupole typically has four cylindrical metal rods. Positive ions are directed from the ion source through a series of electrostatic lenses and focused into the quadrupole [20]. The quadrupole has a direct current (DC) applied to each rod as well as an oscillating radio frequency (RF) current. Opposite pairs of rods are electrically connected, with adjacent rods always having opposite charges for both the DC and RF currents [20]. The electric field cause by the combination of DC and alternating RF potentials results in ions moving along the quadrupole in a corkscrew trajectory [19]. Only ions with a narrow range of mass-to-charge (*m/z*) ratio can pass through the quadrupole at a given set of DC and RF potentials. The DC and RF potentials can be scanned, which allow a range of *m/z* to pass [22]. Ions that pass through the mass analyzer strike the conductive surface of an electron multiplier, creating a cascade of electrons which are then detected and converted into an electronic signal [20].

1.1.2.3. GC-MS Chromatogram

After GC-MS analysis, a chromatogram of the data is generated, consisting of an array of abundances at discrete retention time points. The chromatogram indicates the time at which individual compounds elute from the GC. The abundance at each retention time point in the chromatogram originates from the mass spectrum generated in the MS. Each spectrum shows the ions resulting from the fragmentation of the compound that eluted at that point. The fragmentation pattern relates to the structure of that compound, which allows for identification. The total ion chromatogram is the sum of the abundance of all *m/z* at each retention time point [1].

Another example of a GC-MS chromatogram of diesel fuel is shown in Figure 1-3a. This diesel fuel was analyzed using a slower temperature program than the fuels shown in Figure 1-2. The major normal alkane peaks are labeled for reference. Figure 1-3b shows an expanded region of the pentadecane (C_{15}) peak. The points indicate where mass spectra were collected and summed to create the total ion chromatogram (TIC) abundance. The red line shows where the mass spectrum in Figure 1-4 was obtained. The mass spectrum shows the *m/z* value of the molecular ion and fragment ions resulting from this compound.

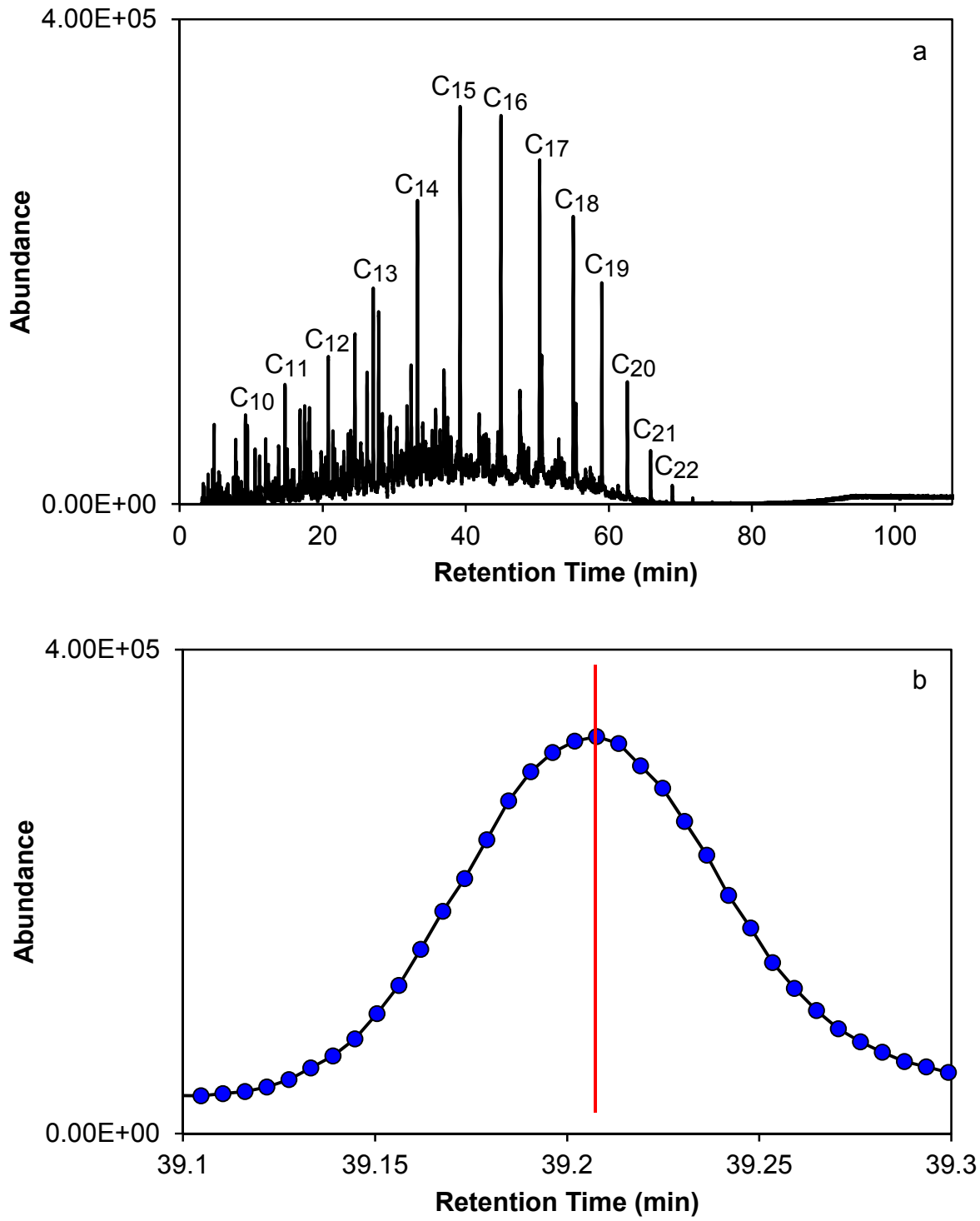


Figure 1-3. A representative chromatogram of diesel fuel with the normal alkane peaks labeled (a) and an expanded region of the pentadecane (C_{15}) peak (b). The blue circles indicate points where mass spectra were collected. The red line indicates the point which the mass spectrum in Figure 1-4 was taken.

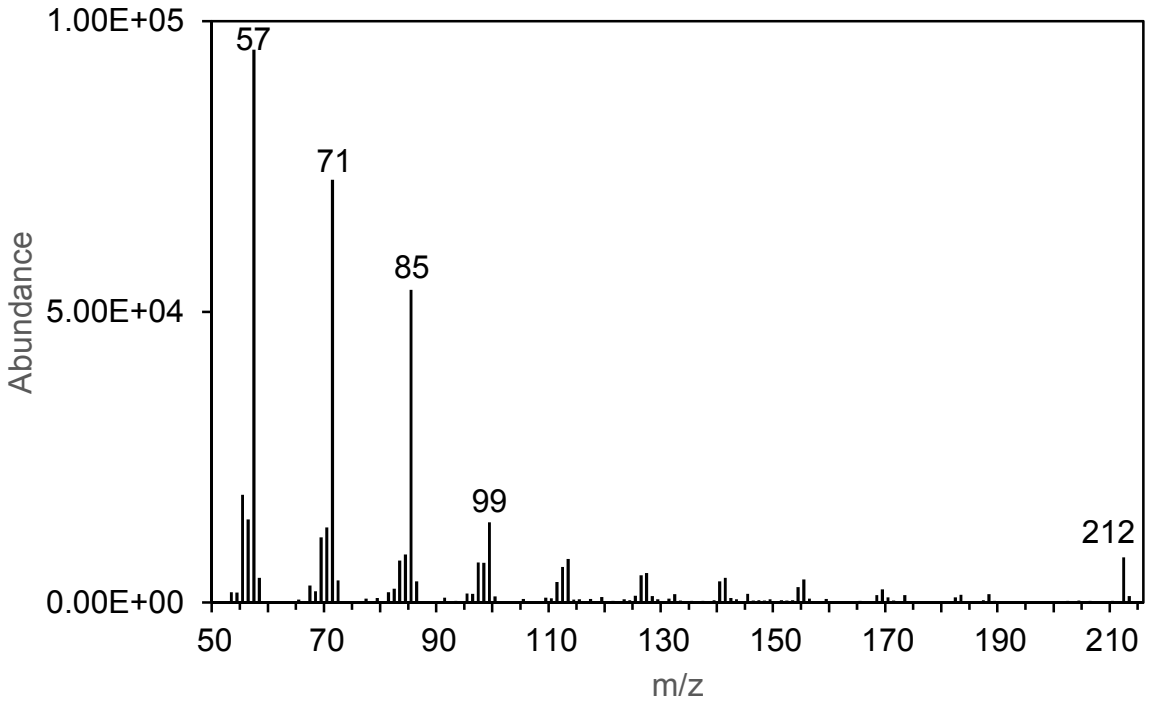


Figure 1-4. The mass spectrum (scan at retention time = 39.219 min, indicated by the red line in Figure 1-3b) of pentadecane (molecular weight = 212 amu).

1.1.3. Data Pretreatment Procedures Applied to Chromatographic Data

The signals in the GC-MS chromatogram can vary from one analysis to the next. Data pretreatment procedures are often employed to correct for these non-chemical variations in chromatographic data [9, 23, 24]. Differences in overall abundance between chromatograms analyzed on the same instrument can result from variation in sample preparation, sample injection, chromatographic conditions, and instrument response [23, 25]. In order to correct this problem, normalization procedures are commonly applied. Another source of variation that occurs in the baseline at the end of the chromatogram, particularly at high temperature, is from the degradation of compounds in the septum, injection port, and GC stationary phase. Differences in stationary phase age and wear can lead to a rise and variation in the signal over time [23]. It is often necessary to apply baseline correction procedures to minimize this variation. Noise, the high-frequency fluctuation in the signal, is another source of variation, which is hard to identify visually when there is a high signal-to-noise ratio. Noise is often the result of instrumental and electronic variation, and can be corrected using a smoothing procedure [23]. Peaks in the chromatogram can elute at slightly different retention times, due to instrumental variation in flow rates, column degradation, and manual injection procedures. These variations can be corrected by applying a retention time alignment algorithm [8, 23].

1.2. Statistical Analysis of Chromatographic Data

After generating the chemical fingerprint, multivariate statistical procedures can be utilized to compare many samples simultaneously. Multivariate statistical procedures are widely used as a research tool and have been applied to a variety of complex samples to help reveal underlying patterns in the data, including applications in lipidomics, metabolomics, proteomics, and petroleomics [12, 25-35]. The most commonly applied multivariate statistical procedure is principal component analysis (PCA). PCA serves as the basis of many other multivariate statistical procedures [36, 37]. Often data pretreatment procedures are applied prior to any multivariate statistical analysis to minimize non-chemical sources of variation.

1.2.1. Principal Component Analysis

PCA is an unsupervised multivariate statistical procedure that helps to identify underlying relationships within complex datasets without any prior knowledge about the data [37]. Often, PCA can identify small differences between samples, which can be over-looked by simple visual inspection of the data [8, 24]. In PCA, latent variables are used to reduce the dimensionality of the data, allowing for the visualization of relationships between samples [9, 36, 38]. The main outputs from PCA are scores plots, which show relationships between samples, and loadings plots, which show the importance of each variable.

In PCA, variables that vary together (covariance) are identified and grouped using eigenanalysis. These groups are identified as the principal components (PCs), which are orthogonal and uncorrelated linear combinations of the original dataset. From the

covariance matrix, the eigenvector and eigenvalue are calculated. The eigenvector for each PC contains the weights of each variable that define that PC, while the eigenvalue is a measure of the amount variance a particular eigenvector describes [9]. The eigenvector with the highest eigenvalue is the first PC. The eigenvector is then multiplied by the mean-centered data in order to obtain the scores for the samples [9, 13, 23, 36, 37]. In this work, the entire chromatogram of each sample was utilized, so each retention time point serves as a variable.

The scores for each sample on the first few PCs can be plotted, which allows for the visualization of clustering patterns. These scores plots (Figure 1-5a), which are a projection of the data onto a lower dimensional space, can then be used to infer relationships among samples [23]. Samples positioned close together in the new PC space are more similar and are associated, while samples that are positioned further apart are different and discriminated from one another [23]. Generally, the chemical differences between samples provide the greatest sources of variance [37]. However, when PCA is applied to chemically similar and highly complex samples, non-chemical sources of variation tend to be the greatest sources of variation.

The eigenvector, or weight, for each PC can also be plotted against each variable resulting in a loadings plot (Figure 1-5b) [23]. Variables with the highest or lowest weightings contribute the most to the positioning of the samples on the scores plot [23, 39]. Loadings plots can be used to infer which variables in the sample are changing or differing among the samples, as these variables will be given the most weight [6].

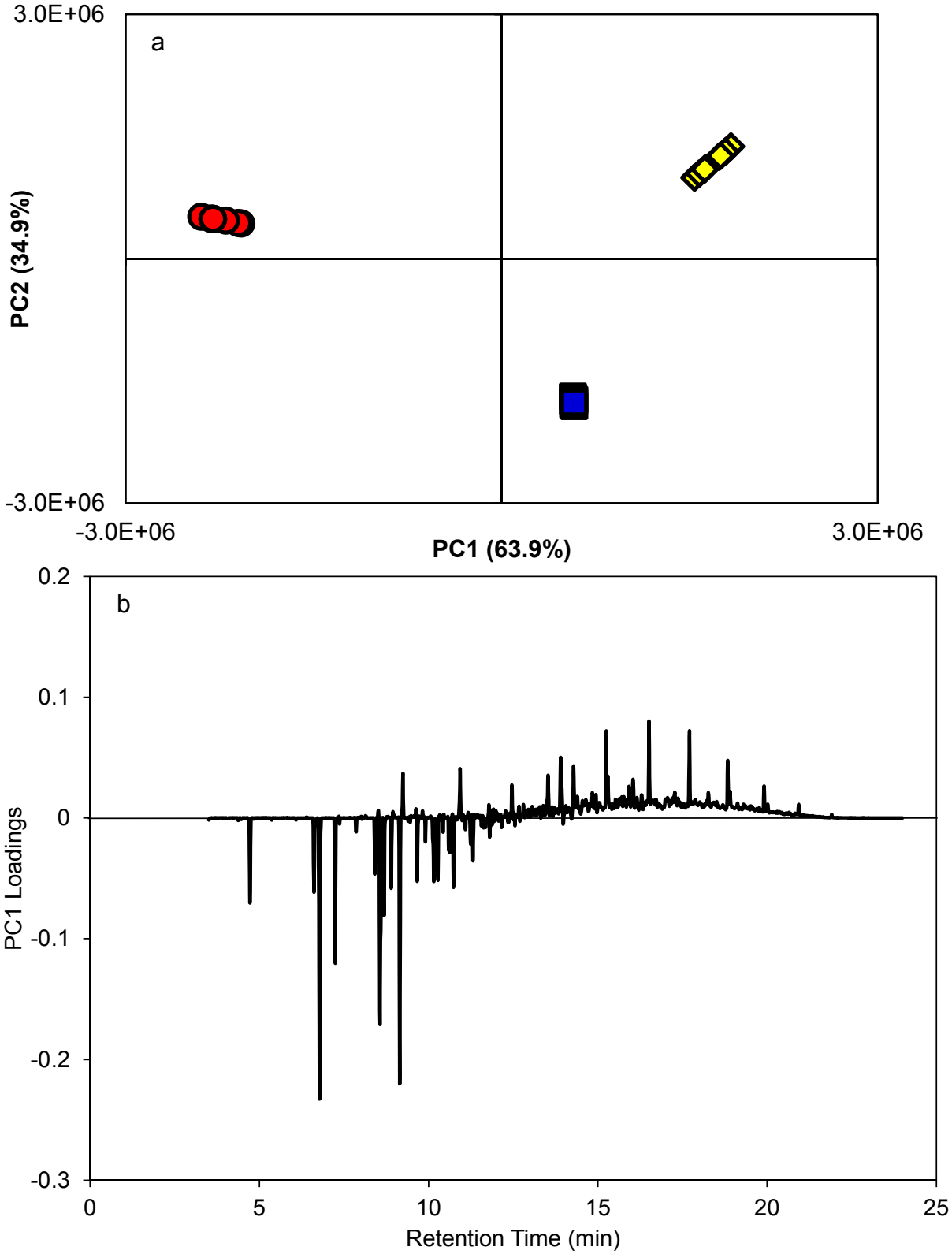


Figure 1-5. PCA scores plot (a) and loadings plot (b) of diesel fuel (yellow), lighter fluid (blue), and gasoline (red) shown in Figure 1-2.

The PCA plots shown in Figure 1-5 were performed using diesel fuel, lighter fluid, and gasoline, each extracted in triplicate with each extract analyzed in triplicate, creating nine total samples for each fuel (Figure 1-2). On the PCA scores plot (Figure 1-5a) replicates are positioned close together, while different fuel samples are positioned further apart from one another. This demonstrates that chemically similar samples are clustered (*i.e.* the replicates), while chemically different samples are positioned further apart (*i.e.* the different fuels).

The loadings plot in Figure 1-5b can be used to identify the variables that are differentiating the samples on the scores plot. For example, many of the compounds found in gasoline (Figure 1-2) are loading negatively on the PC1 loadings plot. This explains why gasoline is positioned negatively on PC1 in the scores plot (Figure 1-5a). Many of the compounds found in diesel fuel (Figure 1-2) are loading positively on PC1, explaining why diesel fuel is positioned positively on PC1 in the scores plot. Similar logic can be used to explain the position of the lighter fluid and all of the samples on PC2.

1.2.2. Application of PCA to Differentiate Complex Samples

There are many areas of research in which PCA and other multivariate statistical procedures are used to differentiate complex samples. As an example, an on-going project in our lab focuses on the use of PCA to associate fire debris with a corresponding ignitable liquid reference standard for use in fire debris analysis [5, 7, 40-42]. Hupp *et al*. demonstrated that TICs and EICs were useful in differentiating diesel fuels from different service stations using PCA, after applying alignment and normalization [7]. However, in this work, distinguishing between chemical variation and non-chemical variation was

challenging because no replicates were used in the PCA. This also made evaluating the improvements from data pretreatment very challenging.

Marshall *et al*. utilized the TIC and extracted ion profiles (EIPs) for the differentiation of five diesel fuels, analyzed in triplicate [5]. EIPs are the sum of several extracted ion chromatograms (EICs) that are characteristic for a compound class. EICs are the plot of the abundance of a single *m/z* at each retention time. In addition, Marshall investigated association of a diesel residue, extracted from a cloth matrix, to the neat diesel using PCA. Marshall showed that replicates of diesel samples could be clustered, but only after retention time alignment and normalization. However, association of the diesel residue to the neat liquid was not possible and the authors suggested that additional data pretreatment procedures would be necessary to further minimize non-chemical sources of variation [5]. In addition, this work demonstrated strategies for identifying retention time misalignments, including a characteristic derivative-shaped peak in the PCA loadings plot.

Baerncopf *et al*. used PCA to differentiate replicate GC-MS chromatograms of six different ignitable liquids (gasoline, diesel, lamp oil, adhesive remover, torch fuel, and paint thinner). By applying only retention time alignment and normalization, chromatograms of residues of each liquid were associated to the neat fuel after being spiked onto carpet and burned, which simulated burning at an arson scene [40]. This shows promise for associating fire debris to a neat source. However, the fuels used in this study are very different in chemical composition, making the chemical variation much greater than the non-chemical variation in the data.

Prather *et al*. investigated the effects of weathering the fuel in addition to matrix interferences on association of burned fire debris samples to a neat liquid, again using PCA [42]. Unevaporated and evaporated samples of gasoline and kerosene were spiked onto a carpet matrix and burned, then the residues were analyzed by GC-MS. After alignment and normalization, the simulated fire debris was associated with the correct ignitable liquid, even in the presence of interference compounds. However association to the correct extent of evaporation was not possible. The authors suggested that a larger dataset with fuels from different chemical classes would be necessary to test the robustness of these procedures for forensic analyses [42].

1.3. Evaluation of Data Pretreatment to Enhance Multivariate Statistical Analysis

The previous work described here demonstrates the utility of applying multivariate statistics to forensic analyses. However, the authors all commented on the small size of the dataset and the need for more thorough investigation of data pretreatment procedures. When chromatograms are collected over a long period of time (several months), data pretreatment procedures become crucial because instrumental drift over time introduces more non-chemical variation. Therefore, it is important to have data pretreatment procedures that can minimize or eliminate these variations as well as having metrics that can be used to evaluate the effect of the applied pretreatment procedures.

Often, data pretreatment procedures are applied to data with little discussion of how the parameters were selected or evaluated [43, 44]. Selection of data pretreatment procedures are facilitated by understanding the sources of the signals that require correction. Incorrect selection of data pretreatment procedures can result in amplifying

small variations in the data and lead to erroneous results [45]. Therefore, data pretreatment cannot be carried out using a “black-box” approach; instead, care must be taken to understand how and the extent to which each data pretreatment procedure corrects the non-chemical sources of variation in the chromatogram [43]. It is critical that the original relationships between variables are preserved even after data pretreatment [46].

There have been many algorithms designed to minimize non-chemical sources of variation in chemical analyses and are too numerous to discuss here [47, 48]. However, in many cases, there is little quantitative assessment of the effect of each procedure on the original chromatographic data. Many of the comparisons are based on a visual examination of the pretreated chromatograms compared to the original chromatograms [25]. This is problematic when trying to optimize the data pretreatment procedure because visual comparison is time consuming and subjective [49].

Common metrics for monitoring data pretreatment are based on a measure of a correlation coefficient or variance, either between samples or among replicates [8, 50-54]. The extent of non-chemical variation can be compared by examining replicate injections of the same sample. As replicates are chemically the same, the only variation must be non-chemical, arising from fluctuations and variations in the instrument [27]. Chromatograms from different samples are also compared, but differences could arise from variation in chemical composition.

Pearson product-moment correlation (PPMC) coefficients, which measure correlation between two samples, have been applied evaluate the effect of data

pretreatment. PPMC coefficients (*r*) are calculated by dividing the covariance between two sets of variables (*x* and *y*) by the product of each variables standard deviation (equation 1) [55].

$$r = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad \text{Equation 1-1}$$

High positive correlation indicates that variables increase and decrease together [55, 56]. In the case of chromatographic data, a high correlation coefficient indicates that the variables, which for chromatographic data is the abundance at each retention time, increase and decrease together [51, 57]. This makes PPMC coefficients an effective metric for evaluating data pretreatment procedures, especially retention time alignment procedures [45, 50, 54, 58]. An increase in the PPMC coefficient is observed when alignment improves. However, due to the large number of data points in chromatographic data, PPMC coefficients can be insensitive to small changes in the chromatogram. In addition, correlation coefficients are unaffected by relative changes in magnitude of the variables, and therefore, could not be utilized to evaluate normalization [56].

Another common method for evaluating data pretreatment procedures is to compare sample or replicate variance before and after pretreatment [8, 50, 53]. The variance (s^2) is calculated using the sum of the squared differences between two sets of values divided and the number of observations (*n*) using equation 2 [55].

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \qquad \text{Equation 1-2}$$

Another metric reported in the literature for evaluating data pretreatment is the standard deviation, or the square root of the variance, which measures deviation from the mean [57]. A smaller variance or standard deviation indicates less variation between samples [55]. In most cases, the variance is calculated based on selected features from the chromatogram, such as retention time or peak height.

In the comparison of three retention time alignment algorithms, van Nederkassel *et al*. utilized PPMC coefficients and the standard deviation of selected peaks to optimize the alignment parameters [50]. Gong *et al*. utilized a correlation coefficient and similarity index to compare aligned chromatograms to a target chromatogram in order to evaluate alignment [54]. Johnson *et al*. employed the average PPMC coefficient of all chromatograms and the standard deviations of selected peaks to optimize alignment parameters [51]. PPMC coefficients are the metric utilized in the correlation optimized warping (COW) alignment algorithm, which aligns chromatograms by maximizing the correlation between a sample and reference chromatogram [58]. Malmquist and Danielsson evaluated a single alignment and normalization using the residual sum of squares between each chromatogram before and after data pretreatment and the average chromatogram [8]. While the authors differ on the “best” alignment algorithm to use, it is generally agreed that alignment is necessary when chromatograms have been collected over a long period of time.

The ratio of the noise in a smoothed verses unsmoothed peak was used to compare parameters of the Savitzky-Golay smoothing algorithm [53, 59] and to compare the result of different smoothing algorithms [52]. In addition, the residual sum of squares between smoothed and unsmoothed voltammograms was utilized by Jakubowska and

Kubiak to evaluate distortion of the signal [53]. The precision of the peak area and height as well as the limit of detection have also been used to compare the effect of different smoothing algorithms, without regard to any peak distortion that may have occurred [60].

In addition to evaluating the raw data, the scores and loadings plots after PCA can also be used to evaluate the effect of each pretreatment procedure. Visual assessment of the clusters on the scores plot is a common method to evaluate data pretreatment procedures [8, 24, 45, 51]. However, visual assessment provides limited quantifiable information. Moreda-Pineiro *et al*. suggested the use of the percent variance accounted for uisng the first three PCs as a method for evaluating data pretreatment [44]. This method is an indirect measure of the association and discrimination of samples and is highly susceptible to influence by outliers. Degree-of-class-separation has been utilized for evaluating the clustering of samples on a scores plot based on a distance between clusters and the distance between samples within each cluster [49].

Despite the critical need for data pretreatment procedures, there has been no direct comparison of applying sequential data pretreatment procedures. When data pretreatment procedures are compared, there typically is not a quantitative comparison because visual examination of the PCA scores plot is used rather than a metric. The development of metrics for the comparison of these data pretreatment procedures would allow for parameter optimization and a means to evaluate the effectiveness of each pretreatment.

1.4. Research Objective

Multivariate statistical procedures are widely utilized in chemical analyses and show great promise for applications in forensic science. However, additional research is still required to develop appropriate methodologies for forensic applications. As previously demonstrated, data pretreatment is a critical aspect of successfully applying multivariate statistical procedures capable of identifying minute differences in the chemical fingerprints of forensic evidence.

The goal in this work is to develop methods for evaluating and optimizing data pretreatment procedures in order to minimize non-chemical sources of variation, resulting in enhanced discrimination of complex samples using multivariate statistical analysis. The goal is not to compare every possible method of data pretreatment, but rather to provide a general overview of common pretreatment procedures and to provide a uniform set of metrics for evaluating these procedures, using both the raw data and the resulting PCA scores and loadings plots. In order to attain this goal, the following aims were outlined:

- Demonstrate methods for objectively selecting and optimizing different data pretreatment methods and associated parameters.

- Develop metrics for evaluating the effect of data pretreatment on the chromatographic data and the PCA results of chemically complex and highly similar samples.

- Demonstrate, using proper data pretreatment procedures, that non-chemical variation in chromatograms can be minimized without altering the discriminatory chemical information.

1.5. Impact on the Forensic Science Community

Multivariate statistics have not been widely applied in a legal setting. In order to be accepted in court, these statistical procedures must pass a Frye or Daubert standard [1]. As part of the basis for meeting these standards, it will be critical to demonstrate that these statistical procedures and the data pretreatment that accompanies them, do not change the fundamental chemical information in the analyses. This research aids in that goal by providing a fundamental understanding of the data pretreatments and metrics to evaluate their effectiveness.

REFERENCES

REFERENCES

[1] R. Saferstein, Criminalistics: An Introduction to Forensic Science, Prentice Hall, Upper Saddle River, NJ, 2004.

[2] S. Bell, Forensic Chemistry, Prentice Hall, Upper Saddle River, NJ, 2006.

[3] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, 2009.

[4] J.D. DeHaan, Kirk's Fire Investigation, Prentice Hall, Upper Saddle River, NJ, 2002.

[5] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith, Anal. Bioanal. Chem., 394 (2009) 2049.

[6] R.G. Brereton, Applied Chemometrics for Scientists, John Wiley & Sons, Hoboken, NJ, 2007.

[7] A.M. Hupp, L.J. Marshall, D.I. Campbell, R.W. Smith, V.L. McGuffin, Anal. Chim. Acta, 606 (2008) 159.

[8] G. Malmquist, R. Danielsson, J. Chromatogr. A, 687 (1994) 71.

[9] S.L. Morgan, E.G. Bartick, in: R.D. Blackledge (Ed.), Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis, John Wiley & Sons, Inc. , Hoboken, NJ, 2007.

[10] M.J. Adams, Chemometrics in Analytical Spectroscopy, Royal Society of Chemistry, Victoria, Australia, 1995.

[11] D.R. Burgard, J.T. Kuznicki, Chemometrics: Chemical and Sensory Data, CRC Press, Boca Raton, FL, 1990.

[12] Z. Wang, J.H. Christensen, Crude Oil and Refined Product Fingerprinting: Applications (Chapter 17), Elsevier, Burlington, MA, 2006.

[13] S.M. Mudge, Environ. Forensics, 8 (2007) 155.

[14] J. Gonzalez-Rodriguez, G. Fowler, Forensic Sci. Int., 231 (2013) 6.

[15] N.V.S. Rodrigues, E.M. Cardoso, M.V.O. Andrade, C.L. Donnici, M.M. Sena, J. Braz. Chem. Soc., 24 (2013) 507.

[16] C. Muehlethaler, G. Massonnet, P. Esseiva, Forensic Sci. Int., 209 (2011) 173.

[17] M. Monfreda, A. Gregori, J. Forensic Sci., 56 (2011) 372.

[18] R.J.H. Waddell-Smith, J. Forensic Sci., 52 (2007) 1297.

[19] D.A. Skoog, F.J. Holler, S.R. Crouch, Principles of Instrumental Analysis, Thomson Brooks/Cole, Belmont, CA, 2007.

[20] M.C. McMaster, GC/MS: A Practical User's Guide John Wiley & Sons, Inc, Hoboken, NJ, 2008.

[21] P.J. Marriott, in: E. Heftmann (Ed.), Chromatography, Elsevier, New York, NY, 2004.

[22] E. de Hoffmann, V. Stroobant, Mass Spectrometry Principles and Applications, John Wiley & Sons, Hoboken, NJ, 2007.

[23] K.M. Pierce, J.S. Nadeau, R.E. Synovec, in: C.F. Poole (Ed.), Gas Chromatography, Elsevier, Waltham, MA, 2012.

[24] M.E. Pate, N.F. Thornhill, R. Chandwani, M. Hoare, N.J. Titchener-Hooker, Bioprocess Eng., 19 (1998) 297.

[25] J.H. Christensen, G. Tomasi, J. Chromatogr. A, 1169 (2007) 1.

[26] J.H. Christensen, A.B. Hansen, U. Karlson, J. Mortensen, O. Andersen, J. Chromatogr. A, 1090 (2005) 133.

[27] J.H. Christensen, J. Mortensen, A.B. Hansen, O. Andersen, J. Chromatogr. A, 1062 (2005) 113.

[28] J.H. Christensen, G. Tomasi, A.B. Hansen, Environ. Sci. Technol., 39 (2005) 255.

[29] L.M.V. Malmquist, R.R. Olsen, A.B. Hansen, O. Andersen, J.H. Christensen, J. Chromatogr. A, 1164 (2007) 262.

[30] N.J. Nielsen, D. Ballabio, G. Tomasi, R. Todeschini, J.H. Christensen, J. Chromatogr. A, 1238 (2012) 121.

[31] M. Steinfath, D. Groth, J. Lisec, J. Selbig, Physiol. Plant., 132 (2008) 150.

[32] J. van der Greef, H. van Wietmarschen, B. van Ommen, E. Verheij, Mass Spectrom. Rev., 32 (2013) 399.

[33] M. Chadeau-Hyam, G. Campanella, T. Jombart, L. Bottolo, L. Portengen, P. Vineis, B. Liquet, R.C.H. Vermeulen, Environ. Mol. Mutagen., 54 (2013) 542.

[34] J. Trygg, E. Holmes, T. Lundstedt, J. Proteome Res., 6 (2007) 469.

[35] D.I. Ellis, R. Goodacre, Analyst, 131 (2006) 875.

[36] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, New York, NY, 2009.

[37] S. Wold, K. Esbensen, P. Geladi, Chemometrics Intell. Lab. Syst., 2 (1987) 37.

[38] P. Gemperline (Ed.), Practical Guide to Chemometrics, CRC Press, Boca Raton, FL, 2006.

[39] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, Inc., New York, NY, 1998.

[40] J.M. Baerncopf, V.L. McGuffin, R.W. Smith, J. Forensic Sci., 56 (2011) 70.

[41] J.M. Baerncopf, V.L. McGuffin, R.W. Smith, J. Forensic Sci., 55 (2010) 185.

[42] K.R. Prather, V.L. McGuffin, R.W. Smith, Forensic Sci. Int., 222 (2012) 242.

[43] M. Daszykowski, B. Walczak, Trac-Trends Anal. Chem., 25 (2006) 1081.

[44] A. Moreda-Pineiro, A. Marcos, A. Fisher, S.J. Hill, J. Environ. Monit., 3 (2001) 352.

[45] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr., 18 (2004) 231.

[46] O.M. Kvalheim, F. Brakstad, Y.Z. Liang, Anal. Chem., 66 (1994) 43.

[47] S.D. Brown, S.T. Sum, F. Despagne, B.K. Lavine, Anal. Chem., 68 (1996) R21.

[48] M. Katajamaa, M. Oresic, J. Chromatogr. A, 1158 (2007) 318.

[49] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A, 1096 (2005) 101.

[50] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y.V. Heyden, J. Chromatogr. A, 1118 (2006) 199.

[51] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A, 996 (2003) 141.

[52] P. Barak, Anal. Chem., 67 (1995) 2758.

[53] M. Jakubowska, W.W. Kubiak, Anal. Chim. Acta, 512 (2004) 241.

[54] F. Gong, B.T. Wang, F.T. Chau, Y.Z. Liang, Anal. Lett., 38 (2005) 2475.

[55] J.L. Devore, Probability and Statistics for Engineering and the Sciences, Duxbury Press, Belmont, CA, 1991.

[56] J.H. Zar, Biostatistical Analysis, Prentice-Hall, Upper Saddle River, NJ, 1999.

[57] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, Pearson, New York, NY, 2000.

[58] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A, 805 (1998) 17.

[59] C.G. Enke, T.A. Nieman, Anal. Chem., 48 (1976) A705.

[60] C.R. Mittermayr, H. Frischenschlager, E. Rosenberg, M. Grasserbauer, Fresenius J. Anal. Chem., 358 (1997) 456.

CHAPTER 2: INITIAL ANALYSIS OF DIESEL SAMPLES

2.1. Introduction

The focus of this work is to evaluate strategies for enhancing the differentiation of complex and chemically similar samples for multivariate statistical analysis. Many statistical procedures have been applied to forensically relevant data, with the goal of implementing those procedures into forensic laboratories to assist in comparisons and assigning a statistical confidence to the forensic analysis [1-5]. One of the most commonly utilized multivariate statistical procedure is principal component analysis (PCA), which discriminates samples based on the greatest sources of variance within the dataset.

Typical data generated in forensic laboratories, including chromatograms of fire debris from gas chromatography-mass spectrometry (GC-MS), are highly complex making differentiation challenging. For example, the chromatograms of different fuel samples can appear similar when compared visually. Often, minute differences, which are hard to find by eye, are necessary to distinguish between samples. Statistical procedures can be introduced to assist in the differentiation. Additionally, by utilizing statistical approaches to analyze forensic samples, there is less subjectivity and greater consistency across forensic laboratories. Moreover, the application of statistical analyses allow for comparisons with statistical confidence rather than simply the analyst's opinion.

In any chemical investigation, variation due to sample preparation and instrumental procedures is often introduced prior to the statistical analysis [6, 7]. When applying PCA to samples with different chemical composition, such as diesel and gasoline, the variation

introduced from sample preparation and analysis is small compared to the chemical differences between the samples. However, when utilizing PCA to differentiate highly similar samples, such as diesel samples from different sources that contain the many of the same compounds, non-chemical differences between analyses are often the largest source of variation, which can mask the chemical differences necessary for differentiation. Data pretreatment procedures are then applied to minimize these variations.

In this chapter, the challenges in differentiating complex and similar samples are demonstrated. First, thirty chromatograms of diesel fuel, a very complex sample, were generated. Then, PCA was applied to highlight the difficulties in discriminating these complex and similar samples. Subsequent chapters will focus on selecting pretreatment procedures, methods to evaluate and select appropriate parameters for each procedure, and the resulting effect of the procedure on the ability to discriminate chemically similar samples.

2.2. Selection of Samples

Diesel fuel was chosen as the model sample due to its complex chemical composition. Diesel fuel consists of small and relatively non-polar molecules, with boiling points ranging from approximately 80 - 350 °C, which is well suited for GC-MS analysis. Diesel fuel consists of hundreds of different chemical compounds, present at varying concentrations, which adds to the complexity of the sample. Diesel fuel is widely available from local service stations and can be used as an accelerant in arson, making diesel fuel forensically relevant. During an arson investigation it is necessary to identify whether an accelerant is present. With additional research, small differences between fuel samples

may provide statistical confidence in a comparison between an accelerant found at a crime scene and an accelerant found in the possession of a suspect. However, an accelerant found at a crime scene cannot be traced to a single service station or brand because diesel fuel found at different service stations (even different brands) are often purchased from the same refinery.

Ten different diesel fuels were collected from service stations in the Lansing, Michigan area during June 2007 and were stored in acid-washed amber bottles at 3 °C until analysis (Table 2-1). Prior to analysis, each sample was diluted 200:1 in dichloromethane (spectrophotometric grade, Sigma-Aldrich, St. Louis, MO), and then analyzed in triplicate using GC-MS, resulting in 30 chromatograms [1, 2, 8].

2.3. GC-MS Parameters

All analyses were performed on an Agilent 6890N gas chromatograph coupled to an Agilent 5975 mass spectrometer detector (Agilent Technologies, Santa Clara, CA). The GC was equipped with an HP-5MS capillary column with a 5% phenyl- 95% methyl-polysiloxane stationary phase (30 m x 0.25 mm x 0.25 μm, Agilent Technologies). Ultra-high purity helium was used as the carrier gas with a nominal flow rate of 1 mL/min. A manual injection with a 10 μL syringe (Hamilton, Reno, NV) was used to deliver 1 μL of diluted diesel with a split ratio of 50:1. A slow, two-step temperature ramp was used in order to maximize the chromatographic resolution. The oven temperature program was

Table 2-1. Diesel samples collected for this work, including the service station and the date of collection.

Sample Identifier	Service Station	Location	Date
Diesel 1	Sunoco	2139 Haslett Road East Lansing, MI	05/31/07
Diesel 2	Sunoco	3000 Dunckel Road Lansing, MI	06/06/07
Diesel 3	Meijer	550 Hull Road Mason, MI	06/07/07
Diesel 4	Meijer	2055 W. Grand River Ave Okemos, MI	06/11/07
Diesel 5	Mobil	1500 Haslett Road Haslett, MI	06/12/07
Diesel 6	Speedway	16819 Marsh Road Bath Township, MI	06/13/07
Diesel 7	Mobil	2704 Lake Lansing Road Lansing, MI	06/15/07
Diesel 8	Marathon	3010 W. Lake Lansing Road East Lansing, MI	06/18/07
Diesel 9	Marathon	401 S. Pennsylvania Ave Lansing, MI	06/19/07
Diesel 10	Speedway	1659 Grand River Avenue Okemos, MI	06/20/07

as follows: 50 °C to 150 °C at 2 °C/min, then 150 °C to 280 °C at 3 °C/min with a final hold of 15 min. The inlet and transfer line were maintained at 300 °C. The mass spectrometer utilized electron ionization (70 eV) with a quadrupole mass analyzer, which scanned mass-to-charge (*m/z*) ratios of 40-550 at a scan rate of 2.91 scans/s [1, 2, 8].

2.4. Visual Assessment of Diesel Chromatograms

After GC-MS analysis, the resulting chromatograms were exported from Chemstation (version E.01.01.335, Agilent Technologies, Santa Clara, CA) and regenerated in Excel (Office 2013, version 15.0, Microsoft Corporation, Redmond, WA). Representative chromatograms of all ten diesel samples are shown in Figure 2-1. The normal alkanes are labeled for reference. The same scale is utilized so that differences in overall abundance can be observed. The overall differences in abundance are likely not chemical, but rather a result of injecting slightly different volumes of sample.

Visual examination of the chromatograms indicates slight chemical differences among each diesel sample. There is a lower abundance of short-chain alkanes (C_9-C_{11}) in Diesels 1 and 2, relative to the other samples. A low abundance of short-chain alkanes is characteristic of summer diesel fuel. In order to increase the cloud point in the winter, diesel fuel is blended with kerosene or jet fuel, which increases the concentration of short-chain alkanes. Therefore, Diesels 1 and 2 are likely summer diesel fuels, due to their lower abundance of short-chain alkanes, while samples 3 - 10 are potentially winter diesel blends. Even though all samples were collected in June, the winter diesel fuel is likely left over from the winter and being distributed until depleted.

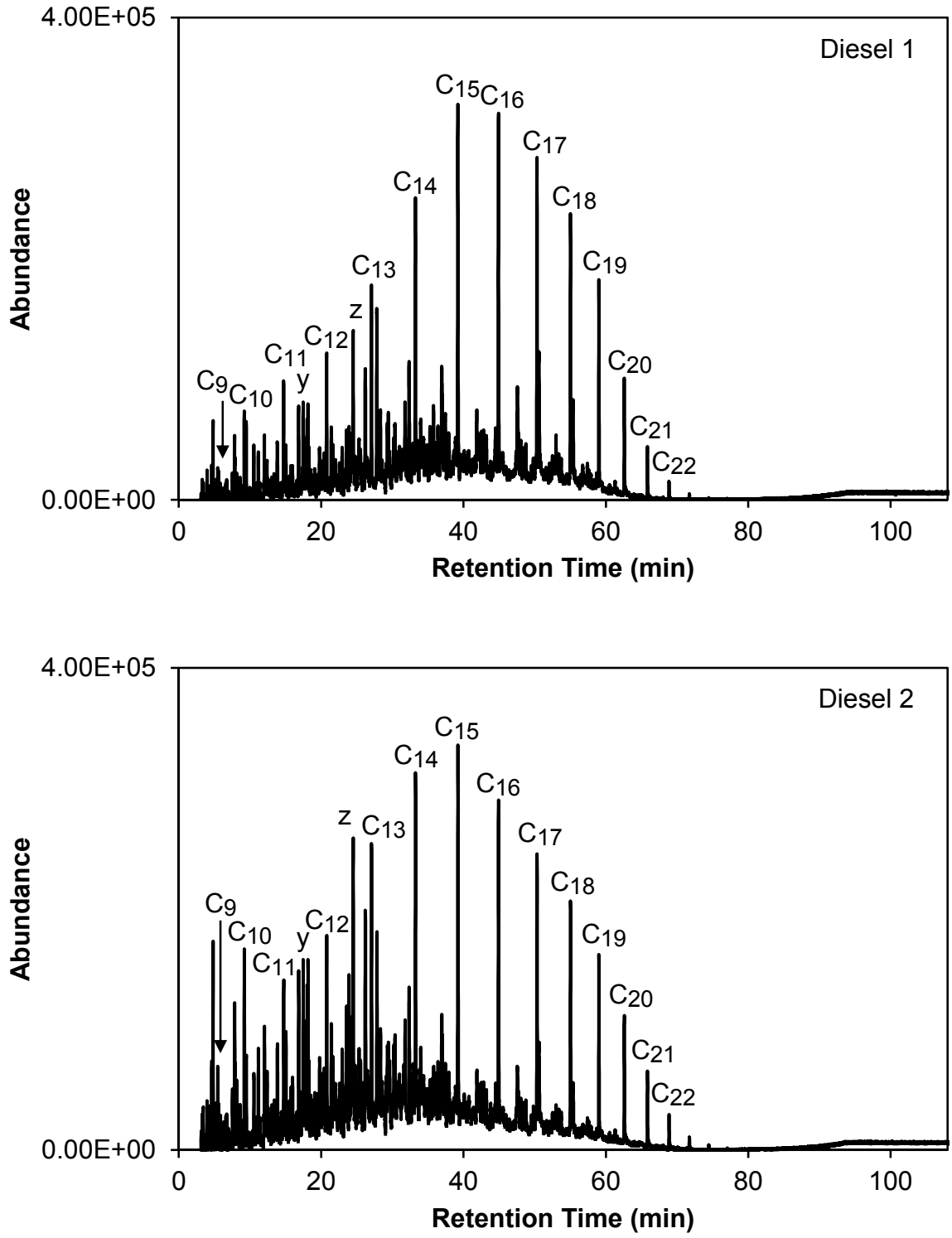
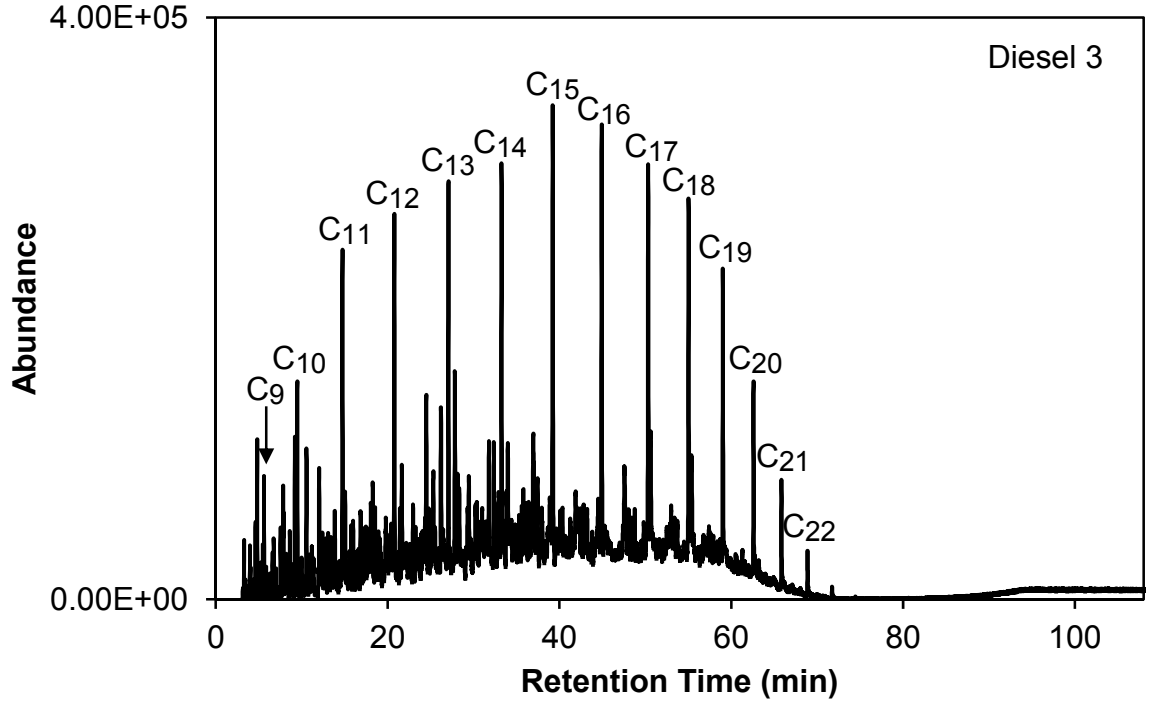


Figure 2-1. A representative diesel chromatogram of each diesel fuel sample (1 - 10) with the normal alkanes labeled. Octane was detected at low abundance, but was not labeled. Labels y and z are used to indicate two clusters of peaks from substituted aromatic compounds observed in diesel 1 and 2.

Figure 2-1 (cont'd).



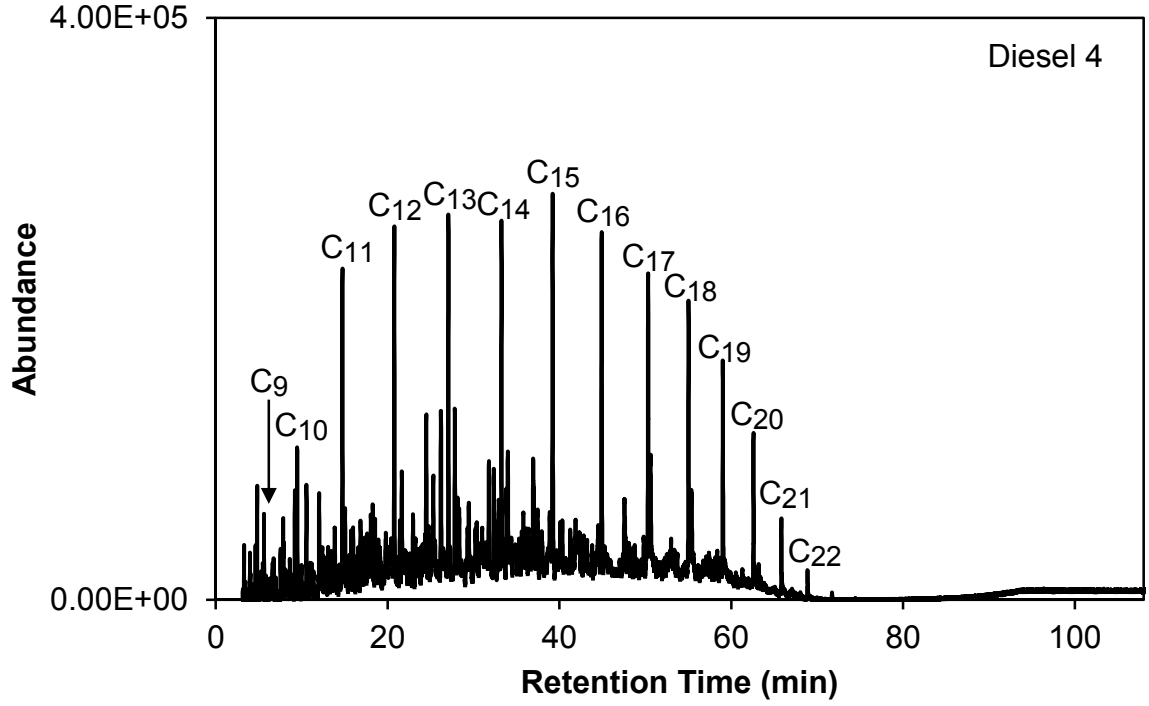
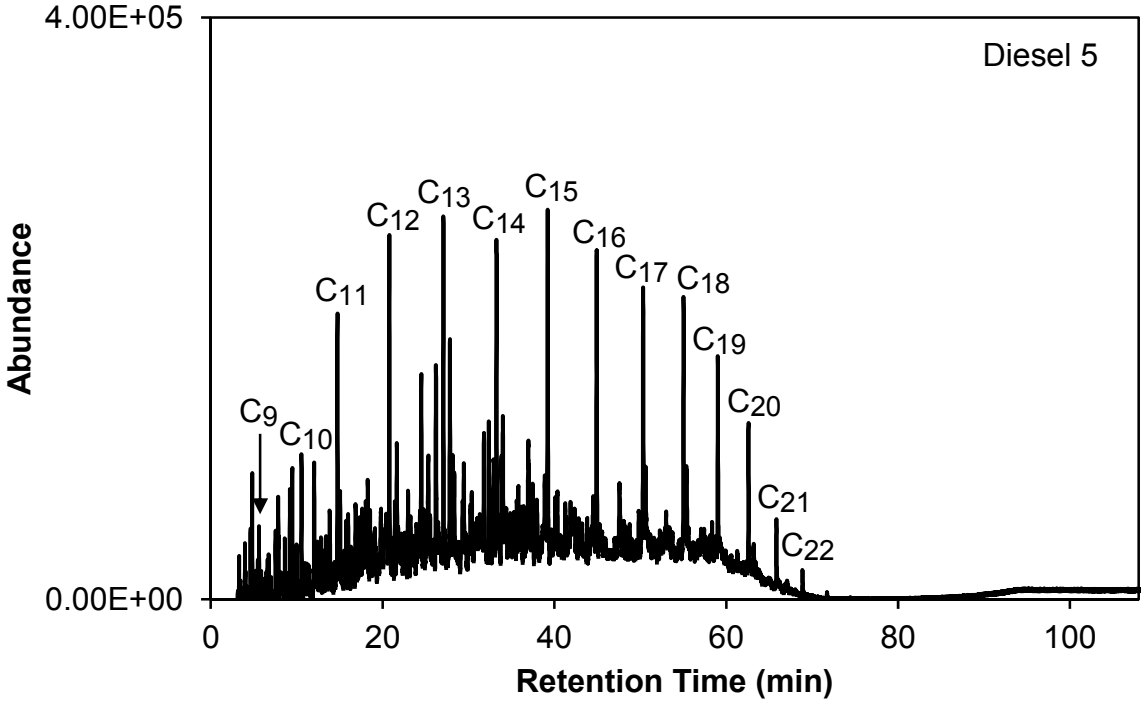


Figure 2-1 (cont'd).



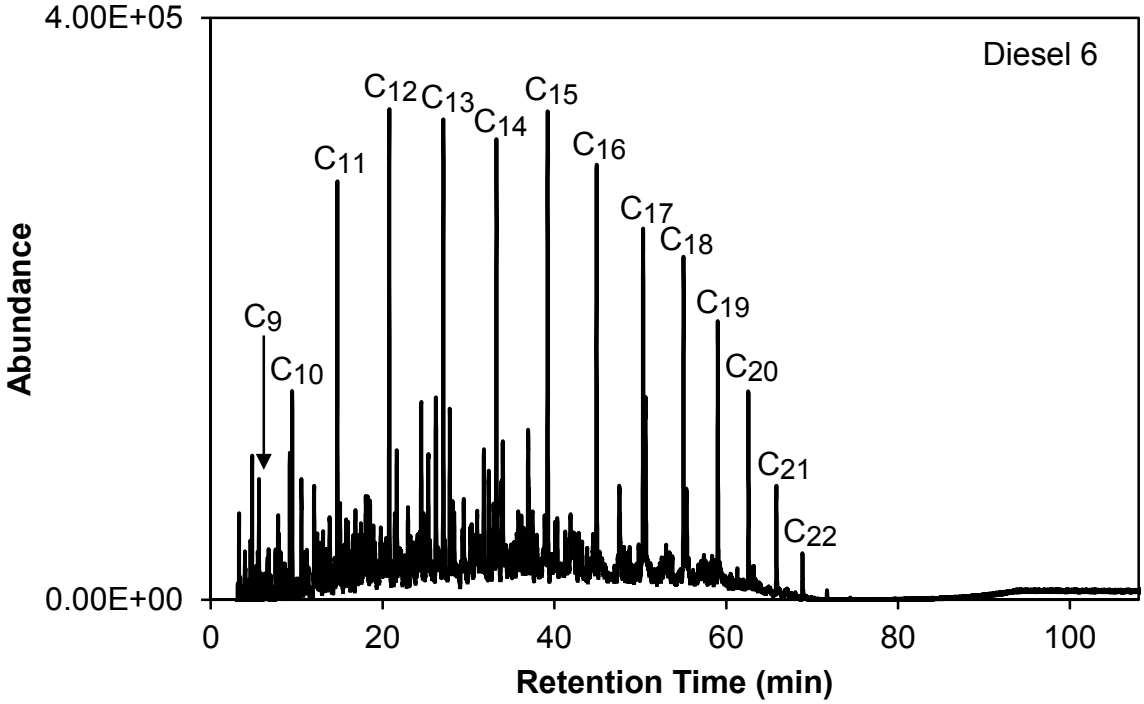
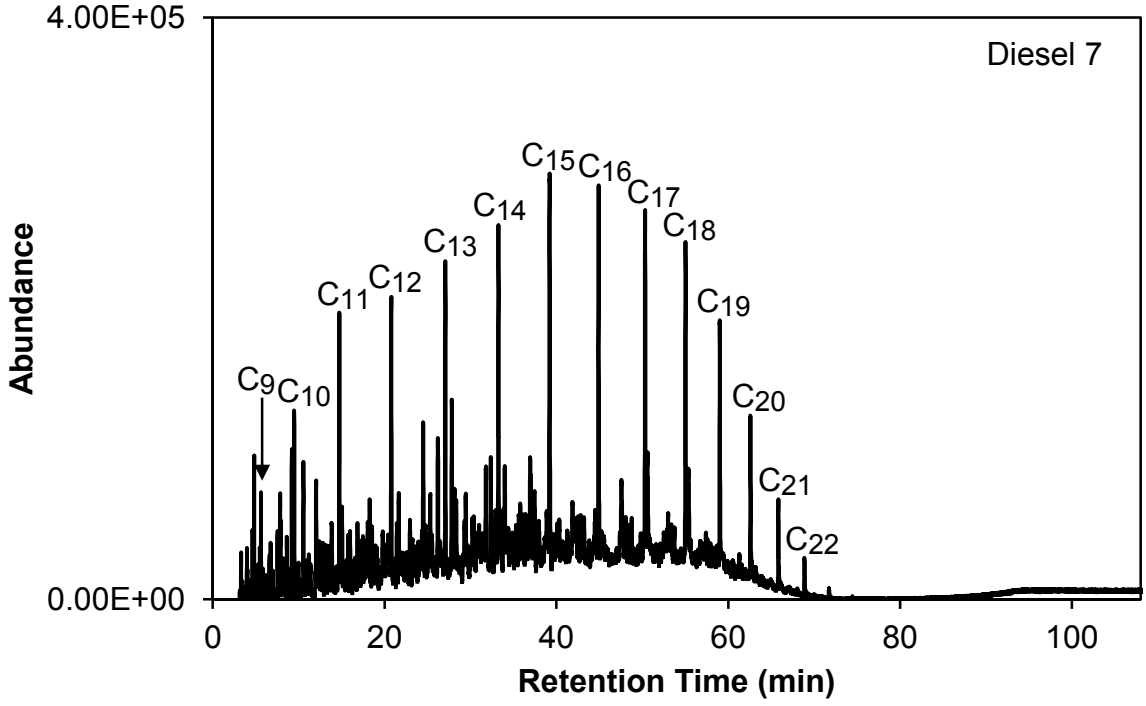


Figure 2-1 (cont'd).



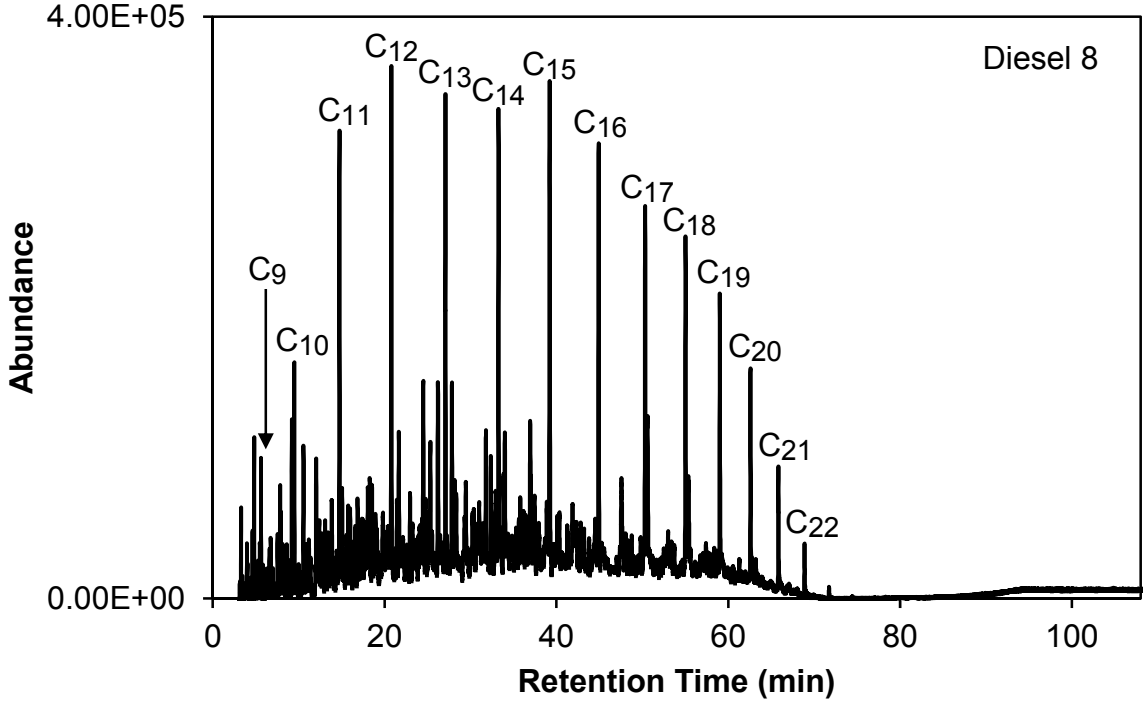
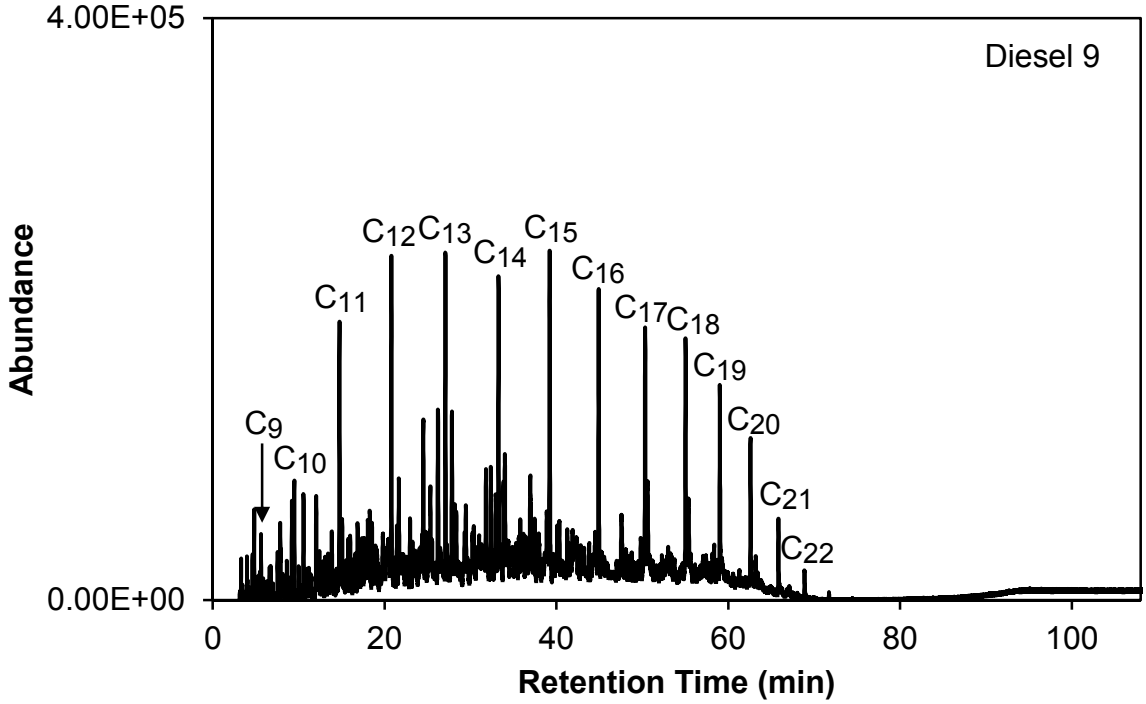
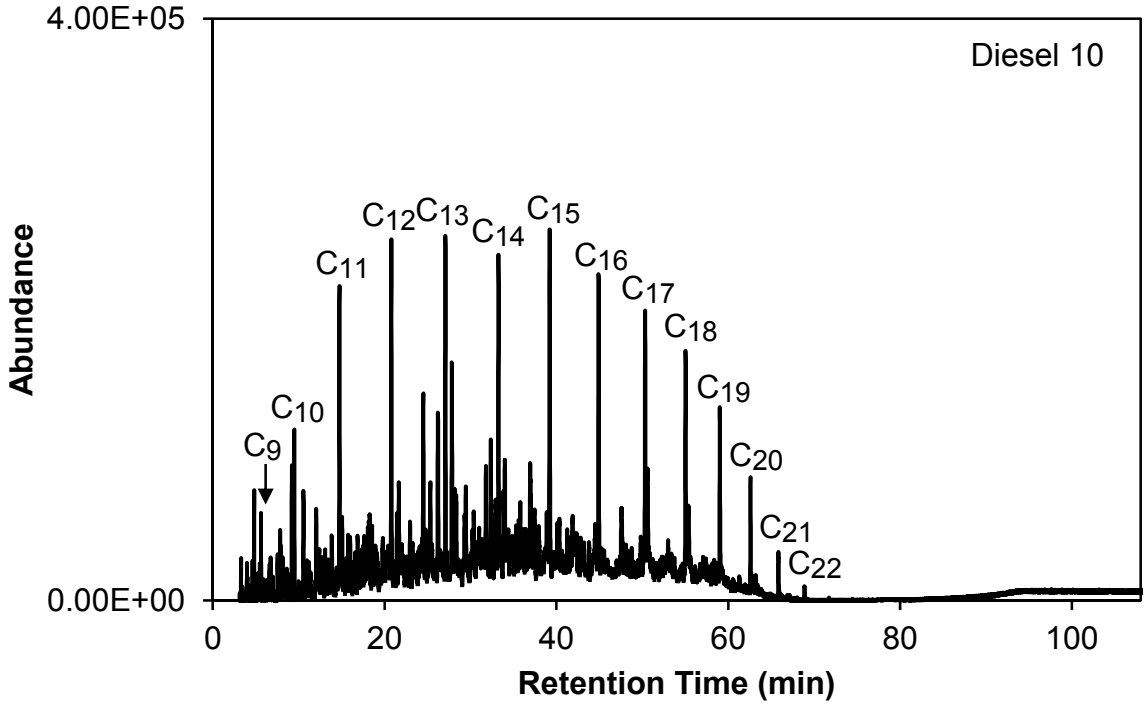


Figure 2-1 (cont'd).





Diesels 1 and 2 also have a higher abundance of two clusters of peaks, provisionally identified as branched and substituted aromatic compounds, labeled as y and z in Figure 2-1. Diesel 2 also has a higher abundance of more volatile (early eluting) aromatic compounds, which is not observed in any of the other samples. The origin of these differences is not known, but may be a result of the starting material or additive, as these compounds were only at higher abundance in samples from Sunoco Service stations (Table 2-1).

Three characteristic groups were observed within the provisionally identified winter diesel fuels (Figure 2-1, samples 3 - 10). Diesel samples 3 and 7 have a unimodal distribution of the normal alkanes, which maximizes at a retention time of approximately 40 minutes (C_{15}). Diesel samples 4, 5, 6, 8, 9, and 10 have a bimodal distribution. Diesel samples 6 and 8 maximize at retention times of approximately 20 and 40 minutes (C_{12} and C_{15}), while diesel samples 4, 5, 9, and 10 maximize at retention times of approximately 28 and 40 minutes (C_{13} and C_{15}). These differences are likely due to differences in the crude oil starting material, refining processes, and blending for each brand as well as the refinery from which the fuel was purchased.

Small chemical differences (such as those described above) that are observed through visual assessment are often overshadowed by non-chemical sources of variation when PCA is utilized. Examples of non-chemical variation are highlighted in Figure 2-2 and Figure 2-3, which shows an overlay of replicate chromatograms and representative chromatograms of Diesels 3 – 10, respectively. In the replicate chromatograms,

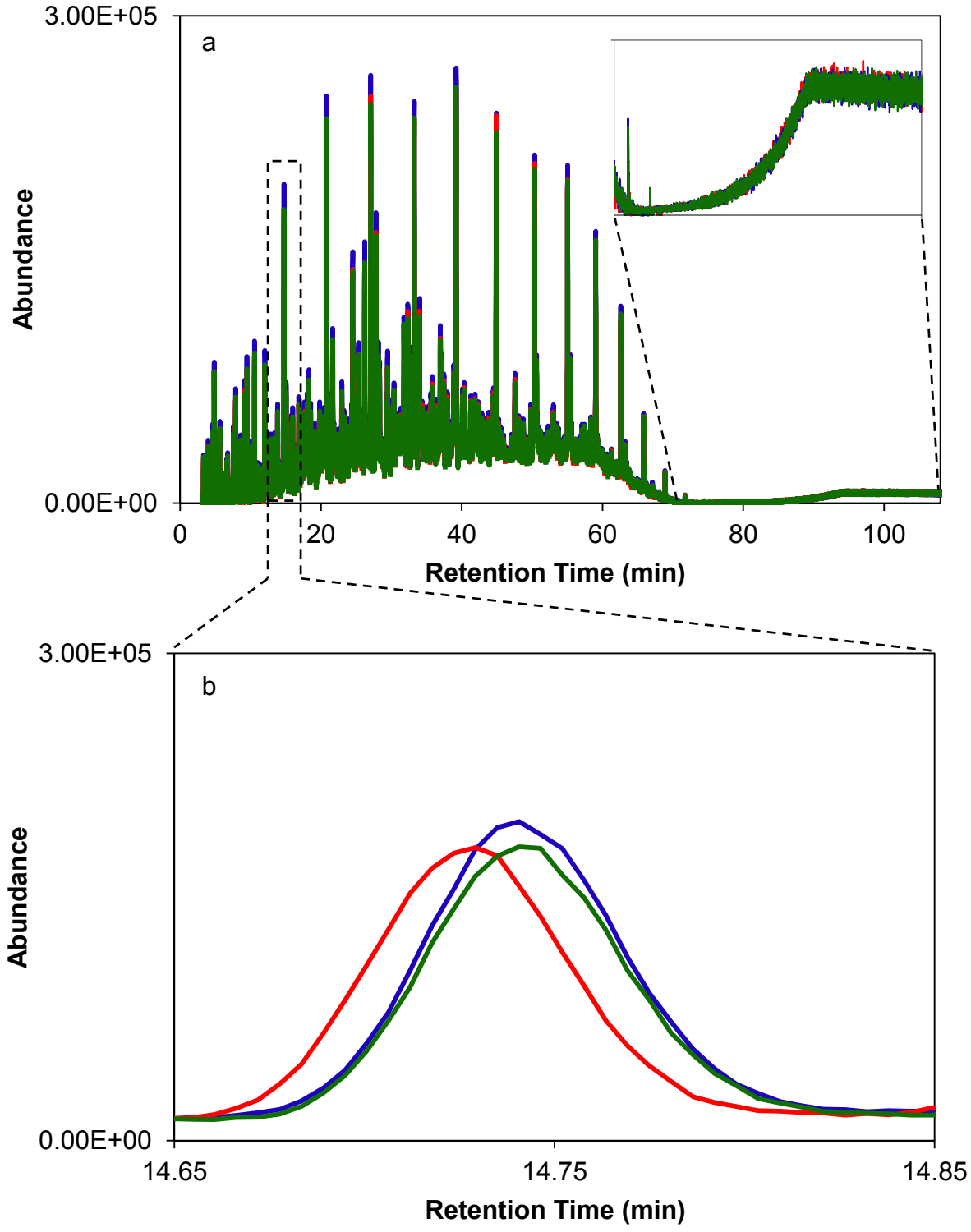


Figure 2-2. Chromatograms of three replicates of diesel 5 (a) and an expanded region of the chromatogram on the undecane peak (b). The inset shows the baseline at the end of the chromatogram.

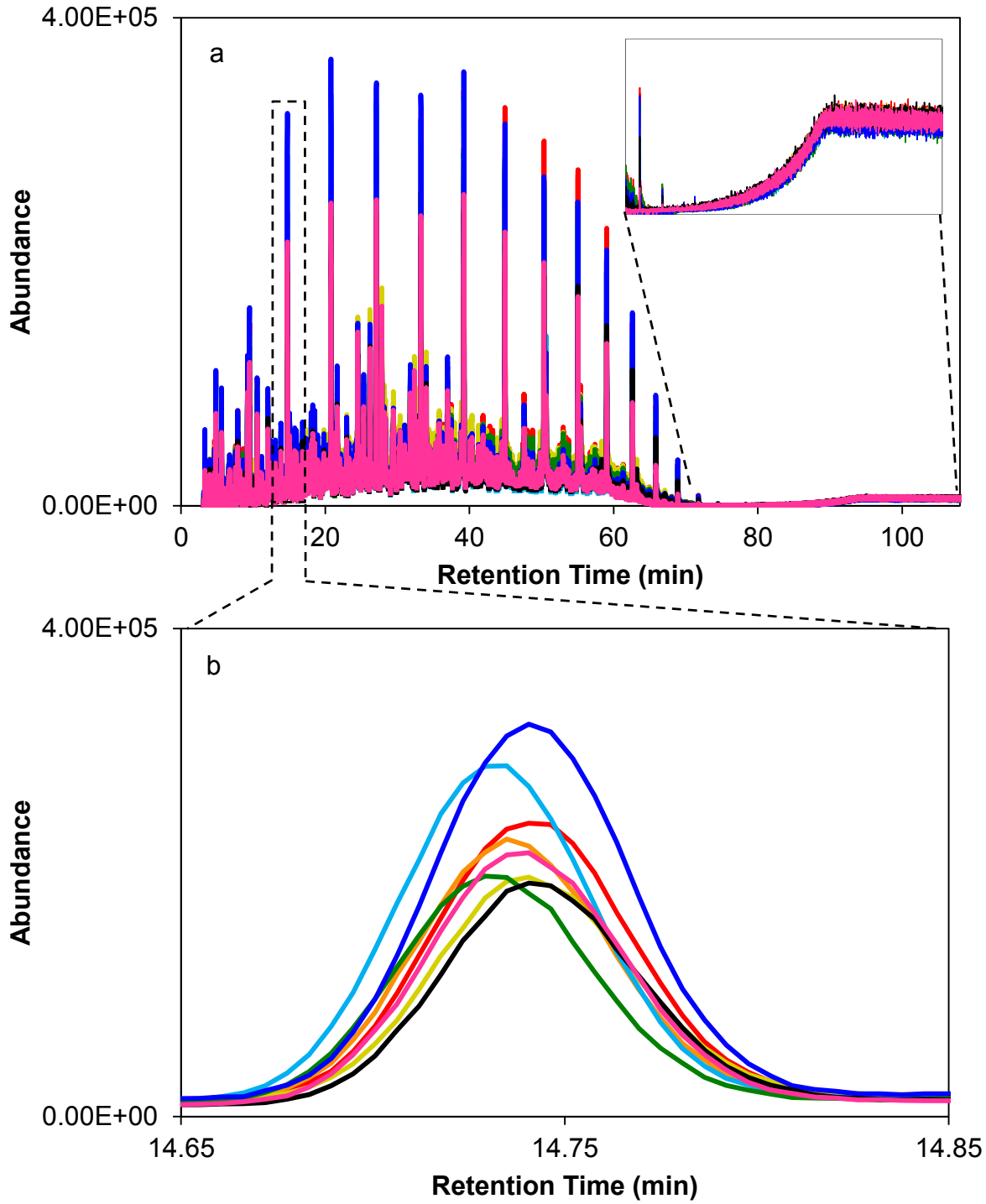


Figure 2-3. An overlay of one chromatogram from each of the eight diesel samples (a) and an expanded region of the chromatogram on the undecane peak (b). The insets in part a show the baseline at the end of the chromatogram. Each color represents a different diesel sample

differences in peak height are observed. In addition, there is also variation in the rise in the baseline and noise observed at the end of each chromatogram. Figure 2-2b and Figure 2-3b show an expanded view of the undecane peak (between 14.65 and 14.85 minutes) where misalignments are observed, both between replicates and between sample chromatograms.

2.5. PCA of Diesel Chromatograms

In this work, PCA was initially performed using three replicates of ten diesel chromatograms. Based on the chemical composition of each fuel in the chromatogram, three clusters are expected on the PCA scores plot. One cluster would contain the summer diesel fuels (Diesels 1 and 2). The second cluster would contain Diesels 3 and 7, which have the unimodal distribution of normal alkanes and the final cluster would contain the winter diesel samples (Diesels 4, 5, 6, 8, 9, and 10) that contain the bimodal distribution.

The scores plot obtained from the PCA of ten diesel samples is shown in Figure 2-4. The x-axis is the first principal component (PC1), while the y-axis is the second principal component (PC2). The number in parentheses indicates the percent variance for each principal component (47.1% by PC1 and 19.1% by PC2, 66.2% for both PCs). Replicates of each sample are not positioned close together, indicating that there are non-chemical sources of variation present. The only replicate samples positioned close together are those of Diesel 2 (grey 4-point stars). Two general clusters are observed, one with the summer diesels: (Diesels 1 and 2), and one with the winter diesels (Diesels 3 - 10). The other 8 diesels are intermingled, even though differences in the distribution of the normal alkanes were observed in the chromatograms.

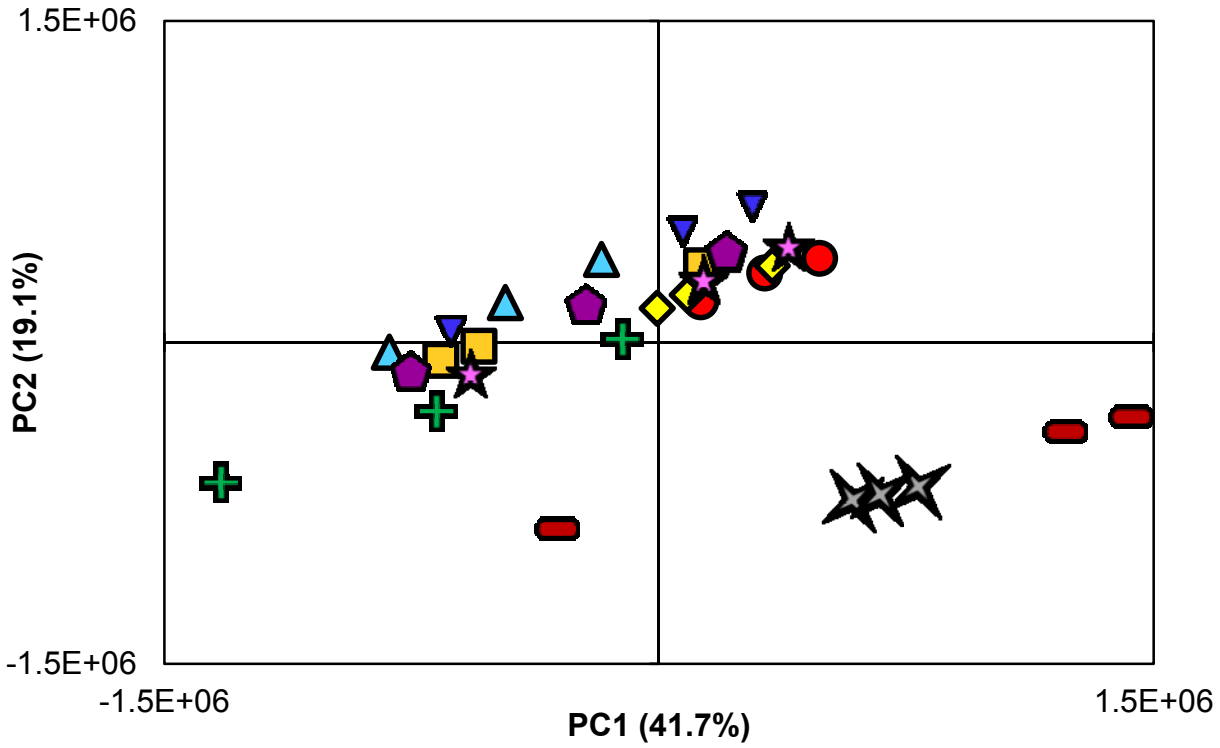


Figure 2-4. PCA scores plot of 10 diesel samples in triplicate. Each diesel sample is represented by a different color and shape: Diesel 1 (dark red ovals), Diesel 2 (grey 4-point stars), Diesel 3 (red circles), Diesel 4 (orange squares), Diesel 5 (yellow diamonds), Diesel 6 (light blue triangles), Diesel 7 (green crosses), Diesel 8 (dark blue inverted triangles), Diesel 9 (purple pentagons), and Diesel 10 (pink 5 point-stars).

This demonstrates that there is as much variation between replicates as there is between samples. This also demonstrates that when PCA is applied to a dataset of chemically similar samples, the non-chemical differences are often identified as the greatest sources of variation between chromatograms.

The loadings plots, which shows the weighting or importance of each variable, can be used to determine which variables are affecting the positioning of the samples on the scores plot. In this work, the loadings plots are presented with the loadings on the y-axis and the variable on the x-axis, which for chromatographic data, is retention time. The loadings plots for PC1 and PC2 are shown in Figure 2-5. In the loadings plot for PC1 (Figure 2-5a), the normal alkanes have the largest influence and contribute most to the positioning of samples on the scores plot. Several short-chain alkanes (C_{11}, C_{12}, and C_{13}) are loading negatively on PC1, affecting the samples with the highest abundance of short-chain normal alkanes. Therefore, winter Diesels 3 – 10, which have a higher abundance of short-chain alkanes, are positioned more negatively on PC1 than the summer diesels.

In the loadings for PC2 (Figure 2-5b), several short-chain normal alkanes (C_{10}-C_{14}) are loading positively. Therefore, compounds with higher abundance of short-chain normal alkanes (the winter diesels) are positioned more positively on PC2 in the scores plot. The two small clusters of peaks on either side of C_{12} (at approximately 17 and 24 min, labeled y and z in Figure 2-1) are present in the loadings plot of both PC1 and PC2 (Figure 2-5). This shows that these peaks were identified as a major source of variation

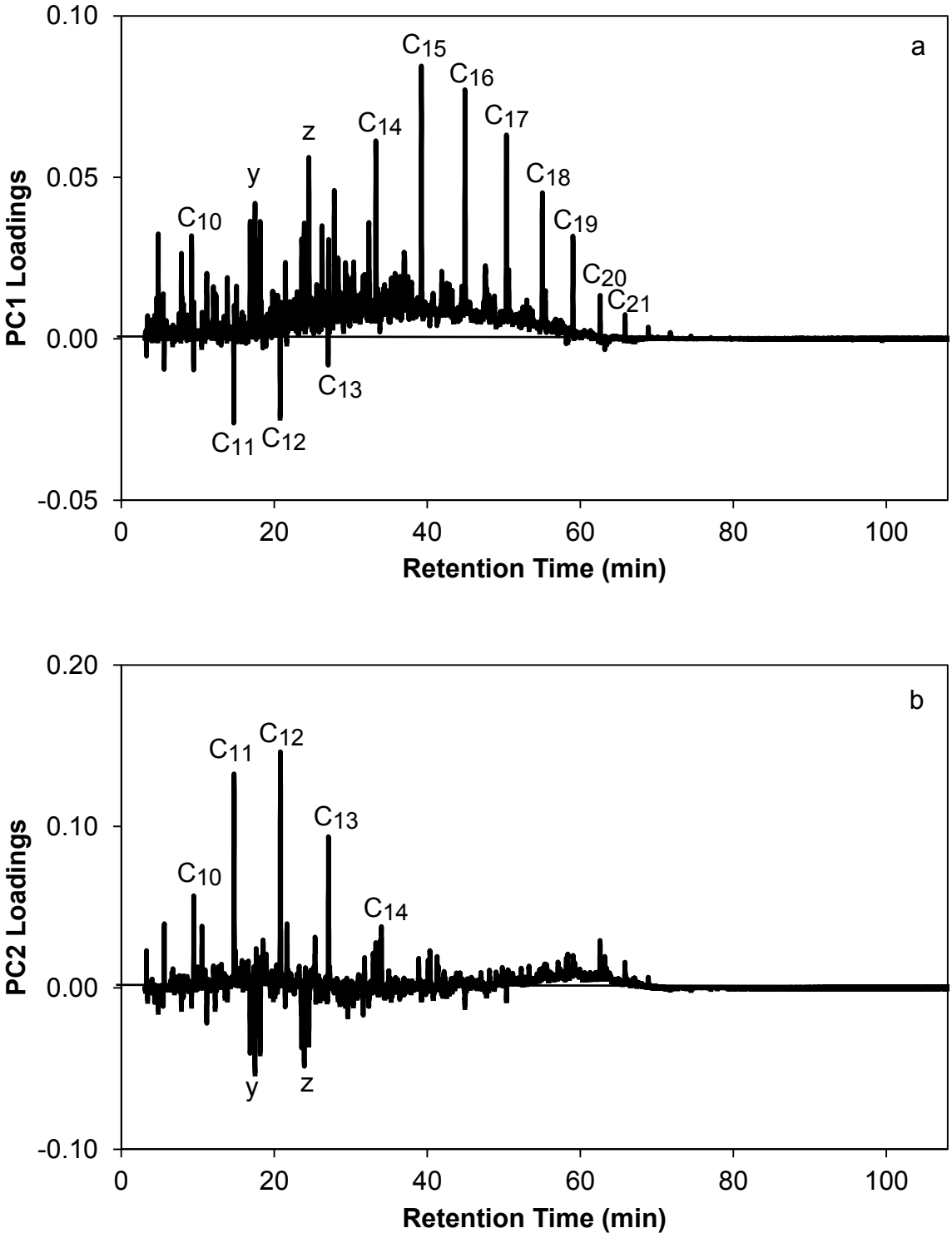


Figure 2-5. Loading plots for PC1 (a) and PC2 (b) after PCA analysis of diesels 1 - 10. The labels y and z correspond to compounds that were provisionally as branched alkanes and substituted aromatic compounds.

between samples. However, these peaks only appear in Diesels 1 and 2, which also explain why these samples are separated from Diesels 3 - 10.

The replicates from Diesels 3 - 10 cluster together, except for one replicate of Diesel 7 (green cross on the left side of Figure 2-4). When the chromatogram of this replicate is compared to all other chromatograms, it has the lowest overall abundance. As most variables in the loadings plot of PC1 and PC2 are loading positively, and this replicate has the lowest abundance, it is positioned most negatively on this PC in the scores plot. This replicate's low abundance is likely due to a lower volume of sample injected into the GC-MS during analysis.

The goal of this work is to investigate different data pretreatment procedures to minimize non-chemical sources of variation, and thereby enhance the discrimination of chemically similar samples, using PCA. To make the differentiation as challenging as possible for this work, samples with no clustering on the scores plot were selected. Therefore, as Diesels 1 and 2 contained chemical differences that were identified by PCA prior to pretreatment, these two samples were omitted from the dataset, and PCA was performed using only replicates of Diesels 3 - 10. Diesel samples 1 and 2 were included in some pretreatment parameter optimization in subsequent chapters, but were omitted from all subsequent PCA scores plots.

The scores plot resulting from PCA of Diesels 3 - 10 is shown in Figure 2-6. No clustering of diesels is observed and most of the replicates are spread along PC1. The loadings plots for PC1 and PC2 are shown in Figure 2-7a and Figure 2-7b, respectively. All compounds are positioned positively in the loadings plot for PC1 and negatively in the loadings plot for PC2. When the overall abundance in the chromatogram varies between

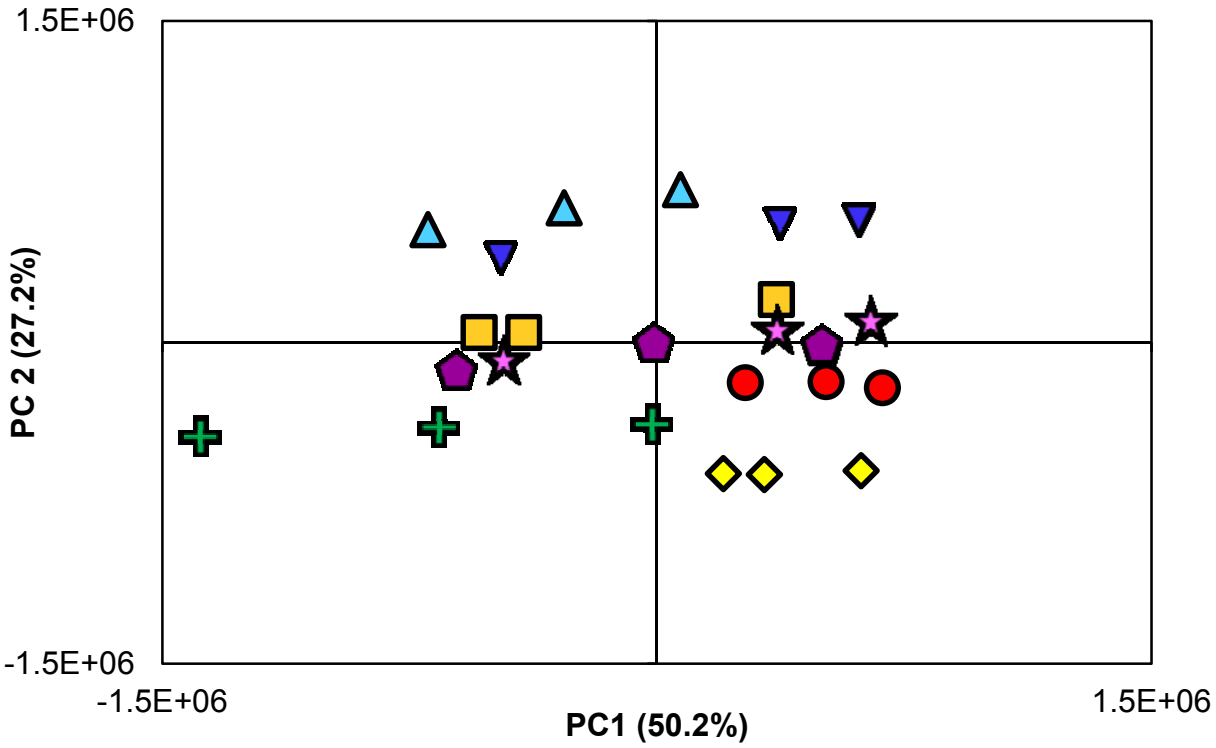


Figure 2-6. PCA scores plot of diesels 3 - 10 in triplicate. Each diesel sample is represented by a different color and shape: diesel 3 (red circles), diesel 4 (orange squares), diesel 5 (yellow diamonds), diesel 6 (light blue triangles), diesel 7 (green crosses), diesel 8 (dark blue inverted triangles), diesel 9 (purple pentagons), and diesel 10 (pink 5 point-stars).

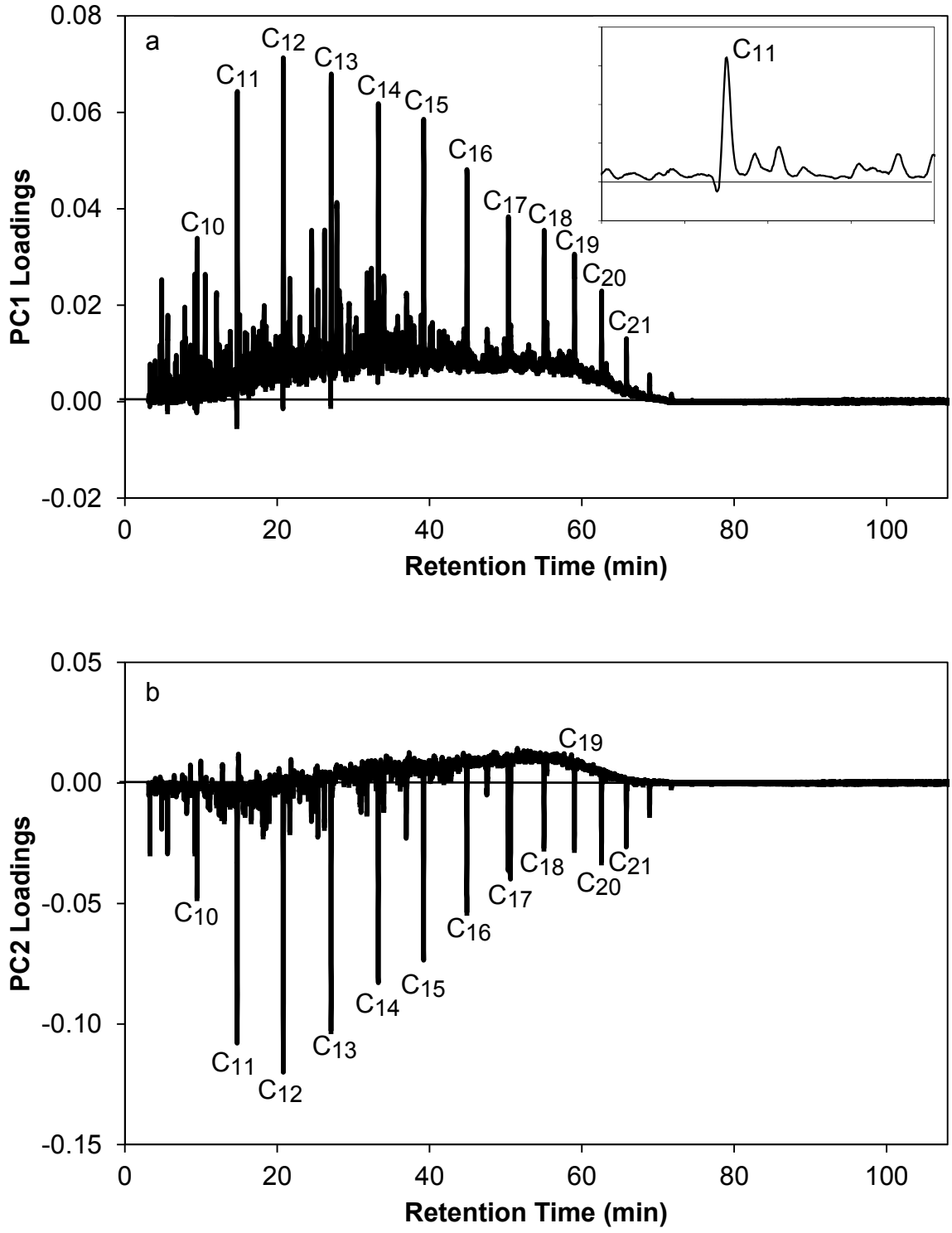


Figure 2-7. Loading plots for PC1 (a) and PC2 (b) after PCA analysis of diesels 3 - 10. The inset shows an expanded region of the undecane (C_{11}) peak to show the derivative-shaped peak, which is characteristic of misalignments.

samples, loadings plots that are mostly positive or mostly negative are common. Further, replicates of each diesel are spread mostly across PC1 in the scores plot, indicating that there are likely differences in abundance among replicates. Hence, the scores and loadings plots indicate that the greatest sources of variation in this dataset are from overall abundance, a non-chemical source of variation, rather than chemical differences.

The loadings plots can provide insight into other non-chemical sources of variation. Derivative-shaped peaks are observed for C_{11}, C_{12}, and C_{13} in PC1 and PC2 (inset Figure 2-7a). The derivative-shaped peaks in the loadings plots result from the peaks in the chromatograms maximizing at slightly different retention times in each sample, indicating retention time misalignment [1, 9]. Therefore, the loadings plots indicate that abundance and alignment are the major sources of variation between chromatograms. The baseline and noise are not prevalent in the loadings plots, indicating that these are not major sources of variation. However, as the largest non-chemical variations are minimized, these lesser sources of variation may become more prominent.

2.6. Summary

Diesel fuel was selected to evaluate the effect of data pretreatment on the PCA of highly similar and chemically complex samples. Eight diesel samples were selected, as they were chemically indistinguishable using PCA, prior to application of data pretreatment. This indicates that the variation from non-chemical sources, such as sample preparation and instrumental analysis, are more discriminatory than the chemical differences between the fuel samples. By minimizing these non-chemical sources of variation, the chemical differences can be utilized to differentiate different diesel samples.

REFERENCES

REFERENCES

[1] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith, Anal. Bioanal. Chem., *394* (2009) 2049.

[2] A.M. Hupp, L.J. Marshall, D.I. Campbell, R.W. Smith, V.L. McGuffin, Anal. Chim. Acta, 606 (2008) 159.

[3] J.M. Baerncopf, V.L. McGuffin, R.W. Smith, J. Forensic Sci., 55 (2010) 185.

[4] J.M. Baerncopf, V.L. McGuffin, R.W. Smith, J. Forensic Sci., 56 (2011) 70.

[5] K.R. Prather, V.L. McGuffin, R.W. Smith, Forensic Sci. Int., 222 (2012) 242.

[6] R.G. Brereton, Applied Chemometrics for Scientists, John Wiley & Sons, Hoboken, NJ, 2007.

[7] K.M. Pierce, J.S. Nadeau, R.E. Synovec, in: C.F. Poole (Ed.), Gas Chromatography, Elsevier, Waltham, MA, 2012.

[8] L.J. Marshall, Association and Discrimination of Diesel Fuels using Chemometric Procedures for Forensic Arson Investigations (Masters Thesis), Michigan State University, Ann Arbor, MI, 2008.

[9] G. Malmquist, R. Danielsson, J. Chromatogr. A, 687 (1994) 71.

CHAPTER 3: NORMALIZATION

3.1. Introduction

Differences in sample abundance are the most common variation observed between chromatograms and can arise for many reasons [1]. In chromatographic data, instrumental fluctuations in flow and injection port temperature can result in abundance differences among samples and even among replicates. The method of injection, including the speed of injection, the amount of time the syringe remains in the injection port, and syringe volume can also result in differences from analysis to analysis. Manual injection is particularly problematic as the injection method (the volume injected, the speed that it was injected, *etc*.) can vary widely for each analysis [2].

Normalization is the most widely applied data pretreatment procedure, even when not utilizing multivariate statistics. Normalization procedures are often used to correct systematic variations in abundance between samples [1, 3]. For chromatographic data, this can be done using some part of the chromatogram, often the height or area of a single peak, or the total area of the chromatogram [4]. However, normalization is often challenging for chromatograms of complex mixtures. Care must be taken when choosing a normalization procedure to ensure that important differences in relative peak abundance among the samples are not lost. In addition, accurate integration to determine peak area for normalization is challenging in complex samples due to co-eluting peaks. Normalization is generally applied after other data pretreatment procedures; however, in this work it is discussed first as this procedure was found to have the greatest effect on the clustering of replicates and discrimination among samples in the scores plot.

3.2. Methods Tested and Evaluation Metrics

In normalization, each data point in a sample (in this work, the abundance at each retention time in each diesel chromatogram) is divided by a unique factor, derived from the sample. There are many methods for determining the normalization factor(s) that are applied: the two most common are the total area and the height from a specific peak in the chromatogram [1, 2, 5-8]. For this work, manual injections were used and no internal standard was included in order to simulate the most challenging normalization scenario.

3.2.1. Total Area Normalization

Total area normalization (also called unit area normalization or constant sum normalization) is performed by dividing the abundance at each retention time (A_t) by the total sum of the abundances in the chromatogram, resulting in the normalized abundance (A_t').

$$A_t' = \frac{A_t}{\sum A_t} \qquad \text{Equation 3-1}$$

In this work, the total area of each chromatogram was approximated by summing the abundance at each retention time in the total ion chromatogram. The abundance at each retention time was then divided by this sum to normalize the chromatogram. Each point in every chromatogram was multiplied by the average total area across all chromatograms in the dataset to return them to the original order of magnitude.

The major assumption using total area normalization is that the total signal response from one sample is equivalent to the total signal from another. In other words, this method assumes that the same volume of each sample has the same instrument

response [2]. While this is rarely true, because response factors differ between compounds, when a large number of compounds are present, this is often a reasonable approximation. A major drawback of this method is that when one peak decreases in size, another peak necessarily increases, which can result in misleading correlations between samples [6]. Area normalization is a good initial method for normalization because it is generally fast and easy to apply and often results in adequate normalization. Additionally, for complex samples, where there are unresolved peaks or a high baseline, area normalization often results in the best minimization of the variation introduced from injection [9].

3.2.2. Single Peak Normalization

In single peak normalization (also called maximum peak or internal standard normalization), each data point is divided by the amplitude of a specific peak of interest (A_I) in the data.

$$A_t' = \frac{A_t}{A_I} \qquad \text{Equation 3-2}$$

For chromatographic data, the peak height or peak area of a selected peak (which is constant across all samples) is used for normalization by dividing each data point by the peak height or peak area of the selected peak.

The most common single peak normalization utilizes an internal standard. Prior to analysis, the same concentration of a non-native compound is spiked into each sample. An ideal internal standard for chromatographic data should have similar physical properties to the compound being analyzed, elute close to the compound of interest. In

addition, the internal standard should be completely resolved in the chromatogram (avoiding increased signal from co-elution), and be at a similar concentration to the compound(s) of interest [7, 8]. Selection of an appropriate internal standard is challenging, especially for a complex mixture where there are compounds with many different properties and at different concentrations [2]. In many cases, a deuterated analogue of each compound is utilized. However, this type of internal standard can be expensive and very challenging to obtain for all compounds in a complex sample, as there are a large number of compounds present. If the internal standard is not properly selected, is present at an inappropriate abundance, or co-elutes with another compound, the internal standard itself can become a major source of variance in PCA. This highlights the need to think about proper data pretreatment procedures, even before sample collection.

When an internal standard is not added, a compound within the sample can be used for single peak normalization. This peak could be the highest abundance peak in each sample, or could be a peak that is common to each sample. However, single peak normalization can skew the relative abundance between samples because the abundance of the selected peak may not truly be the same in all samples. Therefore, normalization to a peak in the sample is often problematic if the abundance of that peak changes between samples.

For this work, each point was divided by the peak height of heptadecane (C_{17}) (at approximately 50.3 min) and multiplied by the average heptadecane peak height across all samples [2]. Heptadecane was chosen because it is a large, retained peak and is less affected by evaporation or variation introduced during injection, due to its lower volatility.

3.2.3. Evaluation Metrics

In order to evaluate the effect of normalization, all 24 diesel chromatograms (Diesels 3 - 10 analyzed in triplicate) were normalized using both total area and single peak normalization methods. Initially, to assess the effect of normalization, a visual inspection of overlaid chromatograms before and after normalization was used. However, comparing overlaid chromatograms is subjective and time consuming as it requires observing only small regions of the chromatograms at one time. In order to quantitatively compare normalization methods, the percent change in total sum of squares of the residuals for replicates (SSR) was developed. To calculate the SSR, an average chromatogram of the triplicates for each diesel sample was calculated. The residual was calculated by subtracting each replicate chromatogram from the corresponding average chromatogram. The residuals were squared and then summed for all 24 chromatograms. The percent change in the SSR between the untreated and the normalized data was then calculated. Using the residuals of replicates allows for monitoring both the peaks and the baseline. Theoretically, for instrument replicates, which are chemically the same, the SSR should be zero. When any variation among replicates is present, the differences must arise from injection and instrumental analysis. Ideally, normalization would remove all differences in peak height, making the replicate samples have the same height at each retention time.

Evaluation of how normalization affected the association of replicates on the PCA scores plot was performed using a visual inspection of clustering patterns. Additionally, the effect of normalization on the clustering or grouping of replicates was quantitatively assessed using the average percent change in the clustering of replicates (PCC). The

PCC was calculated by summing the variance in PC1 and PC2 for replicates of each diesel. The standard deviation was then calculated by taking the square root of the variance. The standard deviations were averaged and the PCC was calculated as the percent change in the standard deviation between the scores plot generated from the chromatograms after data pretreatment and the scores plot generated from chromatograms prior to pretreatment.

3.3. Effect of Normalization on Chromatographic Data

3.3.1. Visual Assessment

An expanded region near the hexadecane (C_{16}) peak of three representative diesel chromatograms is shown in Figure 3-1 before (a) and after each normalization method (b - c). The inset in each figure shows the further expanded baseline, just after the hexadecane peak. Without normalization (Figure 3-1a) there is spread in the abundance along the baseline as well as at the peak maxima. Because these are replicates of the same sample, these differences are due to small differences in injection volume and instrumental variation. After total area normalization (Figure 3-1b), there is less spread in the baseline of the three samples; however, spread in the abundance is still observed at the peak maxima. Using single peak normalization (Figure 3-1c), the peak maxima are close together, while spread is still observed along the baseline. This demonstrates that the normalizations investigated in this work could not correct all of the variation observed.

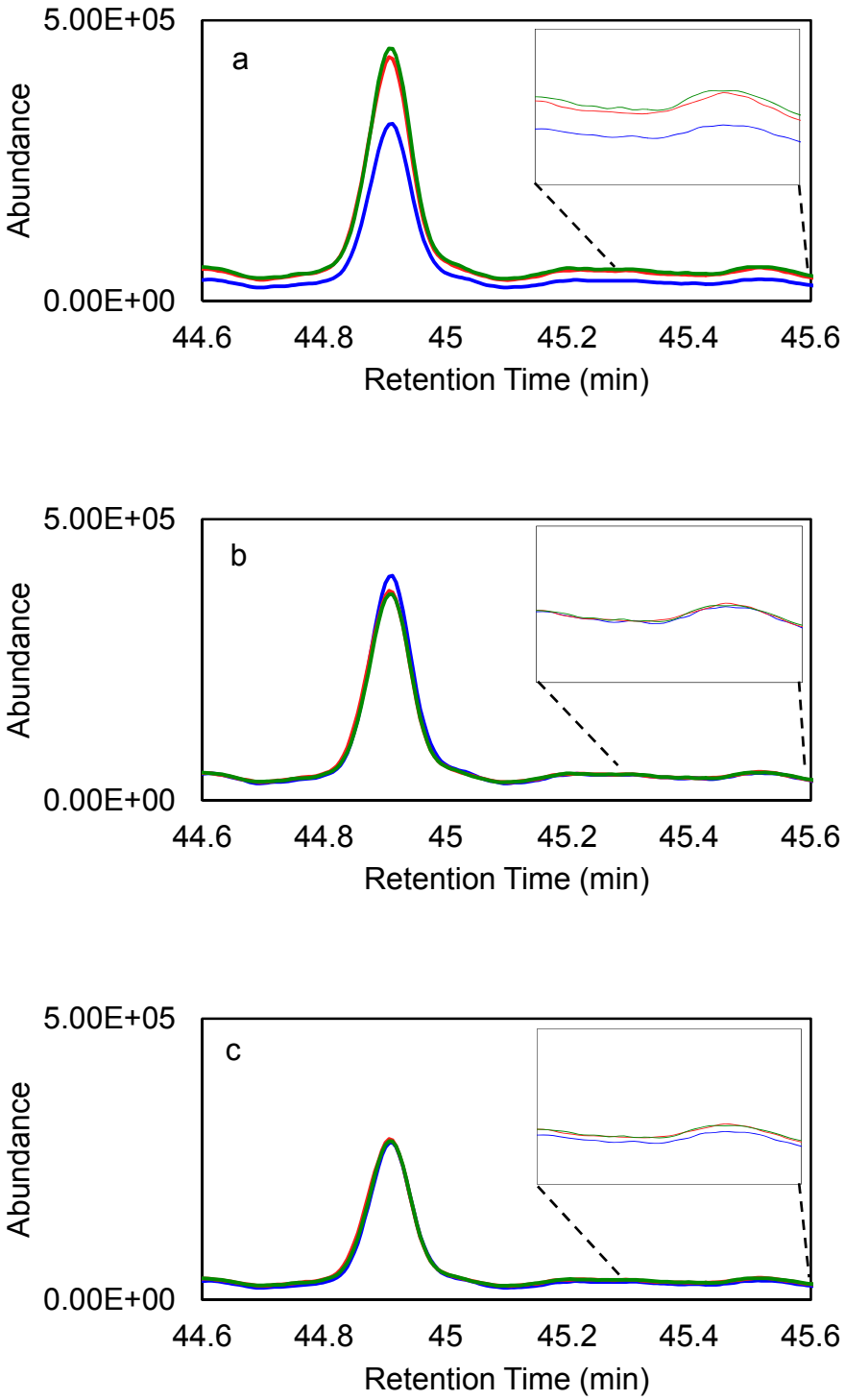


Figure 3-1. An expanded region of the hexadecane peak (C_{16}) in triplicate analysis of a diesel sample, before normalization (a), after total area normalization (b), and after selected peak normalization (c).

3.3.2. Quantitative Assessment

Both normalization methods resulted in a reduction in the spread in the abundance between replicate chromatograms (Figure 3-1). Using the quantitative metric, there is a 92% decrease in the SSR using the area normalization procedure and an 87% decrease using peak normalization. These percent decreases in the SSR are very similar, indicating that both methods result in a large reduction in the variation. However, in this work, the majority of the chromatogram consists of an unresolved baseline; therefore, area normalization resulted in a larger improvement than single peak normalization.

3.4. Effect of Normalization on PCA Scores Plot

Because normalization was determined to be the most important pretreatment procedure for this dataset, both total area and single peak normalization were utilized prior to PCA. Triplicate chromatograms of the eight diesels (Diesels 3 - 10) were normalized using each method, then PCA was performed.

3.4.1. Visual Assessment

Both normalization methods resulted in enhanced clustering of replicates when compared to the PCA scores plot prior to data pretreatment (Figure 3-2a). Using total area normalization (Figure 3-2b), PC1 accounts for 57.9% of the variation and PC2 accounts for 21.4%. After area normalization, there is still spread among replicates, mostly along PC2 on the scores plot, indicating additional sources of non-chemical variation (Figure 3-2b). The loadings plot of PC1 after total area normalization (Figure 3-3a) shows mostly the normal alkane peaks positioned positively, likely due to the variation in peak height shown in Figure 3-1b. However, the largest peaks in the PC1 loadings plot

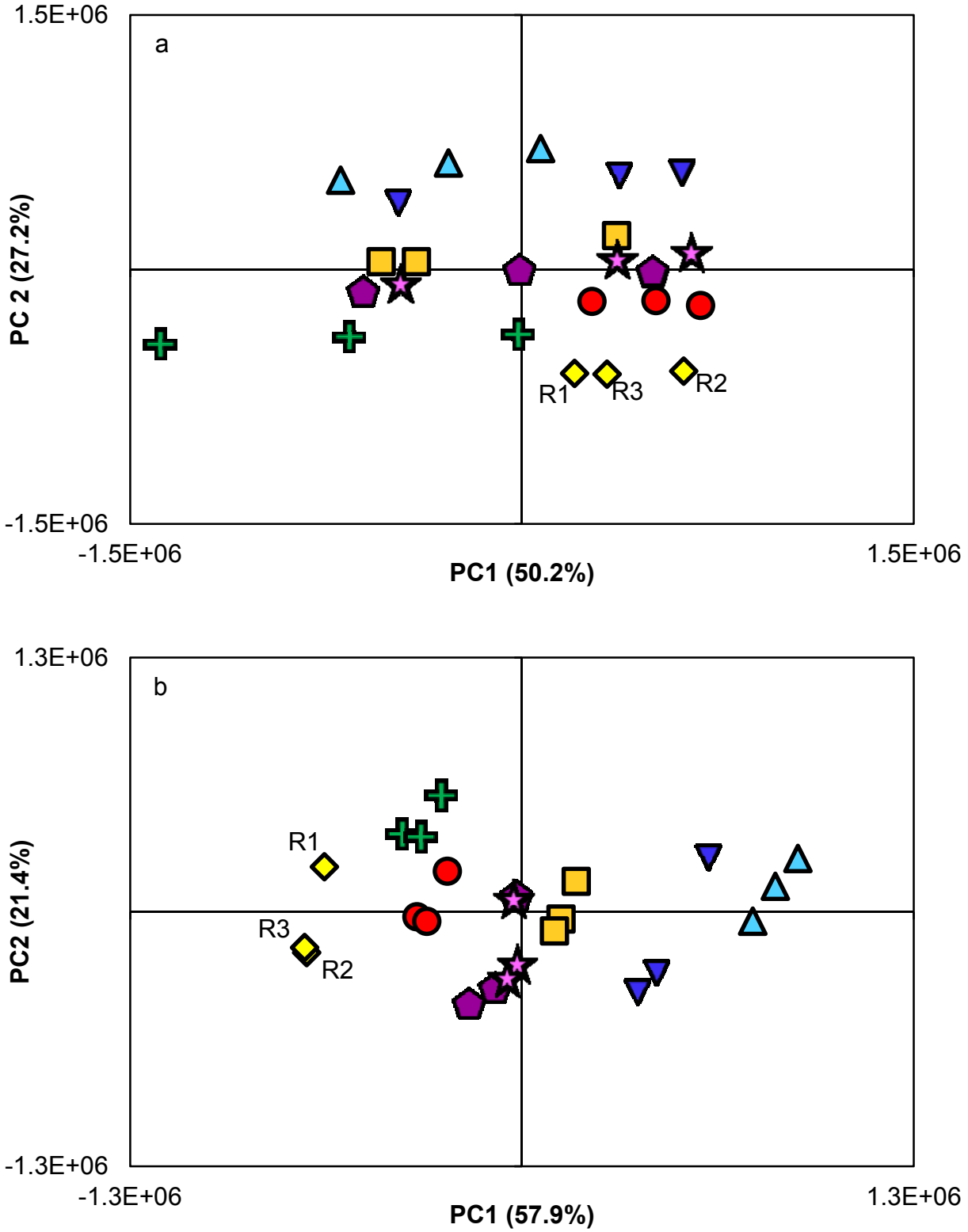


Figure 3-2. PCA scores plot of eight diesel chromatograms in triplicate prior to the application of data pretreatment (a) and after total area normalization (b). Each diesel is represented by a different shape and color.

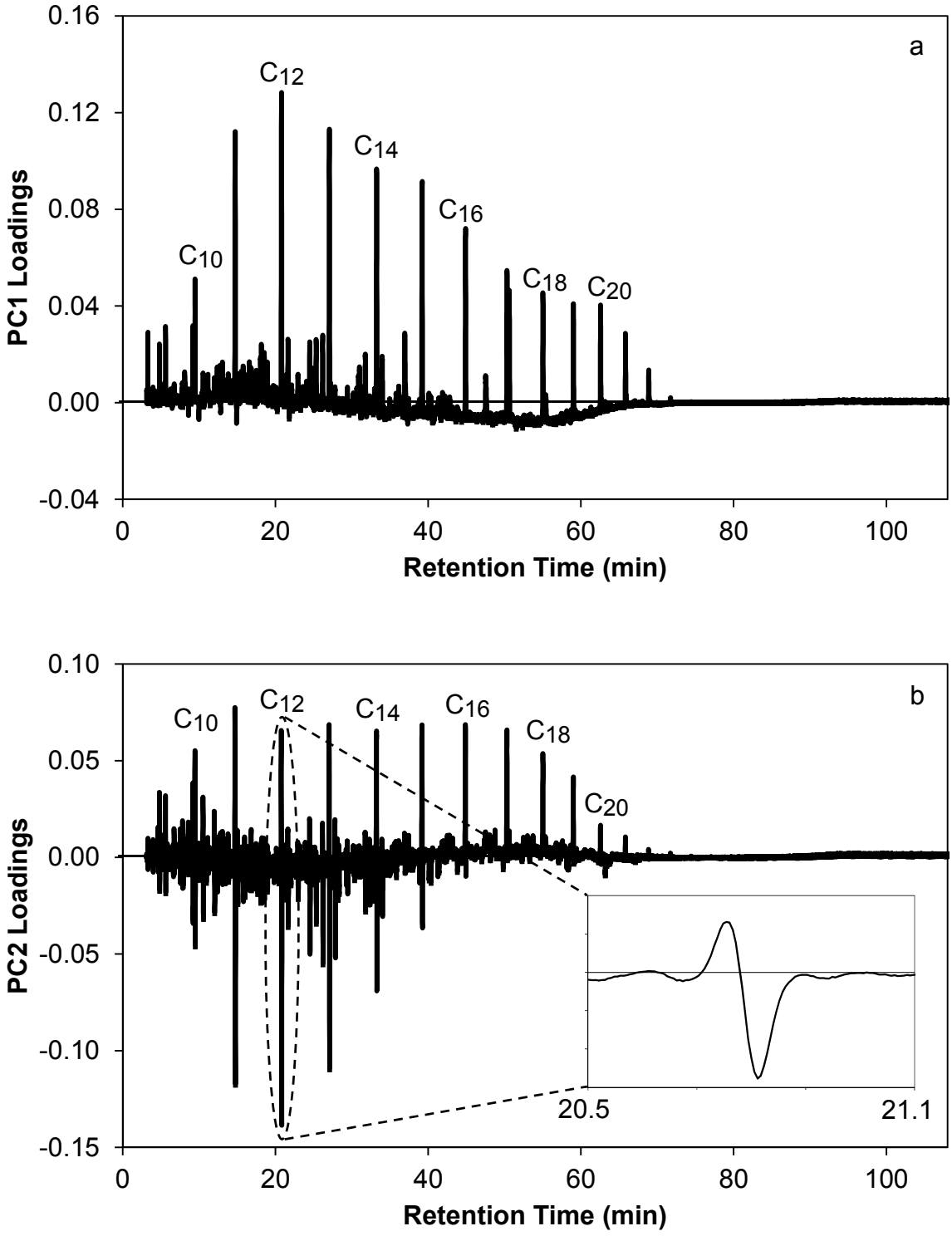


Figure 3-3. Loadings plot for PC1 (a) and PC2 (b) after PCA with total area normalization. The inset in part b shows a derivative shaped peak, indicative of misalignments.

are the short-chain alkanes, C_{11} - C_{13}, with the largest peak at C_{12}, corresponding to the peaks that maximize in the bimodal distribution (see Chapter 2, Figure 2-1). Therefore, some information contained in PC1 is also chemical differences between samples. The loadings plot of PC2 (Figure 3-3b) shows the most dominant peaks as derivative-shaped curves (see inset), indicating that the greatest source of variation on PC2 is retention time misalignments [10]. Most of the spread among replicates occurs on PC2 on the scores plot because misalignments are dominating PC2.

To demonstrate the correction of non-chemical sources of variation, Diesel 5 (yellow diamonds) was chosen and the change in clustering will be highlighted throughout the subsequent chapters. In order to observe the changes in the chromatogram, three replicates of Diesel 5 (labeled R1, R2, and R3) were overlaid and the region around the dodecane peak (C_{12}) was expanded (Figure 3-4). Prior to any pretreatment (Figure 3-4a), R1 is shifted to the left of the other two replicates and R2 is at a higher abundance than the other replicates. After total area normalization, all replicates were at approximately the same height, while R1 was still shifted to the left of the other replicates. This results in the spread in the replicates observed in Figure 3-2b. In the scores plot prior to data pretreatment (Figure 3-2a), the three replicates of Diesel 5 were spread along PC1. From the loadings plot shown in Chapter 2 (Figure 2-7), PC1 included differences in height and misalignments. However, after normalization, peaks were at approximately equal heights (Figure 3-4b). Therefore when PCA was performed, the variation in height was minimized (Figure 3-2b), and R2 and R3 were positioned close together. R1 was still separated along PC2, due to the misalignments. This is supported in the PC2 loadings plot

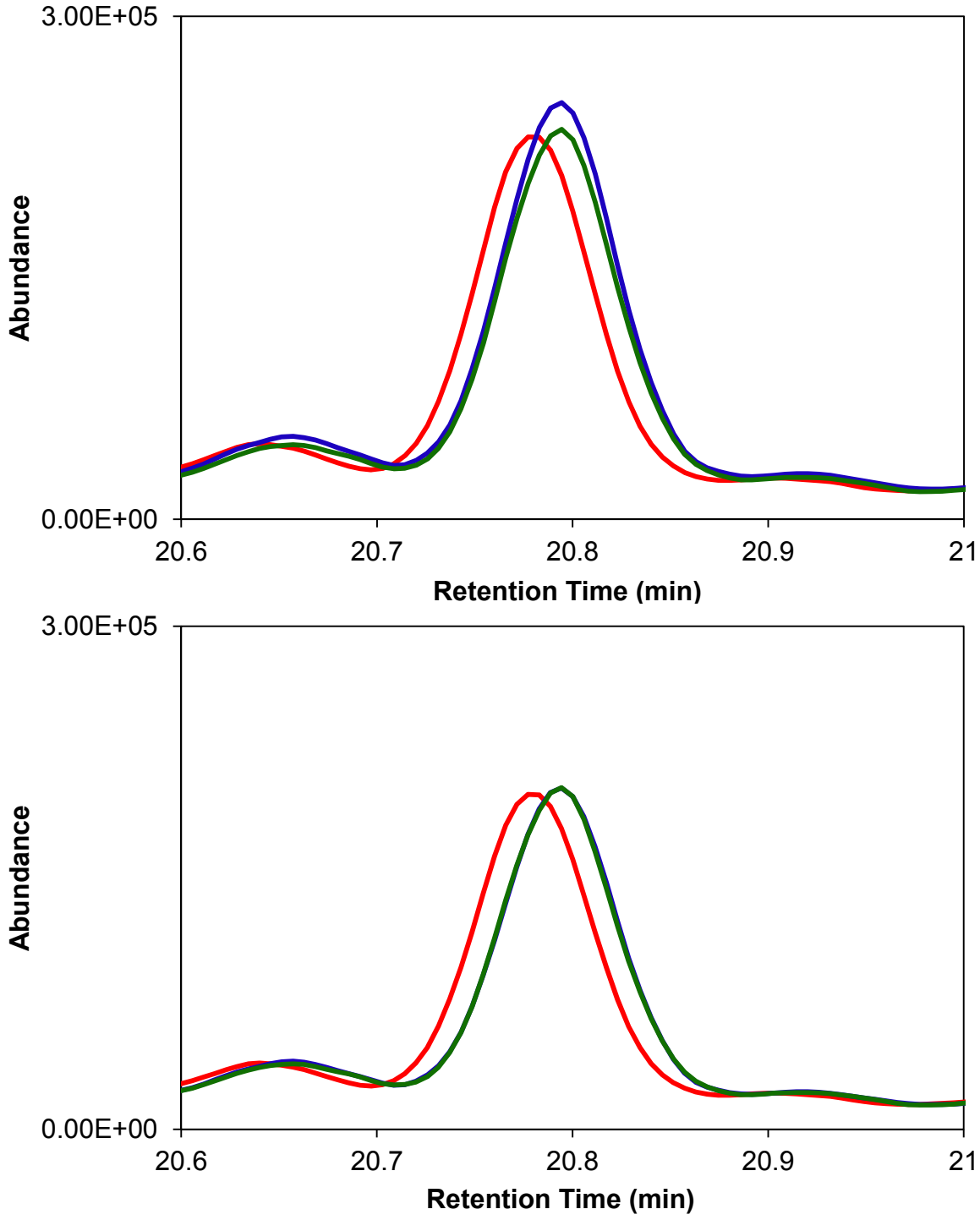


Figure 3-4. An expanded region of dodecane in three replicate chromatograms of diesel 5 before (a) and after (b) area normalization (R2 and R3 are directly on top of one another). Each replicate is indicated by a different color (R1: red, R2: blue, R3: green).

(Figure 3-3b), which shows derivative-shaped curves for many of the normal alkanes that are present.

After application of single peak normalization (using hexadecane), PC1 accounted for 60.2% of the variance and PC2 accounted for 21.4%. Peak normalization results in the spread among replicates occurring along both PC1 and PC2 (Figure 3-5b), indicating that non-chemical variation is present in both principal components. This is supported by the loadings plot for the peak normalized chromatograms (Figure 3-6). Derivative-shaped peaks are observed for many of the alkanes in PC1, indicating that misalignments are still a major source of variation. Additionally, the loadings plot for PC1 shows a large portion of the unresolved region of the chromatogram (20 - 65 min) contributing to the loadings, indicating that the differences in the unresolved baseline region are a major contribution to the variance. This agrees with the visual assessment of the chromatograms, which shows that after peak normalization, variation in the baseline between replicates was still present (Figure 3-1). The loadings plot of PC2 (Figure 3-6b) is very similar to the PC1 loadings plot after area normalization (Figure 3-3). As with the area normalization, this pattern is likely due to the chemical differences between diesel samples with the unimodal and bimodal distribution of normal alkanes.

The dodecane peak (C_{12}) in replicates of Diesel 5 can again be utilized to explain the spread on the PCA scores plot present in the samples and replicates (Figure 3-7). After peak normalization (Figure 3-7b), there are still some differences in the height between R1, R2, and R3. In the PC1 loadings plot (Figure 3-6a), many of the peaks from the normal alkanes, including the dodecane peak, are derivative-shaped peaks, indicating

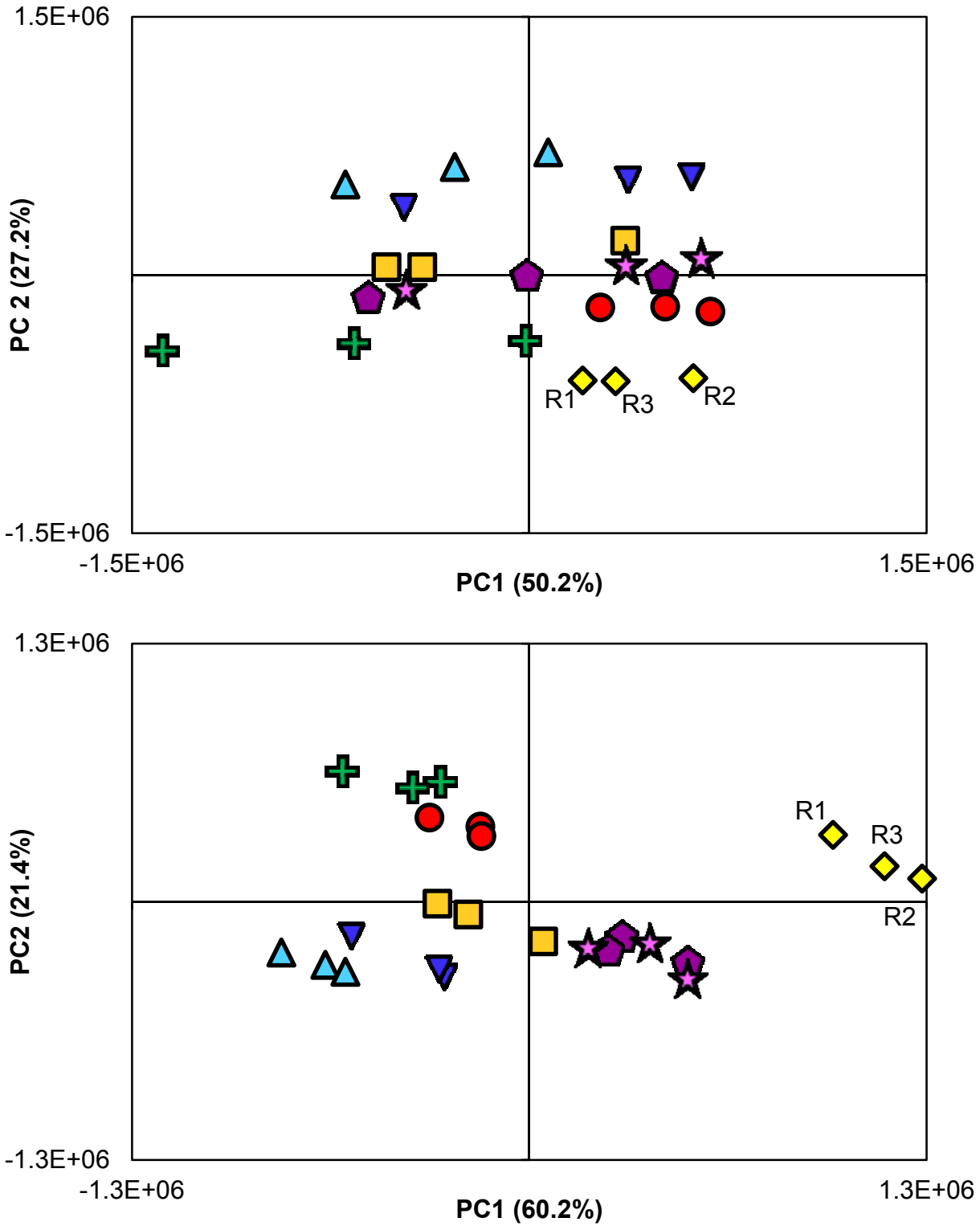


Figure 3-5. PCA scores plot of eight diesel chromatograms in triplicate prior to the application of data pretreatment (a) and after single peak normalization (b).

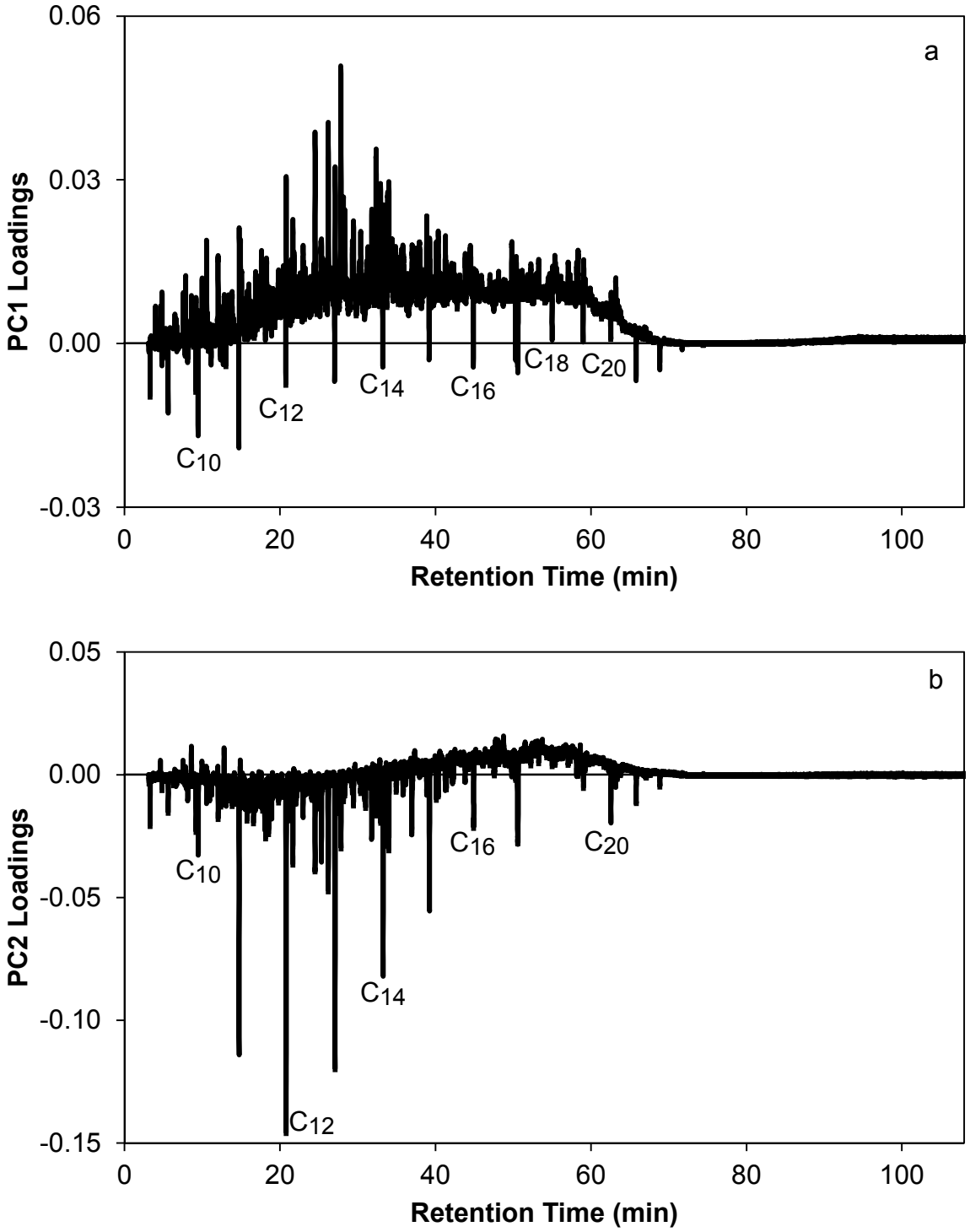


Figure 3-6. Loadings plot for PC1 (a) and PC2 (b) after PCA with single peak normalization.

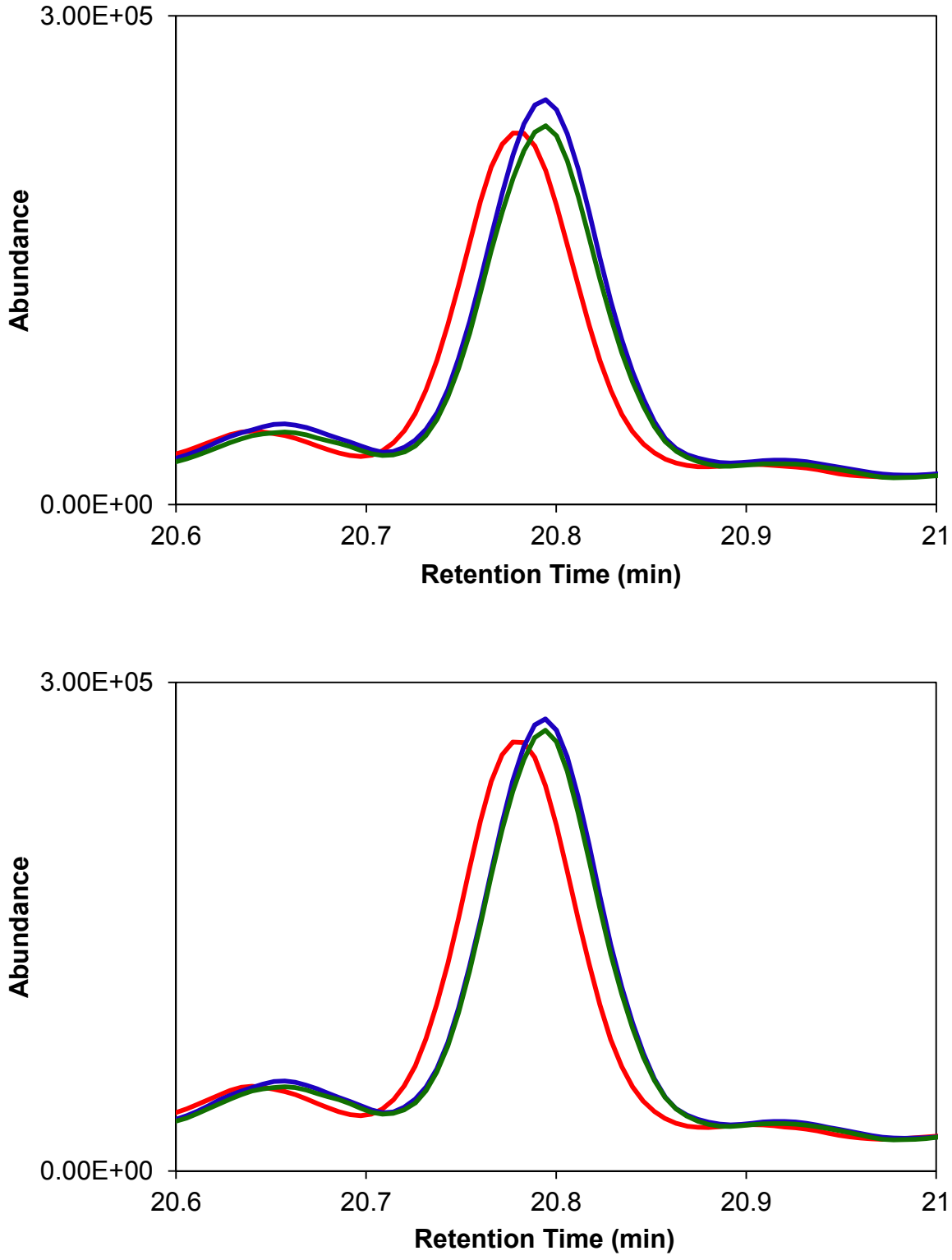


Figure 3-7. An expanded region of dodecane in three replicate chromatograms of diesel 5 before (a) and after (b) peak normalization. Each replicate is indicated by a different color (R1: red, R2: blue, R3: green).

retention time misalignment. After normalization, Diesel 5 R1 (shown in red in Figure 3-7) is still misaligned to the other two replicates, explaining some of the spread observed in PC1. The dodecane peak and several other normal alkane peaks are heavily weighted in the loadings plots (before and after each normalization), indicating that dodecane is greatly influencing the positioning of samples on the scores plot. When large peaks are not well aligned and normalized, there is variation in the resulting PCA scores plot. This explains why R1, R2, and R3 are spread even after normalization, in both PC1 and PC2.

After application of area normalization, chemical differences between samples are beginning to become useful in differentiating samples, based on the loadings plot of PC1. Non-chemical sources of variation are still observed on PC2, resulting in some spread among replicates. Using peak normalization, misalignments and variation in the baseline are observed on the loadings plot of PC1 and chemical differences are observed on PC2. This demonstrates that area normalization minimizes more of the non-chemical variations, thus allowing chemical differences between samples to become the greatest source of variation. However, with peak normalization, additional pretreatment procedures are required to minimize the non-chemical sources of variation.

While differences in the baseline may be small (generally less than 3% of the signal), they contribute to the variation because the differences occur over the entire chromatogram (Figure 2-2). In this work, both a large number of relatively small peaks (compared to the baseline) and a large, unresolved region were present, complicating normalization (Figure 2-1). The baseline must be well normalized, because the baseline accounts for most of the points in the chromatogram. However, the peaks also need to be well normalized because most of the chemical differences in complex samples will

come from differences in peak heights. After utilizing area normalization, real chemical differences were identified as the greatest source of variation, followed by other non-chemical sources of variation. After single peak normalization, the greatest source of variation was still non-chemical, followed by chemical differences between the samples. This demonstrates that area normalization is more effective at correcting the major sources of variation.

3.4.2. Quantitative Assessment

The average percent change in the clustering of replicates (PCC) between the unnormalized and normalized chromatograms was calculated to assess improvements in clustering attained using each normalization procedure. After applying total area normalization, the PCC was 45.1%, indicating that replicates on the scores plot were closer together after normalization than without normalization. Similar results were observed for the single peak normalization, where the PCC was 58.9%. The higher PCC using the single peak is likely due to Diesel 5 (yellow diamonds, Figure 3-2 and Figure 3-5). After area normalization, one replicate of this sample became further separated due to misalignments, resulting in a negative PCC for that sample. If this sample is removed, the average PCC for the area normalization increases to 52.9%, demonstrating similar clustering of replicates as observed using peak normalization.

3.5. Summary

Proper normalization of complex chromatographic data can be challenging; however, normalization is a critical data pretreatment procedure to minimize non-chemical sources of variation. Small differences in abundance, caused by variation in

injection, are often the greatest source of variation for complex and similar samples. In this work, both normalization methods resulted in a reduction in the non-chemical variation. Based on the chromatographic metric, the percent change in the sum of squares of the residual, total area normalization resulted in a larger reduction in variation. However, using the metric for the PCA scores plot, the percent change in the clustering of replicates showed that single peak normalization resulted in better clustering of replicates for this data. This demonstrates that selection of the proper normalization method is dependent on the data and non-chemical variation that is present. The most important point demonstrated in this work is that selection of the particular normalization method is not critical; however, the application of normalization procedures drastically improves discrimination of highly complex samples using PCA. Even though the performance of each normalization method is similar for these data, it is still important that analysts consider selection of an appropriate normalization method, to ensure that the method that is chosen does not skew the data. In some cases, one or more of the assumptions that are made may not always be valid and can lead to erroneous results.

REFERENCES

REFERENCES

[1] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, Inc., New York, NY, 1998.

[2] K.M. Pierce, J.S. Nadeau, R.E. Synovec, in: C.F. Poole (Ed.), Gas Chromatography, Elsevier, Waltham, MA, 2012.

[3] B.K. Lavine, in: S.J. Haswell (Ed.), Practical Guide to Chemometrics, Marcel Dekker, Inc. , New York, NY, 1992, p. 211.

[4] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, New York, NY, 2009.

[5] A. Moreda-Pineiro, A. Marcos, A. Fisher, S.J. Hill, J. Environ. Monit., 3 (2001) 352.

[6] G. Malmquist, R. Danielsson, J. Chromatogr. A, 687 (1994) 71.

[7] J.D. Ingle Jr., S.R. Crouch, Spectrochemical Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[8] D.A. Skoog, F.J. Holler, S.R. Crouch, Principles of Instrumental Analysis, Thomson Brooks/Cole, Belmont, CA, 2007.

[9] S.L. Morgan, E.G. Bartick, in: R.D. Blackledge (Ed.), Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis, John Wiley & Sons, Inc. , Hoboken, NJ, 2007.

[10] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith, Anal. Bioanal. Chem., *394* (2009) 2049.

CHAPTER 4: BASELINE CORRECTION

4.1. Introduction

In temperature-programed gas chromatography (GC), the rise of the baseline at the end of the chromatogram (at high temperature) can vary widely between analyses. This rise is due to the degradation of the polysiloxane stationary phase in the column as well as breakdown of the silicone septum [1]. As these baseline differences are not from the sample, minimizing the baseline should not change the chemical information contained in the chromatogram [1, 2]. In the chromatograms in this work, the unresolved portion in the middle of the chromatogram between 15 and 65 minutes also has an elevated baseline which can introduce variation into the chromatograms. However, for the data considered here, the unresolved portion is sample dependent, originating from the large number of compounds with similar boiling points, which were not well separated by gas chromatography-mass spectrometry (GC-MS). Hence, these variations are chemically important and should not be removed.

There are a number of different baseline correction methods that can be applied. These methods generally fall into one of four categories: (1) transformation of the signal, (2) subtraction of a chromatogram, (3) subtraction of a modeled function, or (4) removal of specific signals in the sample chromatogram [1-4]. Each of these correction methods is discussed in more detail below.

Signal transformation is an indirect method of baseline correction, in which the baseline is not removed, but rather transformed, so that it is no longer meaningful. Although many transforms are available, the most common is using the first-derivative of

the chromatogram [2, 4]. However, this method can often enhance the noise in the chromatogram [4].

In chromatogram subtraction, a chromatogram that does not contain the sample (generally the chromatogram of a solvent blank) is subtracted from each sample chromatogram. This is a fast and simple correction method [1]. However, when the baseline varies between injections, specifically between the blank and sample chromatogram, this method cannot be used because subtraction of a single blank chromatogram will not correct for variations in baseline among all samples and may result in additional variations. This method would then require running a blank after every sample, which would greatly increase analysis time.

Another baseline correction method utilizes a mathematical function to model the baseline in each chromatogram, which is then subtracted from each sample chromatogram [1-4]. The age of the stationary phase, the solvent, and the temperature program can all affect the shape of the baseline and hence, functions used for fitting can range from a simple linear fit to more complex high-order polynomials. As this method is specifically modeled to fit each sample chromatogram, it will more accurately correct for variations in the baseline among samples. However, care must be taken to ensure that the model only accounts for non-chemical signals of the baseline. This is often difficult to achieve, particularly in chromatograms of complex samples where resolution is poor or in cases where peaks of relevance elute during the rise in the baseline.

Removal of individual signals in the chromatogram is specific for data generated by mass spectrometry, where the total signal is the sum of individual ions. In this method,

specific ions are subtracted from the total ion chromatogram (TIC); typically these ions correspond to column degradation and septum bleed. Common mass-to-charge (*m/z*) ratios for ions resulting from column and septum degradation include *m/z* 73, *m/z* 147, *m/z* 207, and *m/z* 281 [5]. However, these ions could also result from fragmentation of chemical compounds in the sample. Therefore, removal of these ions can alter the chemical signal.

The selection of appropriate baseline correction methods is highly sample dependent. Considerations for selecting a baseline subtraction method include the complexity of the sample, the source of the baseline signal, and the compounds present in the sample. Care must be taken to ensure that only signal from the baseline is being removed, and not signal from compounds in the sample.

4.2. Methods Tested and Evaluation Metrics

Three different baseline correction methods were compared in this research. Preliminary results indicated that transformations and subtraction of a solvent blank chromatogram were not appropriate for these data. Using a first-derivative transform led to challenges in applying other data pretreatment procedures, particularly alignment as the peaks were no longer the traditional Gaussian-shaped peaks that are typically observed in chromatography. Subtracting the background from a solvent blank chromatogram resulted in an incomplete reduction of the baseline and was therefore discounted. The methods selected for this work all utilized extracted ion chromatograms (EICs) to remove specific background signals from the chromatogram and included the

background subtracted baseline, subtraction of extracted ion profiles of the background signal, and subtraction of a the baseline using a modeled function.

4.2.1. Background Subtracted Baseline (BSB)

The background subtracted baseline (BSB) method for correcting the baseline involves removing individual ions from the TIC and is included as a function in the instrument software. Chemstation (Version E.02.01.1177, Agilent Technologies, Santa Clara, CA) was used to select a specific mass spectrum, which was then subtracted from each individual mass spectral scan in the chromatogram [6]. For this work, the last scan in the TIC of each chromatogram was used for subtraction (Figure 4-1). Generally, scans at the end of the chromatogram contain ions from only column degradation and septum bleed, making this region useful for evaluating the baseline. Column degradation occurs more readily at high temperature. The GC oven is at the highest temperature at the end of the analysis, resulting in the end of the chromatogram containing most of the ions resulting from degradation. Multiple scans can also be subtracted by repeating this procedure, if additional reduction is required. The BSB function does not allow for negative ion intensities, so the subtraction of any ion that would result in a negative number becomes zero. The function can also be used to remove other background interferences in the chromatogram, such as impurities, by selecting a scan containing the ions characteristic of the interference.

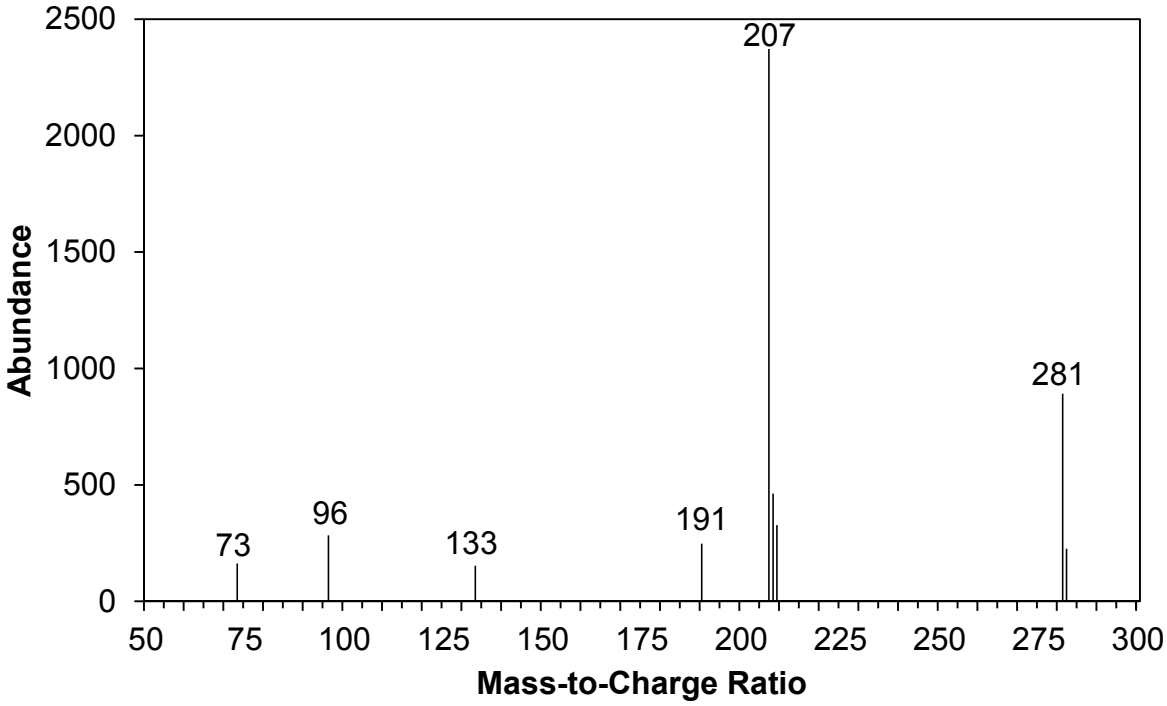


Figure 4-1. Representative mass spectrum from a diesel chromatogram (Diesel 1) at retention time 108.335 minutes, the last scan in the chromatogram.

4.2.2. Subtraction of Extracted Ion Profiles

As some ions in the spectra used for BSB subtraction may also be fragment ions of some of the compounds in the sample, resulting in a reduction in a reduction in the peaks that also contained those ions. Therefore, other subtraction methods may be necessary to prevent the loss of chemical information.

In this work, rather than removing all ions in a given spectrum, the removal of selected ions was also investigated. The EICs of the ions of interest from the baseline were generated in the Chemstation software. The *m/z* that were investigated were the ions present in the last scan of the chromatogram: *m/z* 73, *m/z* 96, *m/z* 133, *m/z* 191, *m/z* 207, *m/z* 208, *m/z* 209, *m/z* 281, and *m/z* 282, all of which are characteristic polysiloxane fragments [5]. The EIC of each *m/z* of interest from Diesel 1 is shown in Figure 4-2. Each EIC was examined to determine if any ions of that *m/z* were also generated from compounds in the sample. As shown in Figure 4-2, *m/z* 96 and *m/z* 133 had high abundances in the peak region of the chromatogram (prior to 70 minutes), which were not present in the same EIC of a solvent chromatogram, demonstrating that these ions also resulted from fragmentation of compounds that are in the sample. Therefore, *m/z* 96 and *m/z* 133 should not be removed from the TIC and were eliminated from further consideration.

The six most abundant EICs were chosen to create a baseline extracted ion profile (EIP) that could be subtracted from the TIC, in order to minimize the baseline. As *m/z* 73 had a relatively low abundance, it was also excluded from the EIP. The selected EICs

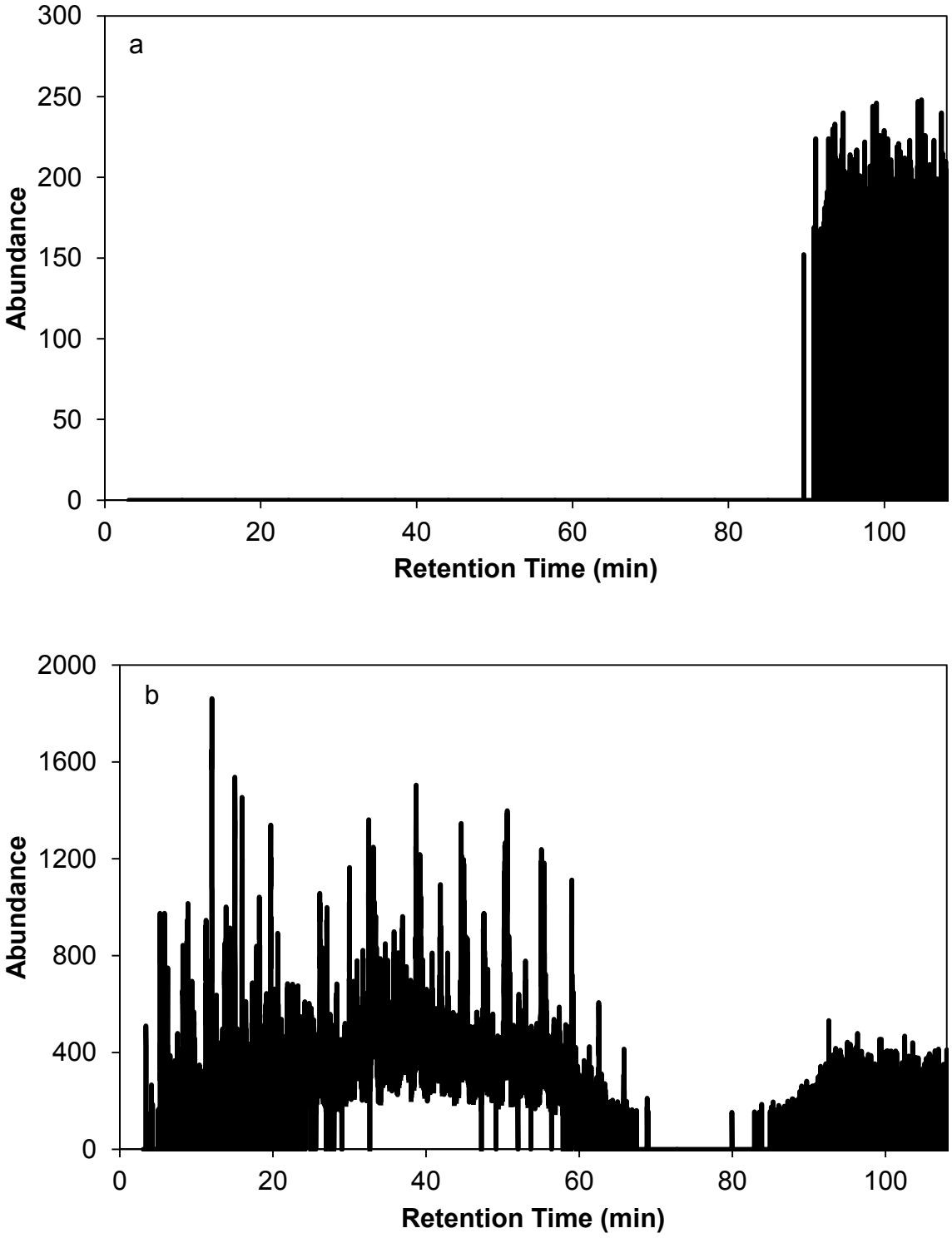
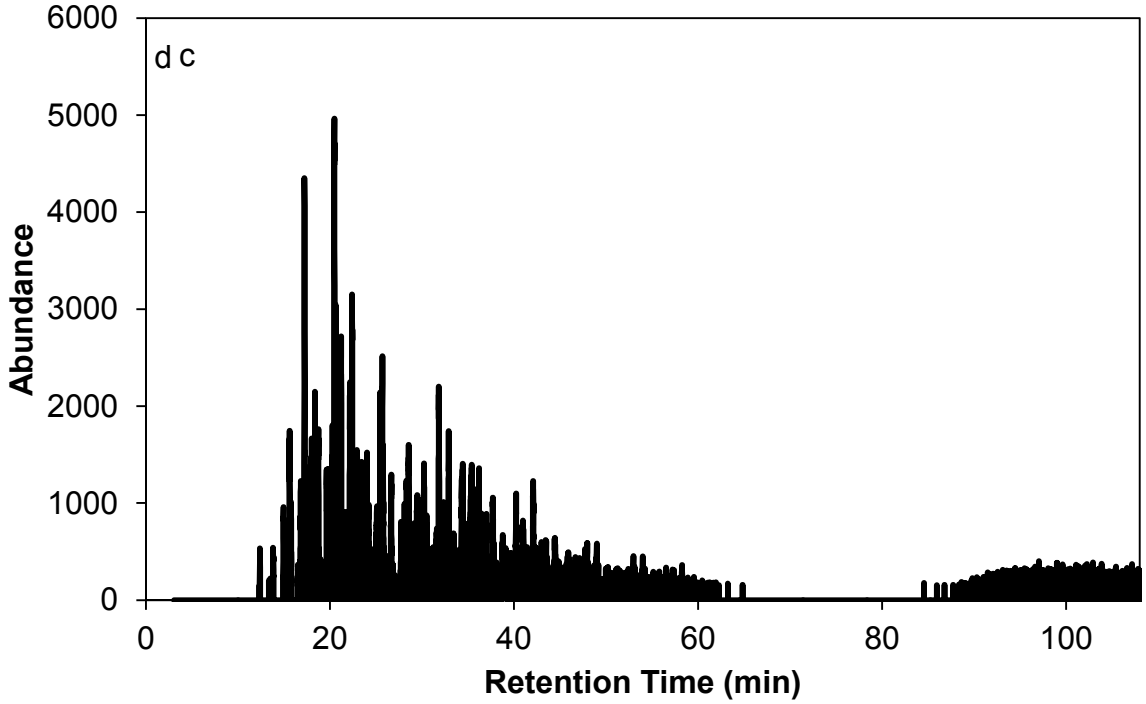


Figure 4-2. Extracted ion chromatograms for ions present in the last mass spectral scan (from Figure 4-1) including mass-to-charge (m/z) 73 (a), m/z 96 (b), m/z 133 (c), m/z 191 (d), m/z 207 (e), m/z 208 (f), m/z 209 (g), m/z 281 (h), and m/z 282 (i). The extracted ion profile generated from these ions is also shown (j)

Figure 4-2 (cont'd).



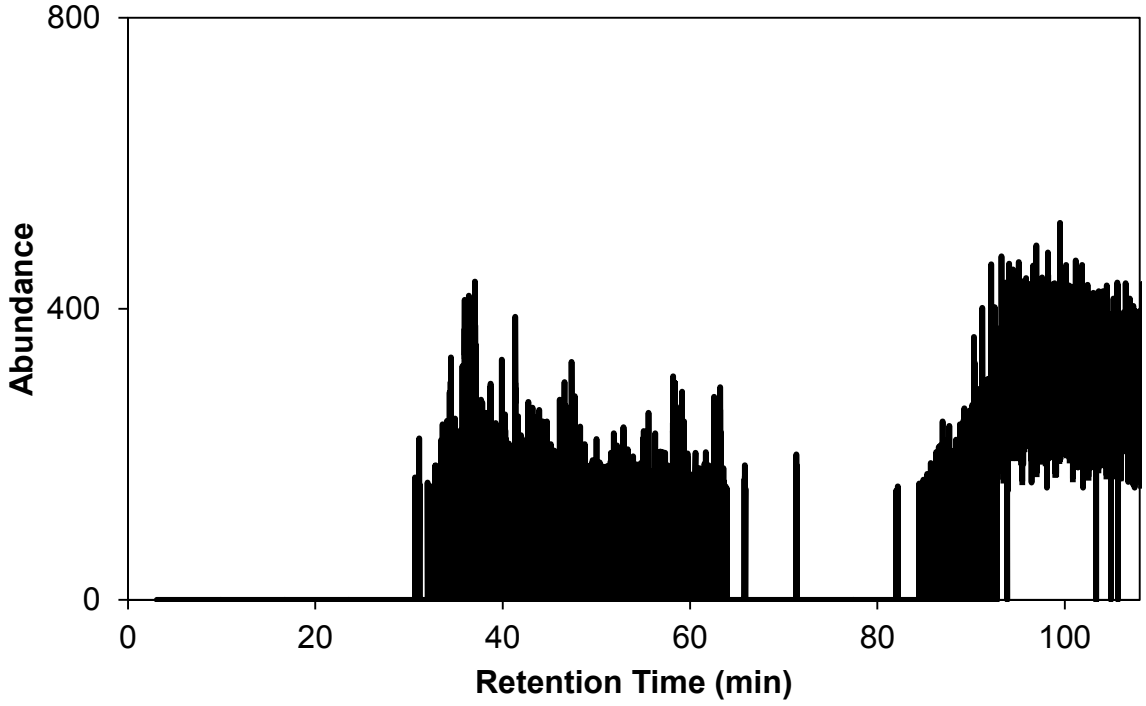
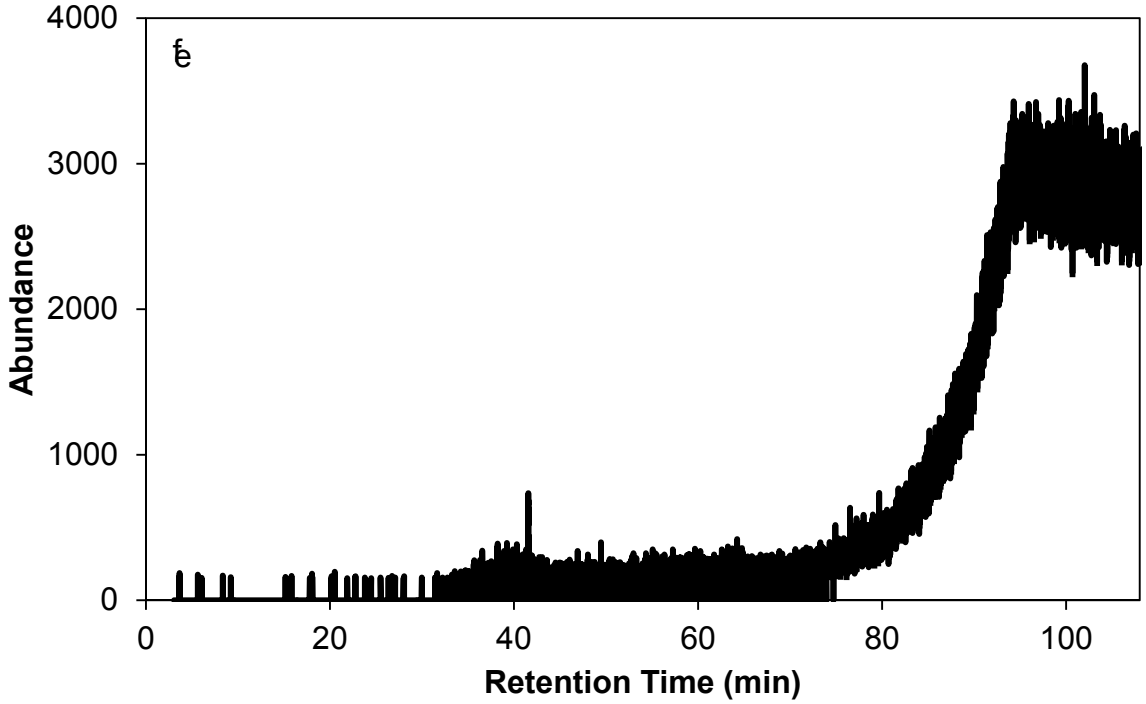


Figure 4-2 (cont'd).



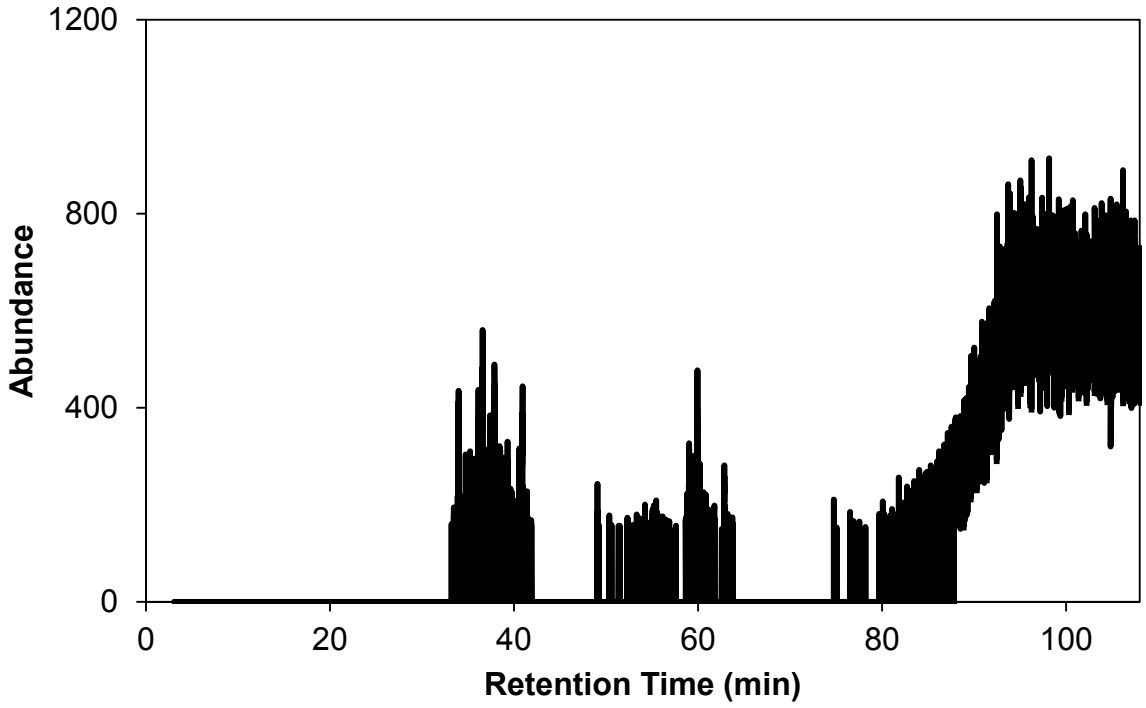
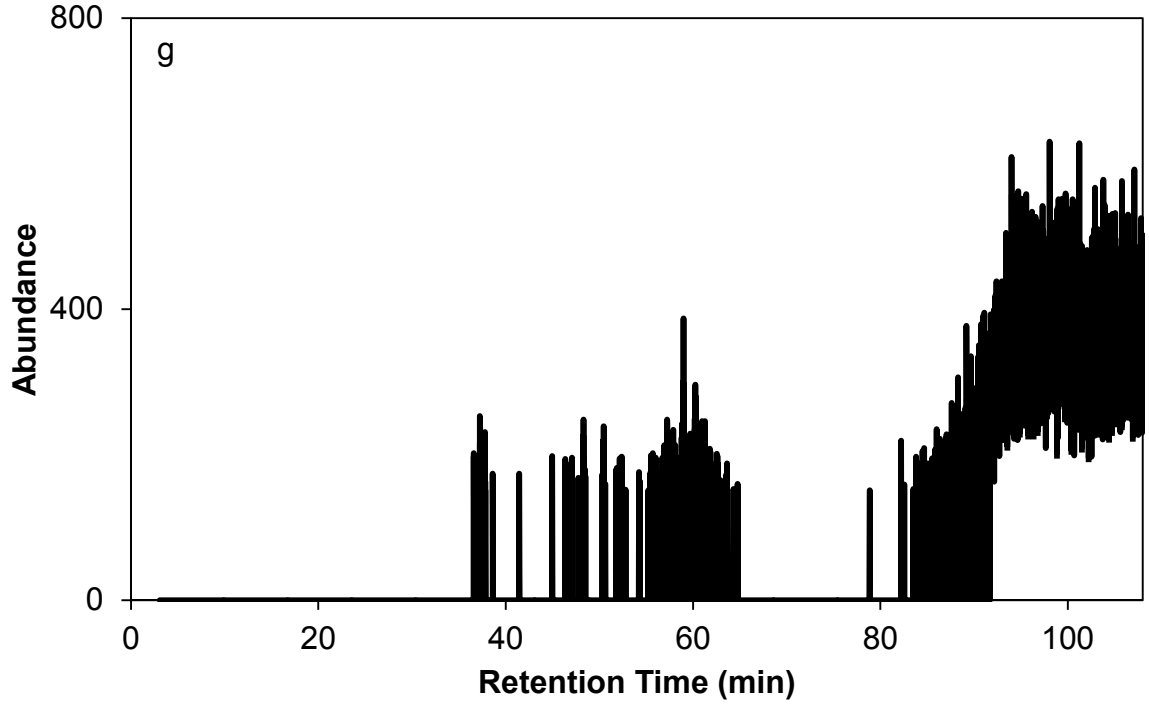


Figure 4-2 (cont'd).



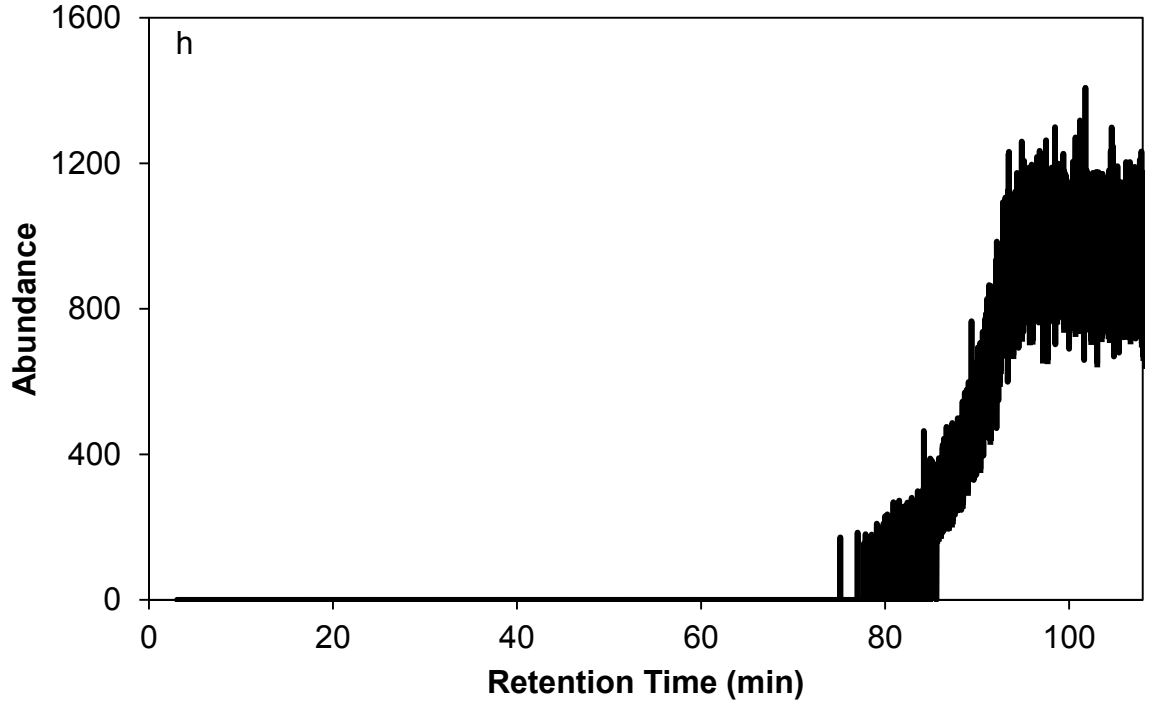
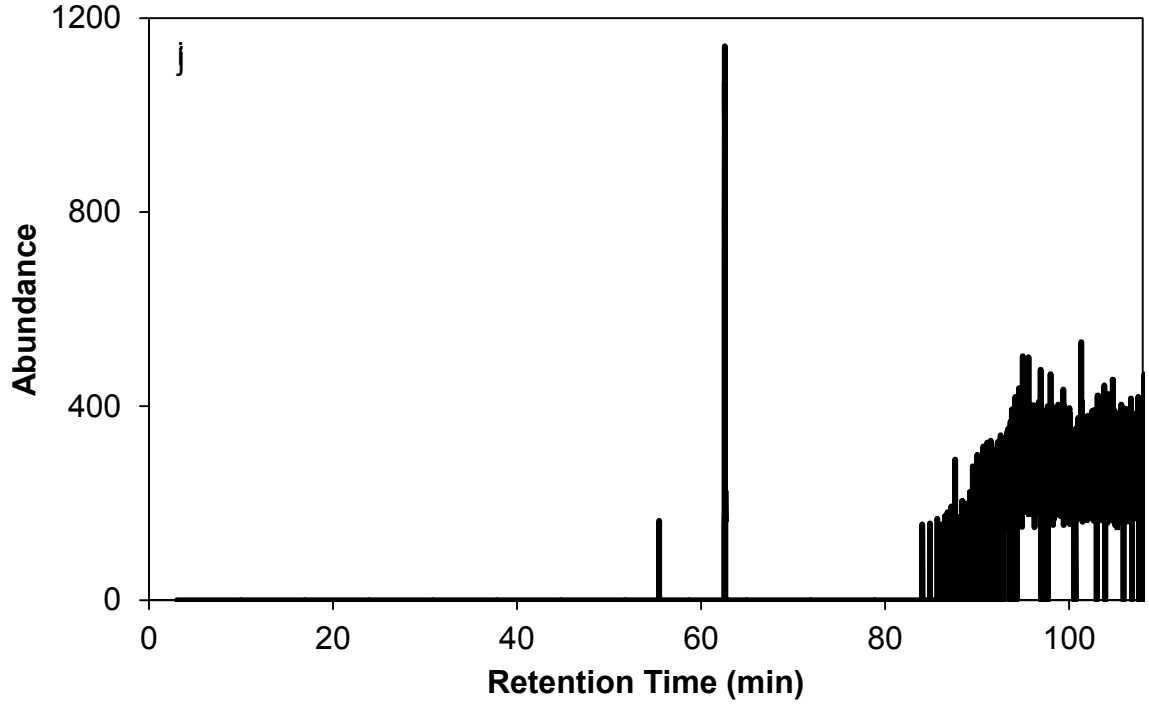
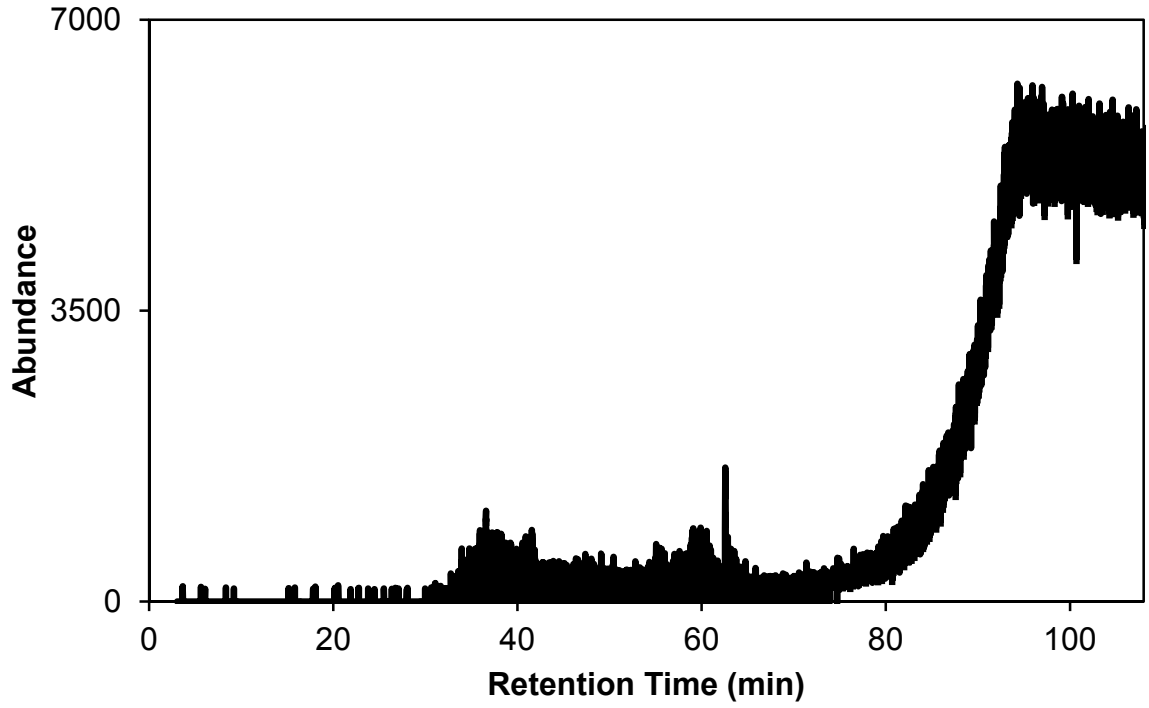


Figure 4-2 (cont'd).





(*m/z* 191, *m/z* 207, *m/z* 208, *m/z* 209, *m/z* 281, and *m/z* 282) were exported from the Chemstation software and imported to Excel (Office 2013, version 15.0, Microsoft Corporation, Redmond, WA). Once in Excel, the EICs were summed at each retention time to generate an EIP (Figure 4-2j). The EIP was then subtracted from the original TIC for each chromatogram of diesel fuel.

4.2.3. Subtraction of a the Baseline using a Modeled Function

A third method for baseline correction was developed and tested in which a model was generated to fit the baseline, using the EIP of the baseline described in Section 4.2.2. When the EIP was generated for the previous method, some ions were not included because those ions also resulted from the fragmentation of compounds in the sample. This resulted in incomplete removal of the baseline. In order to model the EIP, the EIP for each diesel chromatogram was fit in TableCurve 2D (Version 5.01, Jandel Scientific, San Rafael, CA) in order to determine an appropriate equation for the model. An asymmetrical sigmoid function (Equation 4-1) was selected based on the highest coefficient of determination (r^2) value when used to fit a solvent blank. As the solvent blank is the shape of the baseline that is being removed without any other signal present, this can be used to identify the appropriate function to model the EIP of each chromatogram.

$$y = a + \frac{b}{\left[1+\exp\left(-\frac{x-c\ln\left(2^{1/e}-1\right)-b}{c}\right)\right]^{d}} \qquad \text{Equation 4-1}$$

The *y* term is the resulting abundance at each retention time, *x*. The *a* term is the initial height of the function and the *b* term is the transition height (the final height of the function subtracted from the initial height). The *c* term is the retention time that where the inflection point of the curve occurred, while the *d* and *e* terms control the shape of the curved portion of the function. An asymmetric sigmoid allows for different curvature at the top and the bottom of the function, allowing for more flexibility when fitting the baseline. An example of the modeled baseline is shown in Figure 4-3.

The baseline EIP for each chromatogram was imported into TableCurve 2D and fit to generate appropriate *c*-*e* terms. The *a* term was selected as zero so that no signal was removed from the beginning of the chromatogram The *b* term was the average value of the last eight minutes of the chromatogram. This region of the chromatogram is where the baseline is at highest abundance and generally at a constant value. The *b* term was determined from the TIC, rather than the EIP, in order to remove as much of the baseline as possible. The asymmetrical sigmoid function was then regenerated in Excel using Equation 4-1 and subtracted from the original chromatogram, generating a baseline corrected chromatogram. This was repeated for each chromatogram.

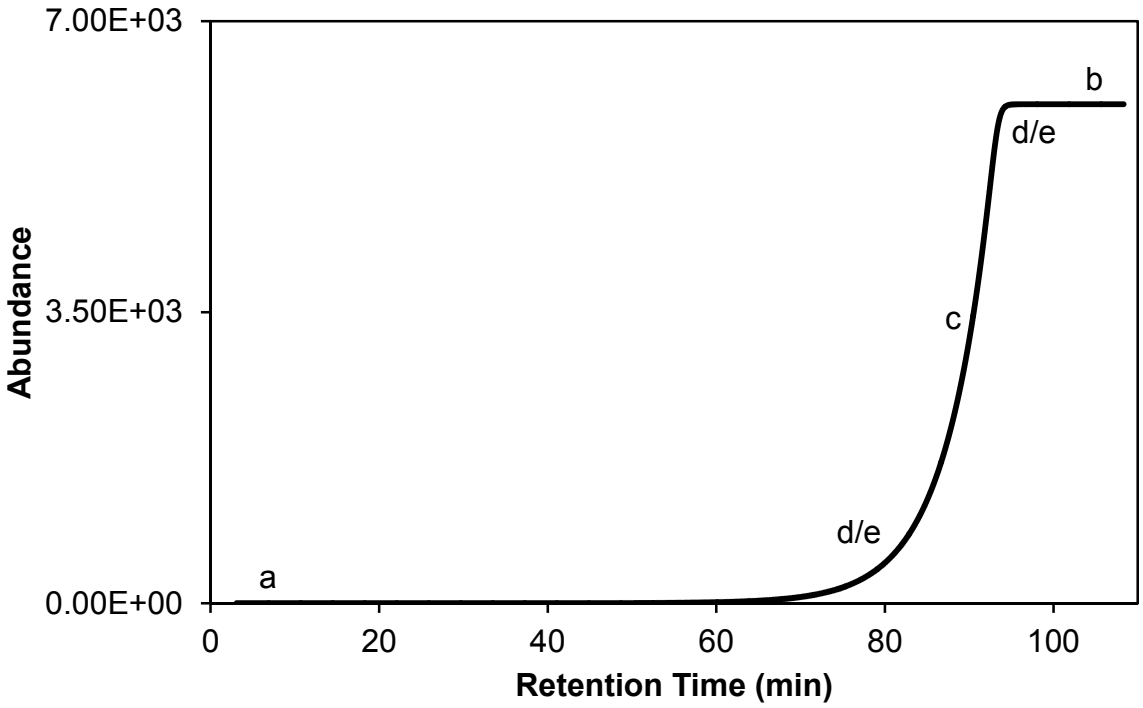


Figure 4-3. Model generated for baseline of a diesel chromatogram, based on Equation 4-1. The *a* term is the initial height of the function, the *b* term is the transition height, the *c* term is the retention time at which the inflection point of the curve occurred, and the *d* and *e* terms control the shape of the curve.

4.2.4. Evaluation Metrics

Chromatograms were visually examined to assess reduction in baseline as a result of correction. The reduction in the baseline was also quantitatively evaluated using the last 40 minutes of the chromatogram, where the rise in the baseline occurs. Ideally, the baseline in this region should be zero. The sum of squares of the abundance for all points in this baseline region was used to measure the magnitude of the baseline. The percent change in the magnitude of the baseline before and after baseline correction was then used to quantitatively compare each baseline reduction method. Visual inspection of the PCA scores plot and average percent change in the clustering of replicates (PCC) was also applied as previously discussed.

4.3. Effect of Baseline Correction on Chromatographic Data

4.3.1. Visual Assessment

Prior to applying any baseline correction method, the chromatograms were overlaid and only very slight differences were observed in the baseline among the sample chromatograms. The percent difference between the average signals in this region was less than 10%, indicating that there were only small differences in the baseline. While it is important to minimize non-chemical sources of variation, this variation is not likely to have a major impact on the resulting PCA.

4.3.1.1. Background Subtracted Baseline

The EICs of each *m/z* present in the baseline of a diesel chromatogram (Figure 4-2) demonstrates the largest drawback with the BSB subtraction. Because the operator

cannot select which *m/z* to remove, all *m/z* in a single scan will be removed. A reconstructed chromatogram, showing the signals that are subtracted using the BSB method is shown in Figure 4-4a. This method resulted in a reduction of the signal in the TIC within the peak region, due to the signals between 30 and 60 minutes in Figure 4-4a. Figure 4-5a shows a representative diesel chromatogram prior to any pretreatment, with an insert showing an expended region of the chromatogram from approximately 70 to 108 minutes, where the rise in baseline occurs. Figure 4-5b shows the same chromatogram after applying the BSB method (removing the signals shown in Figure 4-4a). As shown in Figure 4-5b, after subtraction with the BSB method, not all of the baseline is removed. Therefore, this method removes signals arising from chemical differences in the samples, while not completely removing the baseline.

4.3.1.2. Subtraction of Extracted Ion Profiles

The subtraction of an extracted ion profile allows for the selection of ions to include in the subtraction, permitting the analyst to tailor the subtraction to the sample, thereby overcoming the main limitation in the BSB method described above. In this work, the ions chosen for the baseline EIP (from Section 4.2.2) were selected based on having a large contribution to the rise in the baseline in the noise region but low abundance in the peak region (Figure 4-2). The EIP subtracted from the TIC using this method is shown in Figure 4-4b. Even when selecting specific ion to remove, there is still a reduction in the peak

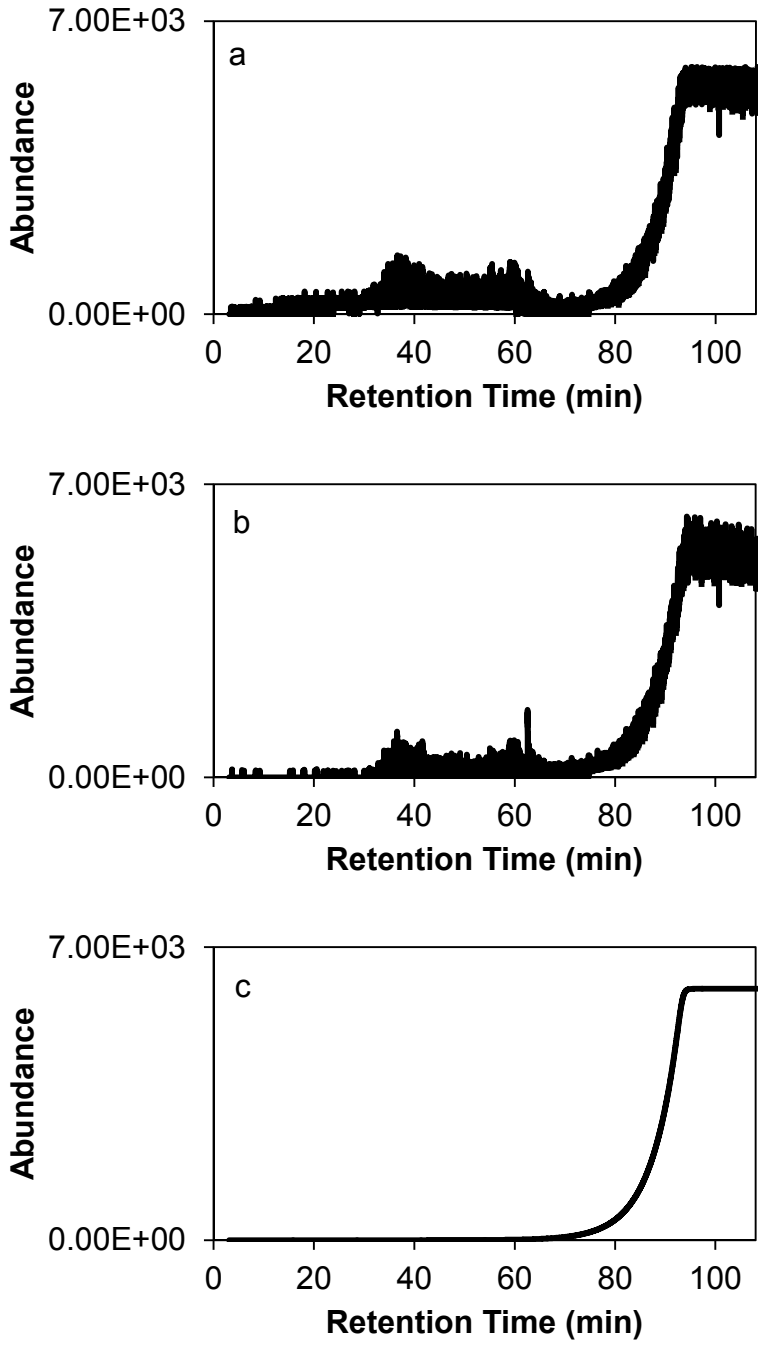


Figure 4-4. The signal that was subtracted from the TIC using the BSB method (a), the EIP (b), and the function fit by the EIP (c)

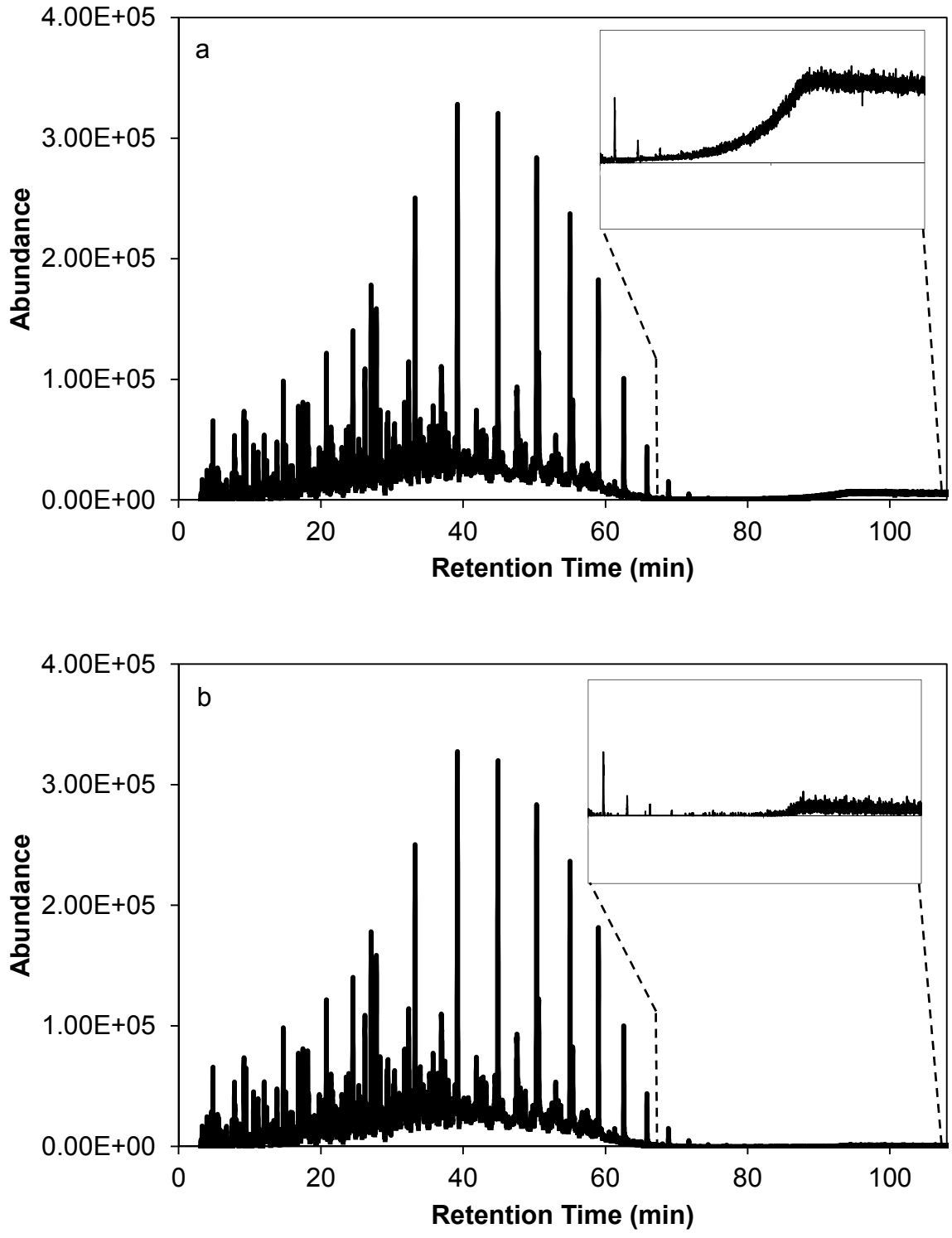
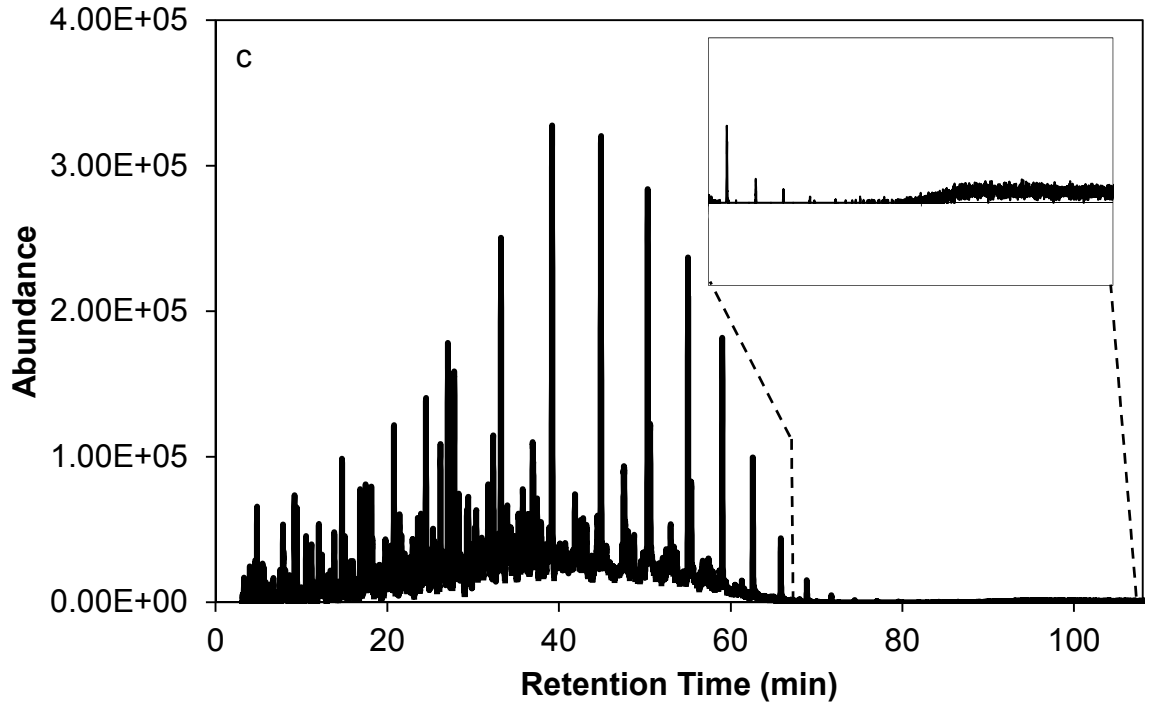
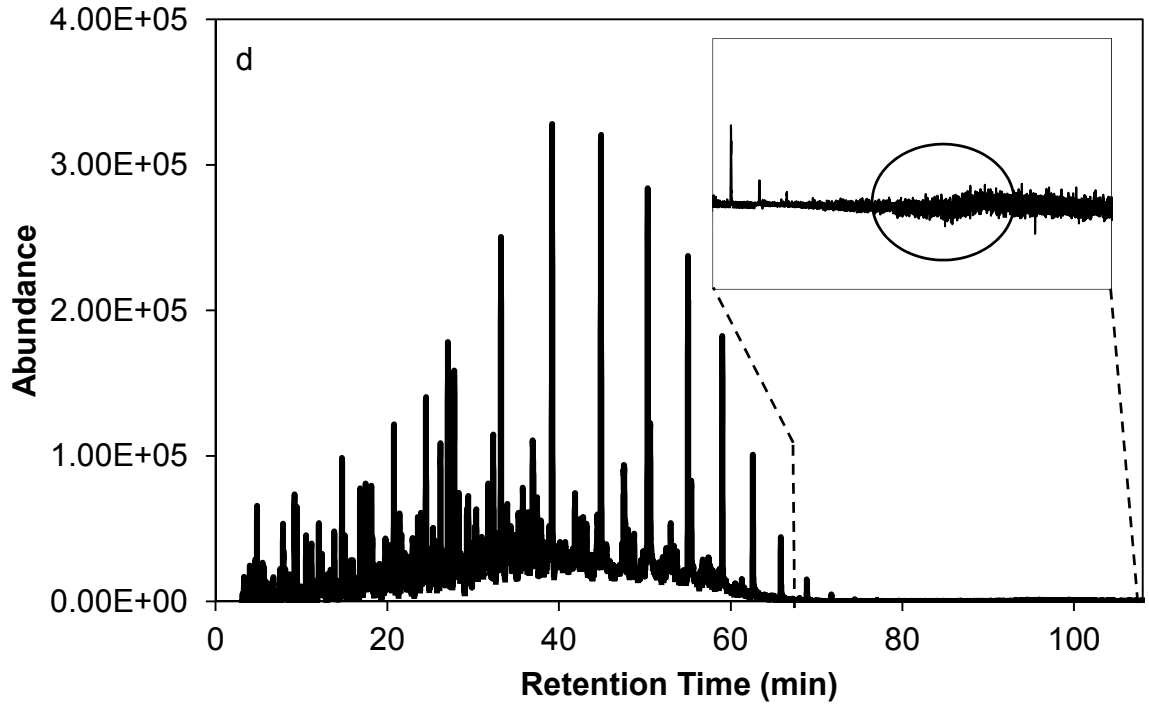


Figure 4-5. The baseline of the TIC before pretreatment (a) and pretreatment using the BSB method (b), the EIP (c), and the function fit by the EIP (d).

Figure 4-5 (cont'd)





region in the middle of the chromatogram. Figure 4-5c shows a diesel chromatogram after subtracting the EIP (removing the signals shown in Figure 4-4b). The subtraction of the EIP also did not completely remove the baseline (Figure 4-5c). More of the baseline remains than when using the BSB method, because several *m/z* were excluded from the EIP. Ions formed from the fragmentation of compounds in the sample will also be removed from the chromatogram, resulting in the observed signal reduction in the peak region. Therefore, for these particular data, subtraction of the EIP is also not effective at baseline correction.

4.3.1.3. Subtraction of a the Baseline using a Modeled Function

The modeling of the baseline also provides the analyst with control over the removal of signals from different regions of the chromatogram. However, the parameters for fitting or modeling the baseline must be determined, which is more labor intensive than other methods. Figure 4-4c shows an example of the modeled baseline that was removed from each chromatogram. Using this method, there is no reduction of signal in the peak region. Figure 4-5d shows a representative diesel chromatogram after subtraction of the modeled baseline. This method results in a more complete removal of the baseline. However, there is a small artifact at approximately 80 minutes (circled in Figure 4-5d) due to improper modeling of the baseline. The *c*, *d*, and *e* terms in the asymmetric sigmoidal fit could be further optimized to reduce this artifact, but would be labor intensive. For these data, this artifact will have little influence in future analyses because of its low abundance.

This method is the only one of the methods investigated that results in both positive and negative baseline signal. The point-to-point fluctuation observed in signal is noise. This is important for evaluation of smoothing, discussed in Chapter 5. In addition, this method did not result in a reduction of signal in the peak region and resulted in the most complete reduction in the baseline.

4.3.2. Quantitative Assessment

As expected from the visual assessment of chromatograms, similar percent changes in the sum of squares of the baseline regions were observed for each of the baseline correction methods that were tested. The BSB method resulted in a 90% reduction in the magnitude of the baseline, and the subtraction of the EIP resulted in an 88% reduction in the baseline compared to the non-corrected chromatograms. The subtraction of the baseline using a modeled baseline EIP resulted in a 92% reduction in the baseline. Because the BSB method contained more ions than the subtracted EIP, the BSB method resulted in a larger reduction in the baseline than the subtraction of the EIP. More ions could have been included in the EIP to increase the reduction of the baseline; however, this could also result in a larger reduction of the signal in the peak region. Employing the fitted EIP allowed for the greatest reduction, due to the overall abundance (*b* term) being from each TIC. Additionally, this method reduces the chance of removing chemical signal as a result of the baseline subtractions.

4.4. Effect of Baseline Correction on PCA Scores Plot

4.4.1. Visual Assessment

Figure 4-6 shows the scores plot for replicates of the eight diesel samples prior to any pretreatment (a), and after baseline correction only (b). There are no differences observed in the positioning of samples on the scores plot after baseline correction. The loadings plots for PC1 and PC2 (Figure 4-7a and Figure 4-7b, respectively) also show no difference before (Figure 2-7) and after baseline correction. In this work, even though the baseline was elevated across a large portion of the chromatogram, this elevation was consistent between samples, and was not a major source of variation. This is likely due to the short period of time over which the samples were analyzed. As expected, because there was no change in the loadings plot, there was also no change in the positioning of samples on the scores plot.

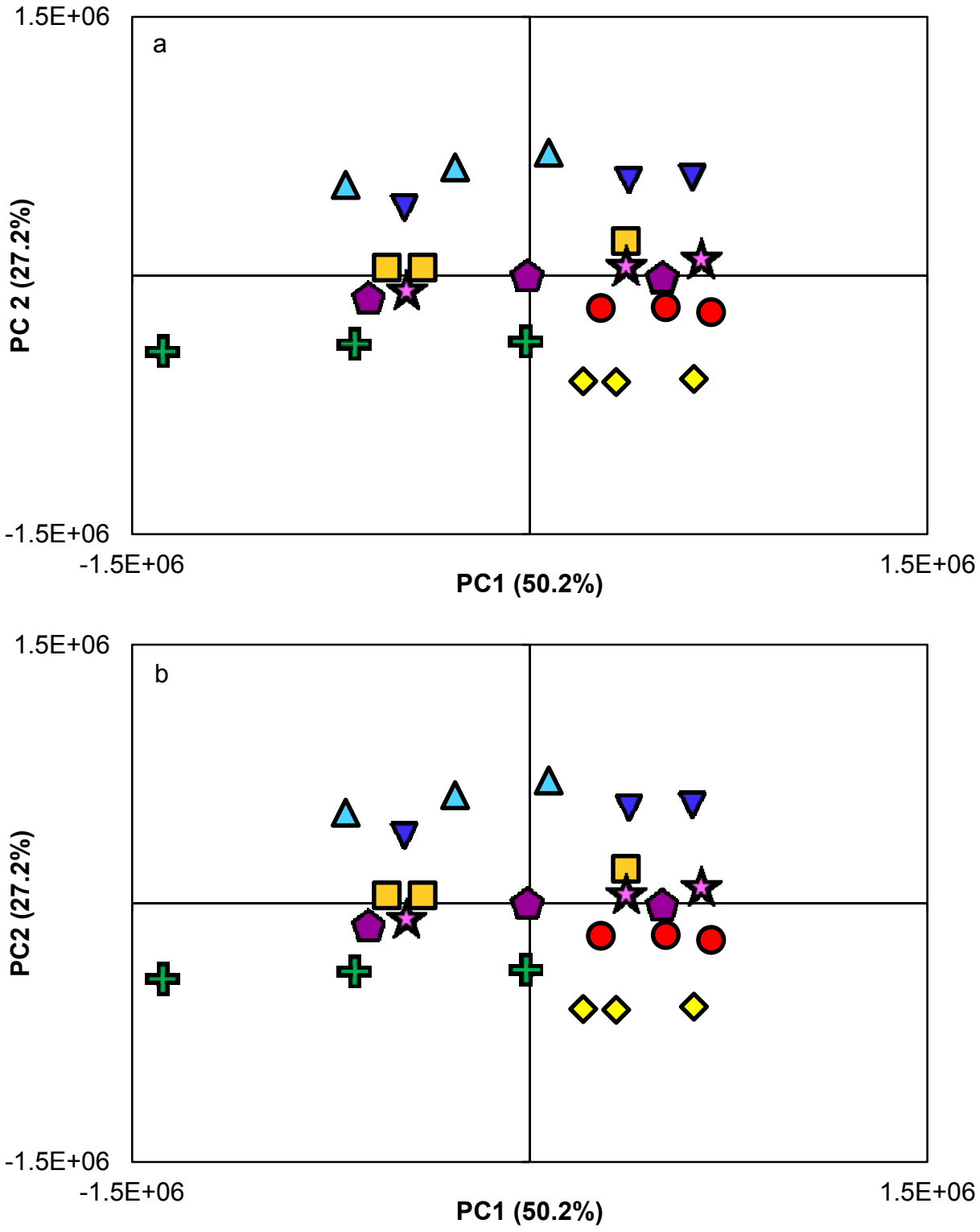


Figure 4-6. Scores plots of eight diesels in triplicate without any pretreatment (a) and after baseline correction (b).

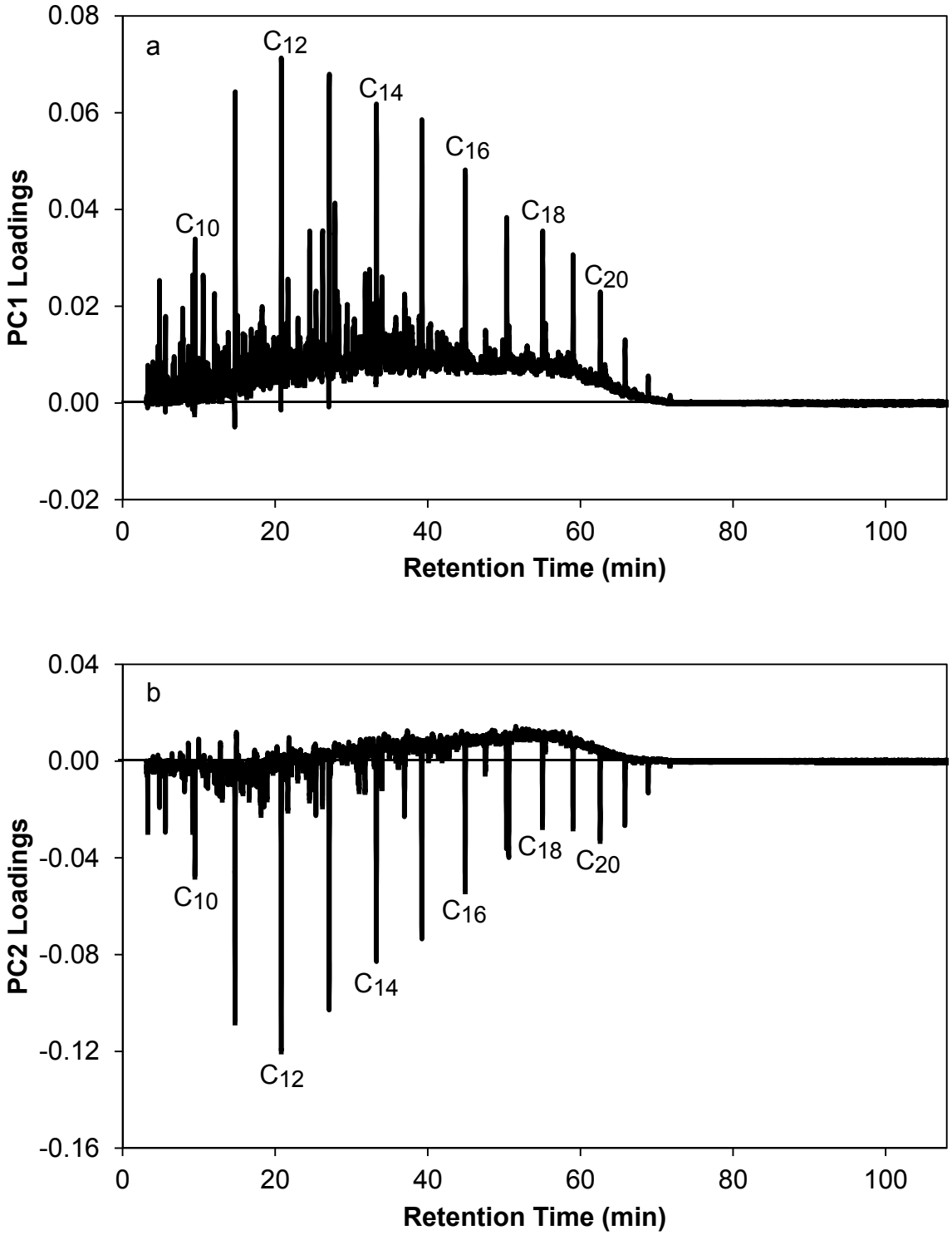


Figure 4-7. Loadings plot for PC1 (a) and PC2 (b) after baseline correction.

PCA was then performed after baseline correction and total area normalization of the data. Figure 4-8 shows the scores plot after area normalization only (a) and after baseline correction followed by area normalization (b). All of the samples were slightly shifted on the scores plot after baseline correction and normalization. This resulted in a few samples appearing to be better clustered (*i.e.* Diesel 10, pink stars), while other samples appeared to be less clustered (*i.e.* Diesel 4, orange squares). The loadings plots for PC1 and PC2 after baseline correction followed by normalization are shown in Figure 4-9a and Figure 4-9b, respectively. Compared to the raw data (Figure 2-7), there is a slight decrease in the baseline in the loadings plots of both PC1 and PC2. The small differences in the loadings plots after baseline correction explains why only small shifts in the samples were observed. After applying baseline correction, the percent variance for PC1 increased from 57.9% to 58.9%, while the percent variance for PC2 decreased from 21.4% to 21.1%. The total percent variance accounted for using the first two PCs increased form 79.3% to 80.0%. In general, the increase in percent variance accounted for using PC1 and PC2 is important because the more variance accounted for on the first two PCs, the more random error is being removed. The few changes in the PCA results after baseline correction indicate that the baseline is not a major source of variation in these data.

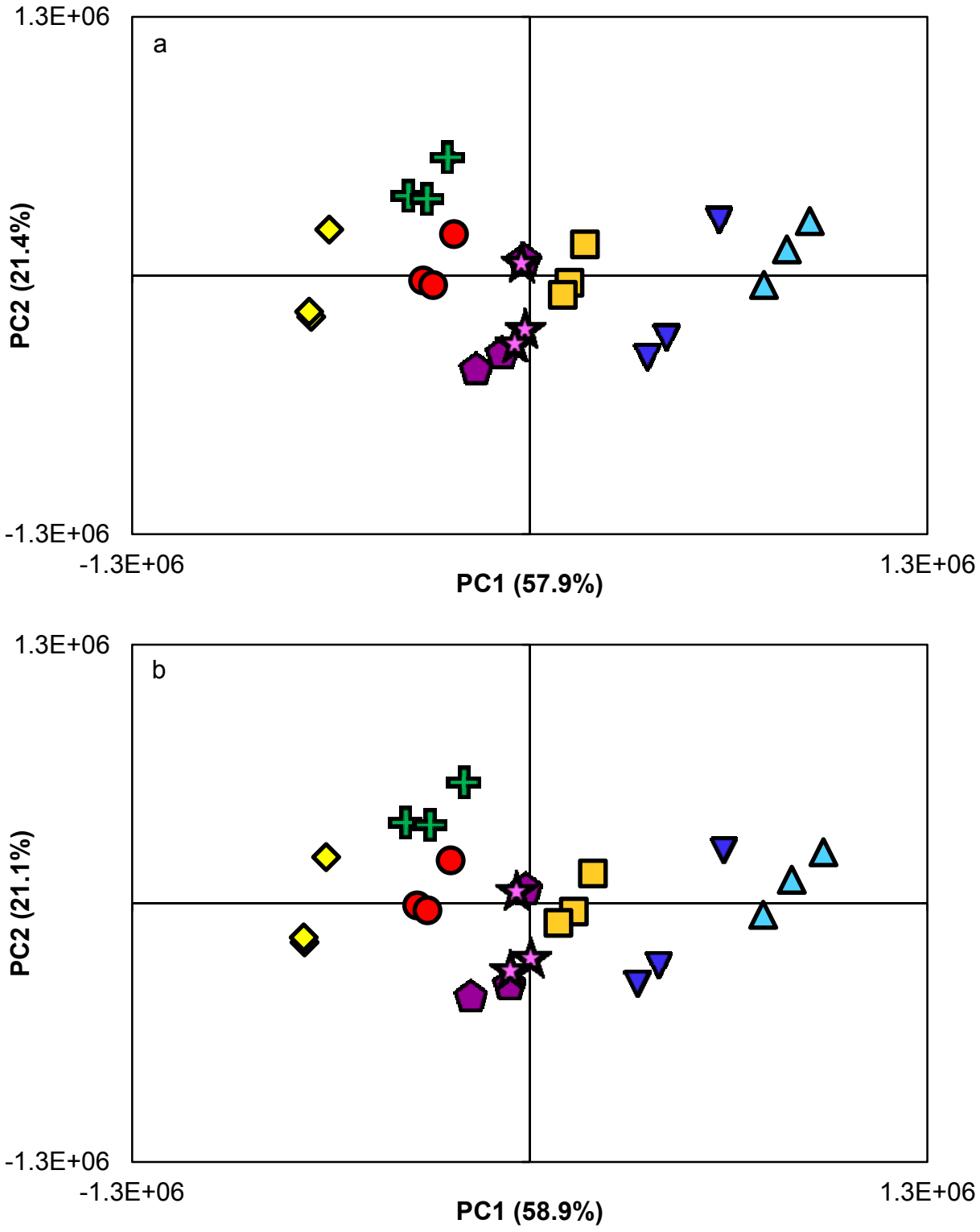


Figure 4-8. Scores plots of eight diesels in triplicate after total area normalization (a) and after baseline correction followed by total area normalization (b).

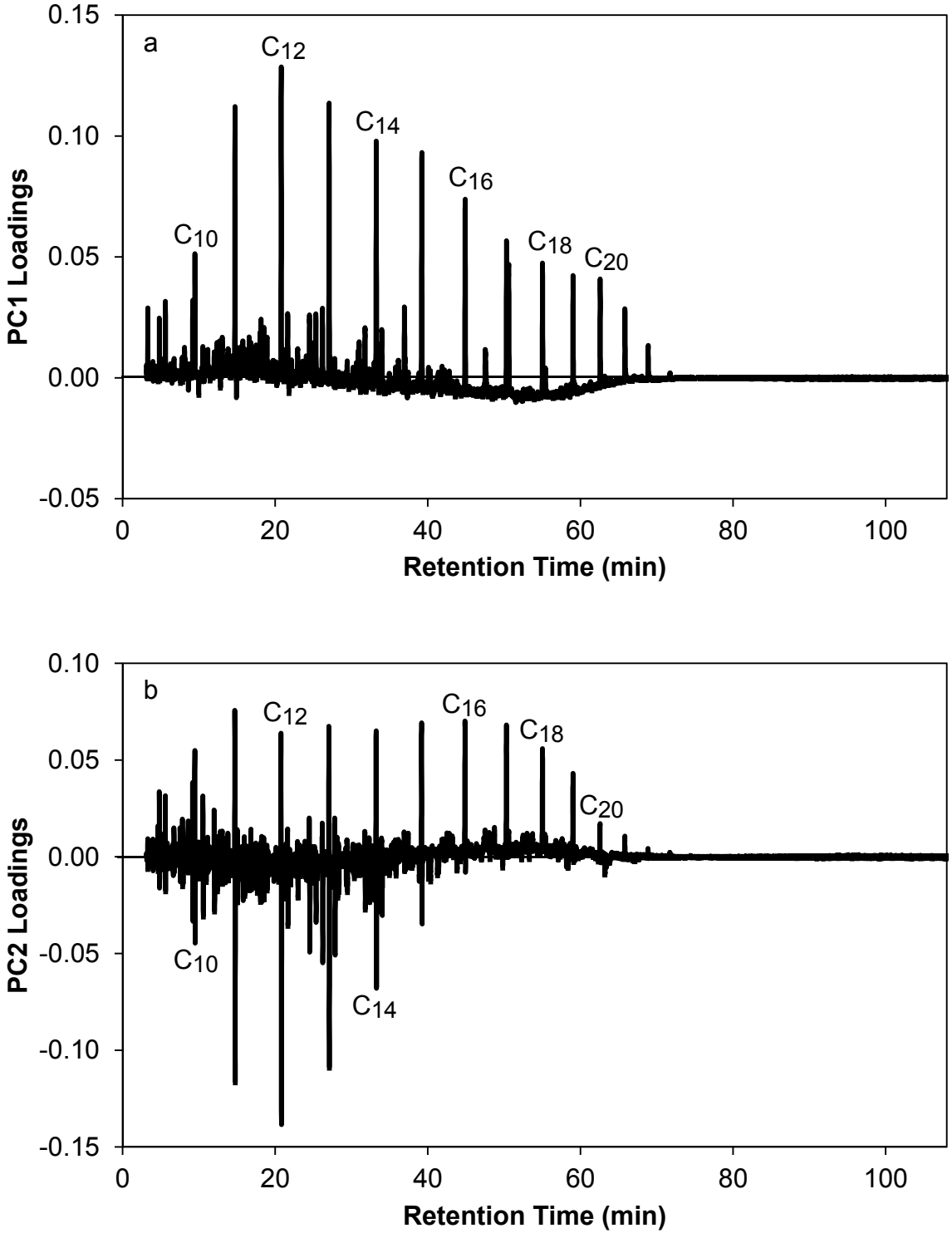


Figure 4-9. Loadings plot for PC1 (a) and PC2 (b) after baseline correction and area normalization.

4.4.2. Quantitative Assessment

Based on the percent change in the clustering of replicates (PCC), baseline correction alone had no measurable effect on the variance of replicates samples in the PCA scores plot (Table 4-1). However, when used in conjunction with either normalization method, there was a slight decrease in the PCC, indicating that baseline correction resulted in slightly poorer clustering of replicates. The PCC using only area normalization was 45.1% and 59.0% for peak normalization. After baseline correction, the PCC decreased slightly to 44.6% and 58.7%, respectively. As previously discussed, baseline correction may not be necessary for this work. However, this is not always the case. If different chromatographic columns were used or the samples were analyzed over a longer time period, there may be greater variability in the baseline. However, in this work, the samples were collected over a few weeks and only small, insignificant variation was observed in the baseline. Also, baseline correction may have been necessary if peaks eluted during the rise in the baseline.

Table 4-1. The average percent change in the clustering (PCC) of replicates after the listed pretreatment procedures including baseline correction using the extracted ion profiles (EIP fit) and normalization using total area (Area) and single peak (Peak) normalization methods.

Baseline Correction	Normalization	PCC
-	Area	45.1
-	Peak	59.0
EIP fit	***-***	***0.0***
EIP fit	***Area***	***44.6***
EIP fit	***Peak***	***58.7***

4.5. Summary

The baseline extends over a large number of data points in the chromatogram and therefore could be a major source of variation in PCA, even though the baseline is often small compared to the peaks. In this work, there was little quantifiable difference in the baseline correction methods examined. Additionally, baseline correction was shown to have little effect on the clustering of replicates when analyzed using PCA. This is likely due to the similarity of the rise in the baseline for the different diesel samples tested. Baseline correction may be useful when applied to other datasets. Based on the three methods evaluated in this work, the fitted EIP to remove the baseline provides the analyst with the most control when selecting the signals to remove.

REFERENCES

REFERENCES

[1] K.M. Pierce, J.S. Nadeau, R.E. Synovec, in: C.F. Poole (Ed.), Gas Chromatography, Elsevier, Waltham, MA, 2012.

[2] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, Inc., New York, NY, 1998.

[3] M. Daszykowski, B. Walczak, Trac-Trends Anal. Chem., 25 (2006) 1081.

[4] S.L. Morgan, E.G. Bartick, in: R.D. Blackledge (Ed.), Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis, John Wiley & Sons, Inc., Hoboken, NJ, 2007.

[5] M.C. McMaster, GC/MS: A Practical User's Guide John Wiley & Sons, Inc., Hoboken, NJ, 2008.

[6] Agilent Technologies Inc., Agilent Technologies, Inc, 2011.

CHAPTER 5: SMOOTHING

5.1. Introduction

Noise, or point-to-point fluctuations in signal, is another source of non-chemical variation. Like the baseline, the noise must also be minimized to allow for comparison of chemical variations between samples [1]. The goal of smoothing is to minimize the random fluctuations in the chromatogram without distorting the chemical signal. Smoothing methods can be classified into two general categories: running and filtering smoothers [1-4]. Running smoothers remove point-to-point fluctuation using the data points around a central point (called a window). The position of the central point is calculated using an average of the data points in the window (called a boxcar smooth) or by fitting the data points in the window using a polynomial function (called the Savitzky-Golay smooth) [3, 4]. The center point is incremented along the chromatogram and the process is repeated for each point, resulting in a smoothed chromatogram. The filtering smoothers removes specific signals from the chromatogram. The most common example is the fast Fourier transform smoother, which filters the high-frequency signals from the chromatogram. Noise is rapid changes in signal that occurs from point to point and therefore is high frequency in nature [3].

5.2. Methods Tested and Evaluation Metrics

One smoothing algorithm from each of the general types of smoothers was compared. The Savitzky-Golay (SG) smooth was utilized as the running smoother and the fast Fourier transform (FFT) smooth served as the signal filtering smoother. These smoothing algorithms were selected due to their popularity and wide availability [1, 3, 4].

Many commercially available data analysis and chemometric software packages contain one or more of these smoothing algorithms. Origin Pro (version 7.5 OriginLab Corporation, Northampton, MA) contains both the Savitzky-Golay and the fast Fourier transform and was used to compared both algorithms using a single diesel chromatogram. TICs were exported from Excel after baseline correction and imported into Origin for smoothing.

5.2.1. The Savitzky-Golay Smooth

The Savitzky-Golay algorithm uses a moving average with a least-squares polynomial equation to fit the chromatogram [5]. The order of the polynomial and the number of data points in the smooth can be varied. For this work, different combinations of polynomial order and number of points were investigated. The order of the polynomial ranged from 1 to 6 and the total number of data points varied from 3 to 25, with equal number of points on each side of the central point. Only even-order polynomials (after the first order) were considered, as the central point smooth results in the equivalent smoothing for even and the following odd-order polynomial [6]. The SG smoothing algorithms were applied in Origin and exported back to Excel for further investigation.

5.2.2. The Fast Fourier Transform Smooth

To apply a FFT smooth, the data are transformed from the time domain to the frequency domain using a fast Fourier transform. Then, a low-pass filter is applied to the data in the frequency domain to remove the high-frequency noise component. The point at which the filter is applied is called the cutoff frequency. The cutoff frequency for the low-pass filter is inversely related to the number of points in the chromatogram [7], at a fixed scan rate. This means that the number of points in the chromatogram, which is

affected by the temperature program, the solvent delay, and the scan rate, will result in different degrees of smoothing. Many software packages favor the running smoothers, because their performance is not dependent on the number of points or scan rate. For this work, FFT smooth from 1 to 10 points was applied in Origin, corresponding to cutoff frequencies between approximately 2.91 and 0.29 Hz.

5.2.3. Metrics Used for Evaluation

The performance of each smoothing algorithm was evaluated based on the signal enhancement and extent of peak distortion. The signal enhancement was quantitatively measured by calculating the percent change in the noise and the signal-to-noise ratio in the TIC before and after smoothing. The standard deviation of the last 13 minutes of the TIC was defined as the noise (s_{noise}), because only noise is present in this region of the chromatogram. The signal-to-noise (S/N) was calculated by dividing the maximum abundance of the pentadecane peak (A_{C15}) in each TIC by the previously defined noise.

$$\frac{S}{N} = \frac{A_{C_{15}}}{s_{noise}} \qquad \text{Equation 5-1}$$

The percent change in the noise was also used to determine the degree of smoothing. A higher degree of smoothing resulted from a larger reduction in the noise.

Smoothing can result in peak broadening, which can be observed as a widening of the peak and a reduction in the peak height. EICs were used to determine peak distortion, as EICs provide better resolution and enhanced signal-to-noise, allowing for a more sensitive evaluation of peak distortion. In order to monitor peak distortion, the percent change in the peak height, peak width, and resolution before and after smoothing

were calculated. The maximum peak height of the pentadecane peak in the EIC of *m/z* 71 was utilized for the peak height. The pentadecane peak was selected because it was the largest peak in the chromatogram, resulting in a high signal. The ion with *m/z* 71 was selected because it was the highest abundance ion with baseline resolution for pentadecane. The peak width for pentadecane was determined using the second statistical moment (M_2), which measures the variance of the peak, using the abundance (A) at each retention time (t) in the peak [8, 9].

$$M_2 = \frac{\int_0^\infty t^2 A_t dt}{\int_0^\infty A_t dt} \qquad \text{Equation 5-2}$$

When peaks become broader, there is also a reduction in the resolution between peaks. The resolution (R_S), or separation between peaks, [10] was calculated between tetralin and pentylbenzene using the retention time (t) and width (w) of each peak [11].

$$R_s = \frac{2(t_2 - t_1)}{(w_1 + w_2)} \qquad \text{Equation 5-3}$$

In the TIC, these two peaks overlap; however, by utilizing EICs of *m/z* 132 (for tetralin) and *m/z* 148 (for pentylbenzene), the peaks can be separated, allowing an accurate calculation of resolution (Figure 5-1). The resolution of these peaks in Diesel 1a prior to

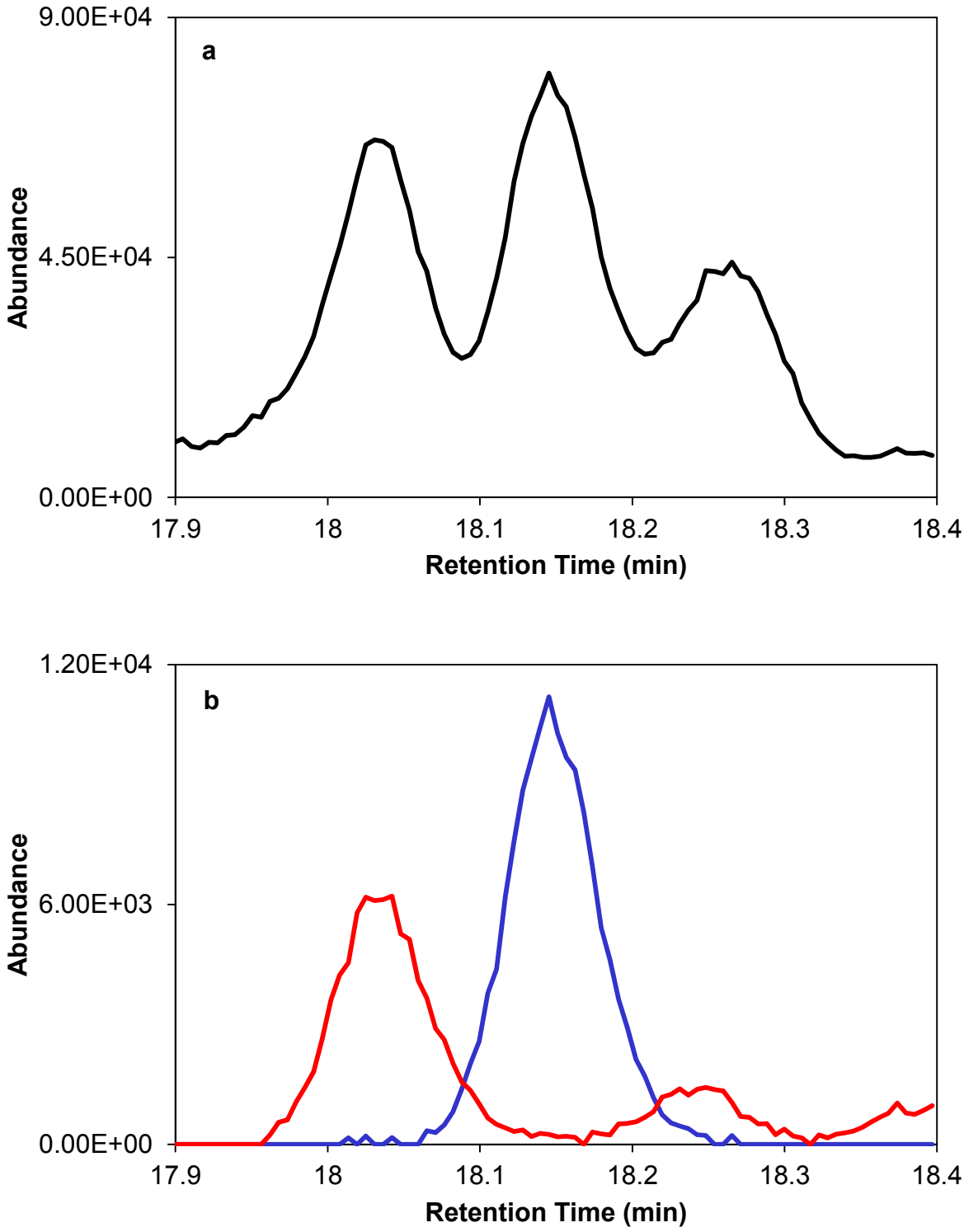


Figure 5-1. A representative diesel chromatogram showing the TIC (black) (a) and EICs (b) of m/z 132 for tetralin (blue) and m/z 148 for pentylbenzene (red).

smoothing is 0.9. Baseline resolution is 1.5. Ideally, smoothing should enhance signal-to-noise while only causing minimal peak distortion (*i.e*., minimal reduction in height and broadening of peaks, resulting in loss of resolution).

5.3. Effect of Smoothing on Chromatographic Data

5.3.1. Visual Assessment

The small fluctuations that are due to noise are often difficult to visually identify in the chromatogram. Figure 5-2 shows an expanded region of a diesel chromatogram before (a) and after (b) smoothing using a 2-point FFT smooth. The inset on the left shows an expanded view of the baseline. Prior to smoothing, the peak and baseline are jagged, showing point-to-point variation in the signal. After smoothing, the random fluctuations have been reduced. The inset on the right shows the end of the chromatogram where the signal is approximately constant and the point-to-point fluctuations are due to noise. After smoothing, the noise at the end of the chromatogram is also reduced. The higher degree of smoothing that is applied, the more the noise is reduced. However, the small differences resulting from different smoothing parameters were challenging to identify using visual assessment.

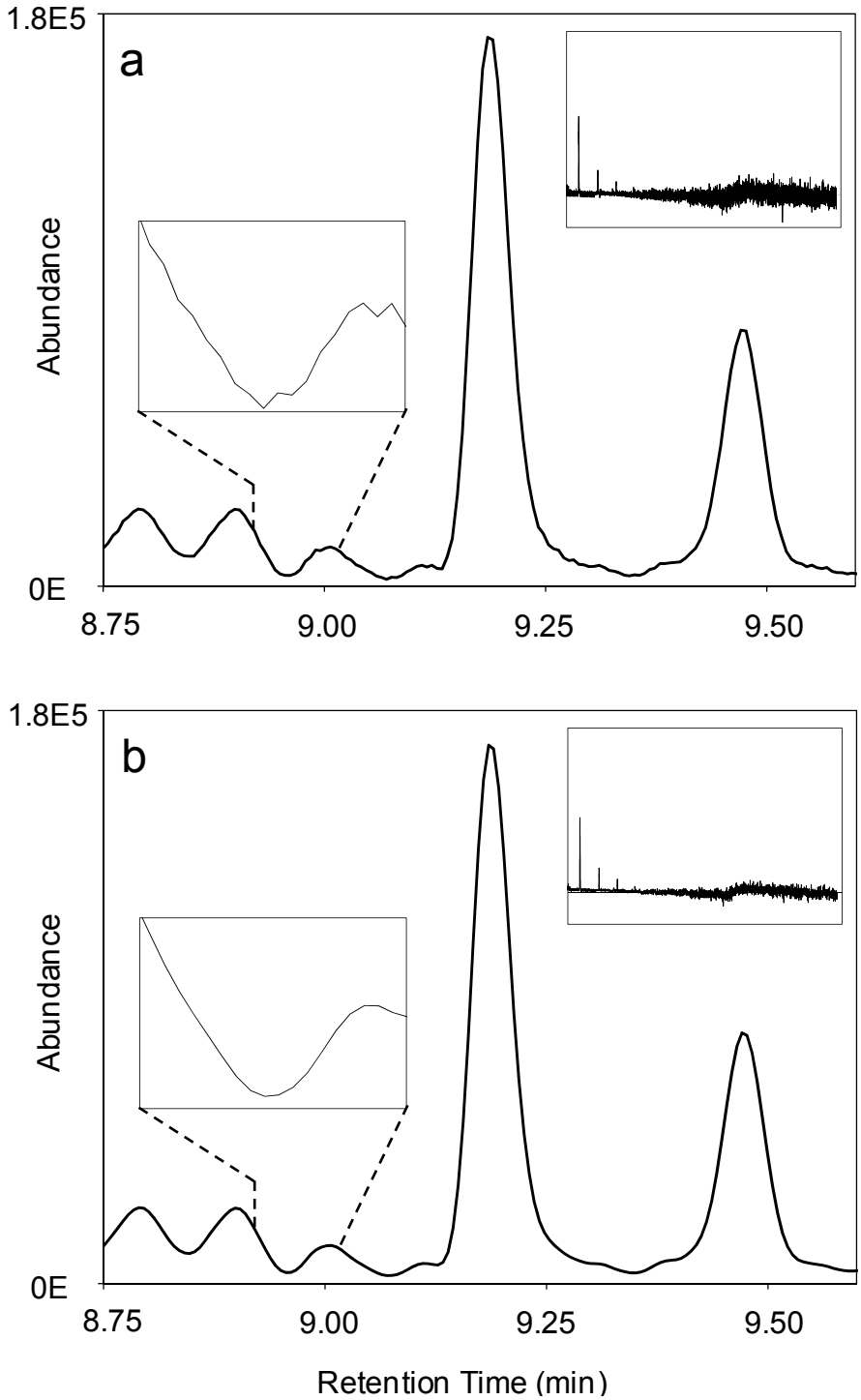
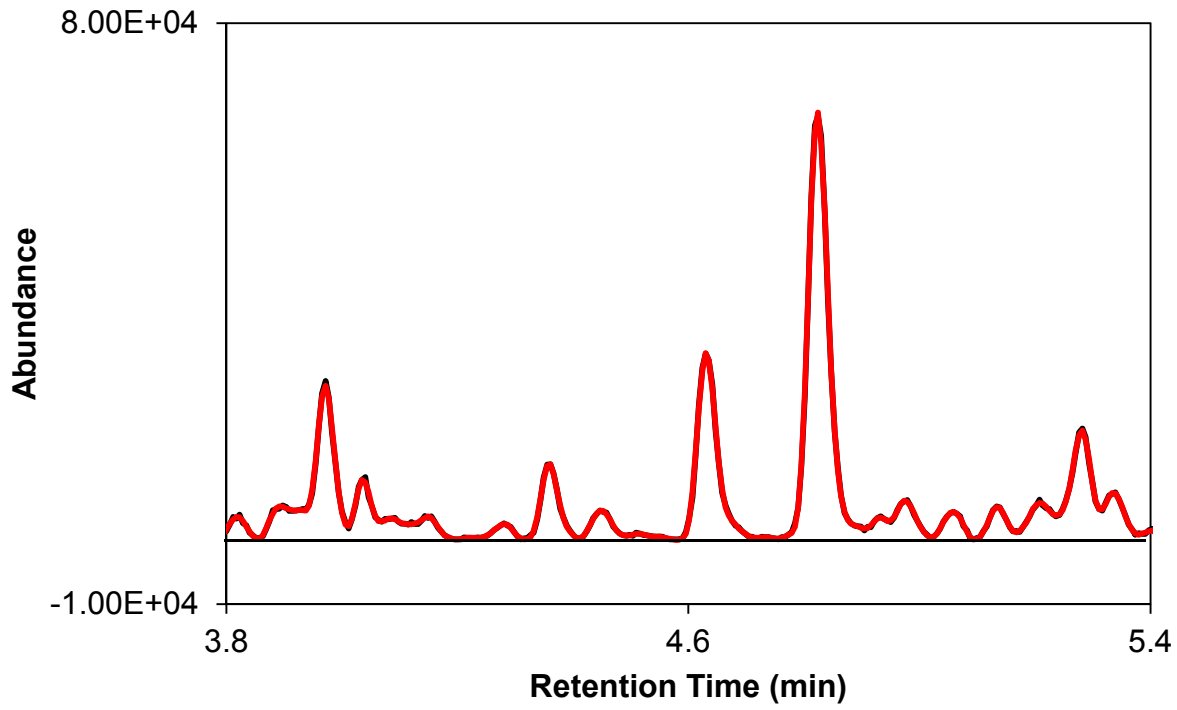


Figure 5-2. An expanded region of 1, 3, 5-trimethylbenzene in a representative diesel chromatogram after baseline correction (a) and after baseline correction and smoothing, using FFT 2 (b). The inset on the left is a further expanded region of the baseline, demonstrating the point-to-point variation before and after smoothing. The inset on the right shows the region at the end of the chromatogram, including the region defined as noise.

For some of the diesel samples, after smoothing had been applied, undesirable changes were observed in the chromatogram. At higher degrees of smoothing, reductions in peak heights and increases in peak widths were observed. These changes were often small, but could be identified after the chromatograms were overlaid. Additionally, artifacts were often observed on the edges of the peaks. Figure 5-3 shows an expanded region of Diesel 1a overlaid before (black line) and after smoothing (red line) with a Savitzky-Golay, 4^{th}-order polynomial with 11 total points (a) and a 6^{th}-order polynomial with 31 points (b). In Figure 5-3a, there were only slight differences observed between the unsmoothed and smoothed chromatograms, using a moderate level of smoothing. When a high degree of smoothing was applied, the peaks became broader and peak heights decreased. In addition, artifacts were usually observed near the edges of large peaks and appeared as valleys in the negative direction on either side of the peak. In Figure 5-3b, the artifacts are shown on the red trace between 4.40 and 4.90 minutes. These artifacts, which are characteristic of over-smoothing, are not easily identified using the metrics, so visual inspection may be still be necessary to ensure that over-smoothing is not occurring. These artifacts are not usually observed at low degrees of smoothing.



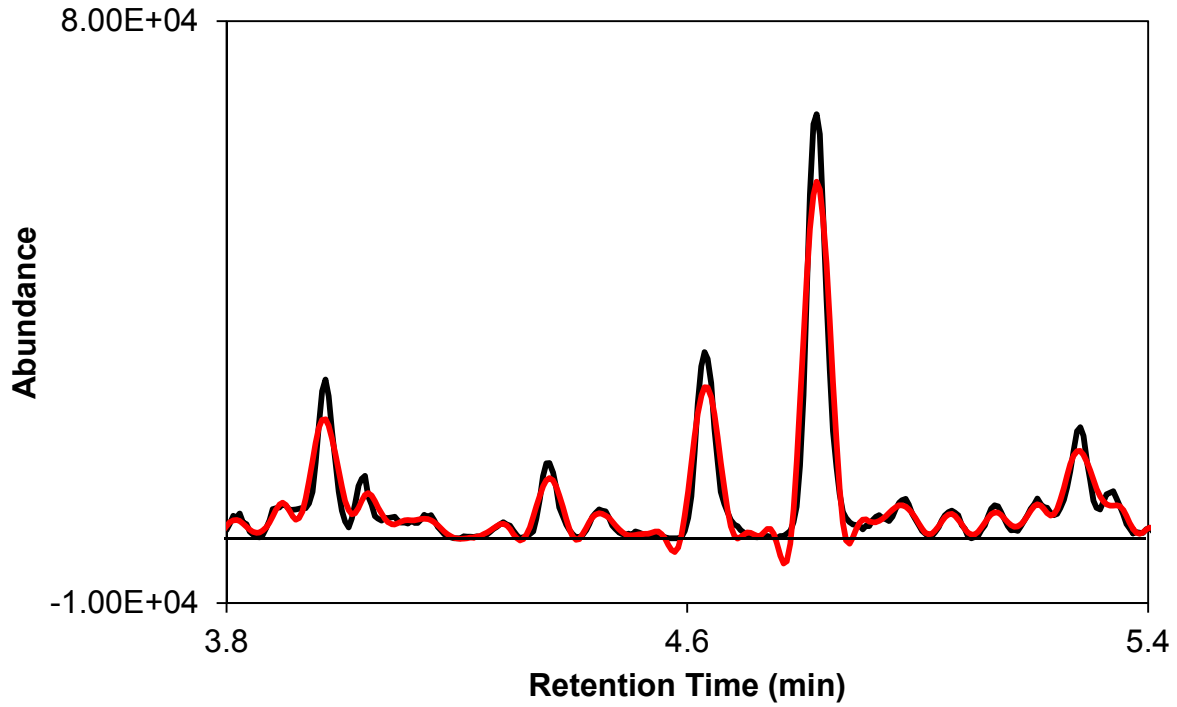


Figure 5-3. An expanded region of a diesel chromatogram without smoothing (black line) and with smoothing (red line) using a Savitzky-Golay smoothing algorithm. Part a shows a good smooth (polynomial order of 4 and 11 total points) while part b shows the broadening of peaks, decrease in peak height, and artifacts on the peak edges associated with oversmoothing (polynomial order of 6 and 31 total points).

5.3.2. Quantitative Assessment

The specific smoothing parameters that were tested and the results of the quantitative assessment are shown in Table 5-1. The percent change in the noise and signal-to-noise were used to evaluate the degree of smoothing, while the percent reduction in the peak height, peak variance, and resolution were used to indicate the extent of peak distortion. Similar percent reductions in noise were observed for different combinations of the smoothing parameters for both the fast Fourier transform and Savitzky-Golay smoothing algorithms. For example, a fast Fourier transform smooth with 1 point had a similar reduction in noise and increase in signal-to-noise to a Savitzky-Golay smooth with a 4^{th}-order polynomial with 7 points (Table 5-1). Parameters were grouped, based on the degree of smoothing. Different parameters that resulted in the same reduction in noise (degree of smoothing) were assigned to the same group (groups 1 - 5, Table 5-1). Each group has a similar reduction in noise but contains both smoothing algorithms and several different combinations of parameters for the SG smoothing algorithm. These relationships are more apparent in Figure 5-4, when the standard deviation in the noise is plotted versus the number of points in the smooth, on a log-log scale. Each group (Figure 5-4, represented by color) has approximately the same noise, but a different number of points included in the smooth. Having several combinations of parameters at each smoothing level provides the analyst with more control over the smoothing and the ability to minimize some peak distortion that may be observed.

Table 5-1. Percent change in each metric for different smoothing parameters. The parameters are grouped based on the level of smoothing.

Group	Applied Smooth	Noise[a]	Signal-to-Noise Ratio[b]	Peak Height[c]	Peak Variance[c]	Resolution[d]
1	FFT 1[e]	-20	24	-1.7	0.6	-0.2
	SG 2,5[f]	-25	34	-2.8	0.2	2.1
	SG 4,7	-21	26	-2.1	0.2	2.1
2	FFT 2	-38	58	-5.1	4.9	1.7
	SG 1,3	-35	53	-5.0	4.5	-0.3
	SG 2,7	-35	54	-4.7	1.8	2.8
	SG 4,11	-35	54	-4.6	2.7	2.6
	SG 6,15	-35	54	-5.0	2.9	2.6
3	FFT 3	-45	79	-7.0	7.8	-0.7
	SG 1,5	-45	78	-8.5	10	-1.5
	SG 2,11	-45	80	-5.5	3.1	-0.7
	SG 4,17	-45	80	-4.5	2.5	1.6
	SG 6,23	-43	75	-4.2	2.1	1.6
4	FFT 4	-50	92	-9.0	12	-3.5
	SG 1,7	-50	91	-11	18	-7.3
	SG 2,15	-50	94	-7.5	3.7	0.5
	SG 4,25	-50	98	-7.8	2.7	0.1
5	FFT 10	-58	91	-31	58	-22
	SG 1, 15	-57	88	-31	74	-25

a. Calculated as the standard deviation of the last 13 minutes of the chromatogram.
b. The maximum height of the C_{15} peak divided by the noise.
c. Using the EIC of m/z 132 for Tetralin.
d. Between EIC of Tetralin (m/z 132) and EIC of pentylbenzene (m/z 148). Baseline resolution corresponds to a value of 1.5.
e. FFT: Fast Fourier transform smoothing. The number indicates how many points were used for the smooth.
f. SG: Savitzky-Golay smoothing. The first number indicates the order of the polynomial, while the second number indicates the total number of points in the smooth.

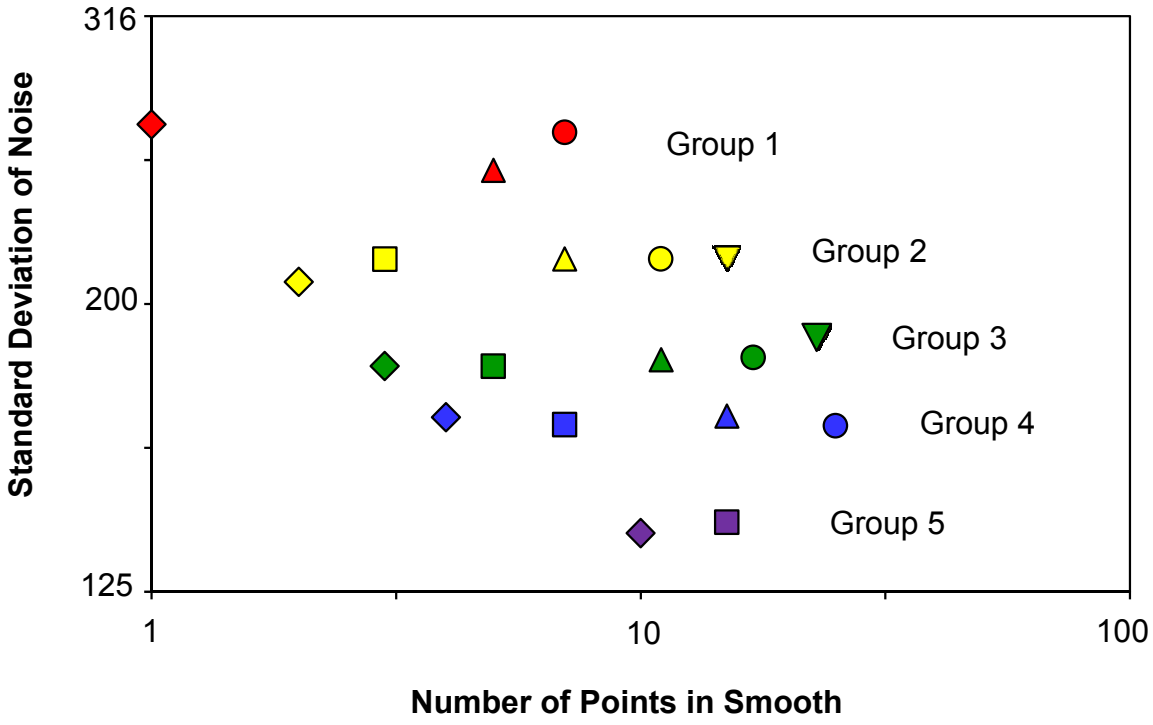


Figure 5-4. A log-log plot of the standard deviation of the noise region versus the total number of points in the smooth. Different smoothing parameters are represented by each symbol: FFT (◆), SG 1st order polynomial (■), SG 2nd order polynomial (▲), SG 4th order polynomial (●), SG 6th order polynomial (▼). Groupings were assigned based on the standard deviation in the noise region after smoothing. The color represents the groups in Table 5-1.

Over all 5 groupings, the percent reduction in noise ranged from 20 - 58%, with similar reductions in noise observed within each group. Both the FFT and SG algorithms performed similarly within each group (Table 5-1). This demonstrates the wide range of smoothing options available. Group 1 had the smallest reduction in noise (20 - 25%) while group 5 had the greatest reduction (57 - 58%). Group 1 also had the smallest improvement in the signal-to-noise ratio (24 - 34%) while groups 4 and 5 had similar improvement in the signal-to-noise ratio (88 - 98%). This shows that fewer points and a lower order polynomial result in a lower degree of smoothing and more points and a higher order polynomial result in a higher degree of smoothing.

Peak distortion was also considered when evaluating the smoothing algorithms. For all levels of smoothing, there is a reduction in the peak height, ranging from 2% to 30%. Variations in peak height of 5% were observed for replicate analyses prior to smoothing; therefore, reductions in peak heights greater than 5% were considered significant and detrimental. This only occurs in groups 3, 4, and 5. The percent change in the peak variance ranged from 0.2% to 74% for groups 1 to 5; however, significant and detrimental changes (>5%) were also only observed in groups 3, 4, and 5. In most cases, groups 1 - 4 had only small changes in resolution (generally less then ± 3%) and were likely not significant. Larger and detrimental changes in resolution were observed when the change in peak variance exceeded 18%, which is observed in groups 4 and 5. These metrics were not able to identify the artifacts that were visually observed on the edges of peaks (Figure 5-3) when the percent change in the noise exceeded approximately 40% (group 3).

In general, the FFT smoothing algorithm resulted in a monotonic decrease in noise when more points were considered in the smooth (Figure 5-4, diamonds). When a higher degree of smoothing was used (*i.e.* more points, corresponding to a lower cutoff frequency), there was also more peak distortion observed. The SG smoothing algorithm resulted in a decrease in noise when the order of the polynomial was decreased and the number of points was constant (Figure 5-4). A decrease in noise was also observed when the number of points included in the smooth was increased while the order of the polynomial was constant (Figure 5-4, squares, triangles, and circles). As with FFT smoothing, SG smoothing resulted in more peak distortion at higher degrees of smoothing. However, within a group, there was less distortion using a SG smooth with a higher order polynomial and more points. The performance of the corresponding FFT within that group usually fell in the middle of the SG parameters. The smoothing parameters in group 2 result in the largest reduction of noise (35%), while introducing only minimal (5%) peak distortion. For this work, the FFT smoothing algorithm with 2 points (FFT 2) was selected and used for all subsequent pretreatments.

Using the FFT smoothing algorithm, the ideal cutoff frequency was approximately half of the scan rate (in this work corresponding to FFT 2). For the SG smoothing algorithm, a polynomial order of 2 or 4 is sufficient for smoothing most chromatographic peaks. For a given order polynomial, an approximately 25% increase in signal-to-noise ratio is obtained by increasing the number of points in the smooth by 2n, where n is the order of the polynomial. To decrease the peak distortion and maintain the same degree of smoothing, the order of the polynomial can be increased and the number of points in

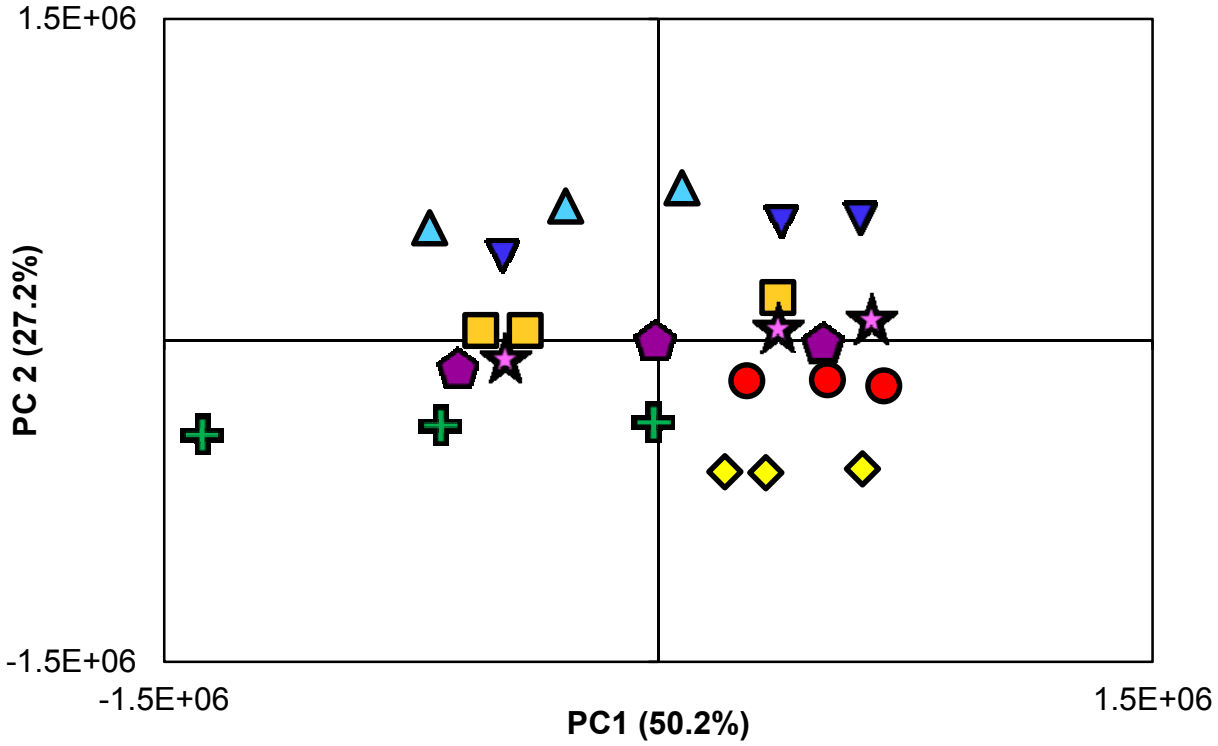
the smooth increased by 2n. After group 4, there is no increase in the signal-to-noise ratio and substantial peak distortion.

5.4. Effect of Smoothing on PCA Scores Plot

5.4.1. Visual Assessment

The PCA scores plots (Figure 5-5) after baseline correction only (a) and after baseline correction and smoothing (b) show little visual difference in the positioning of samples. This indicates that the noise is not a major source of variation in the chromatograms of these samples. After baseline correction and smoothing, PC1 and 2 account for 78.3% of the variation, only slightly increased from 77.4% for baseline correction alone.

The small change in the positioning of samples is also reflected in the loadings plots (Figure 5-6) of PC1 (a) and PC2 (b). There is little visual difference between the loadings plots after only baseline correction (Figure 4-7) and after baseline correction and smoothing. The only notable difference is that the point-to-point fluctuations observed in the noise region of the loadings plots (95 to 108 minutes) have been reduced. The small effect that smoothing has on clustering of replicates is expected, given the small contribution of this region in both the chromatogram and the loadings plots. However, if a lower signal-to-noise ratio was observed in the chromatogram, the noise would become a more significant source of variation.



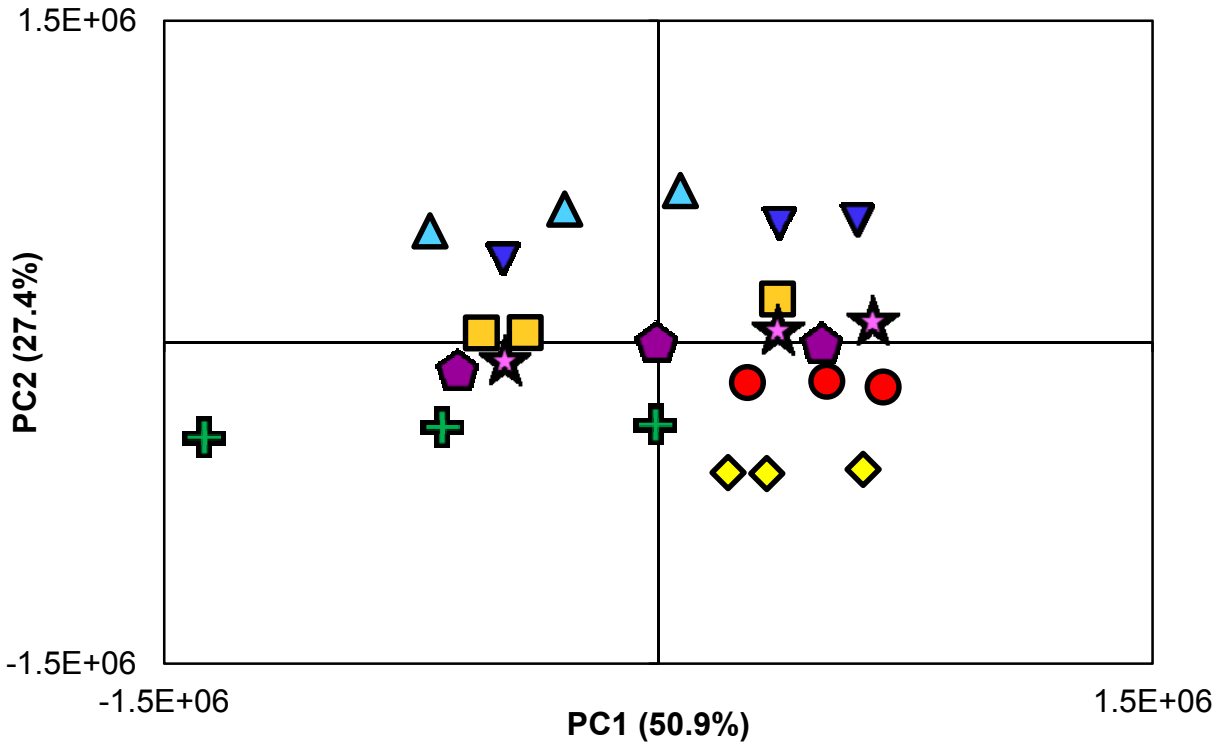


Figure 5-5. PCA scores plot of eight diesel chromatograms in triplicate after baseline correction (a) and after smoothing using FFT 2 (b).

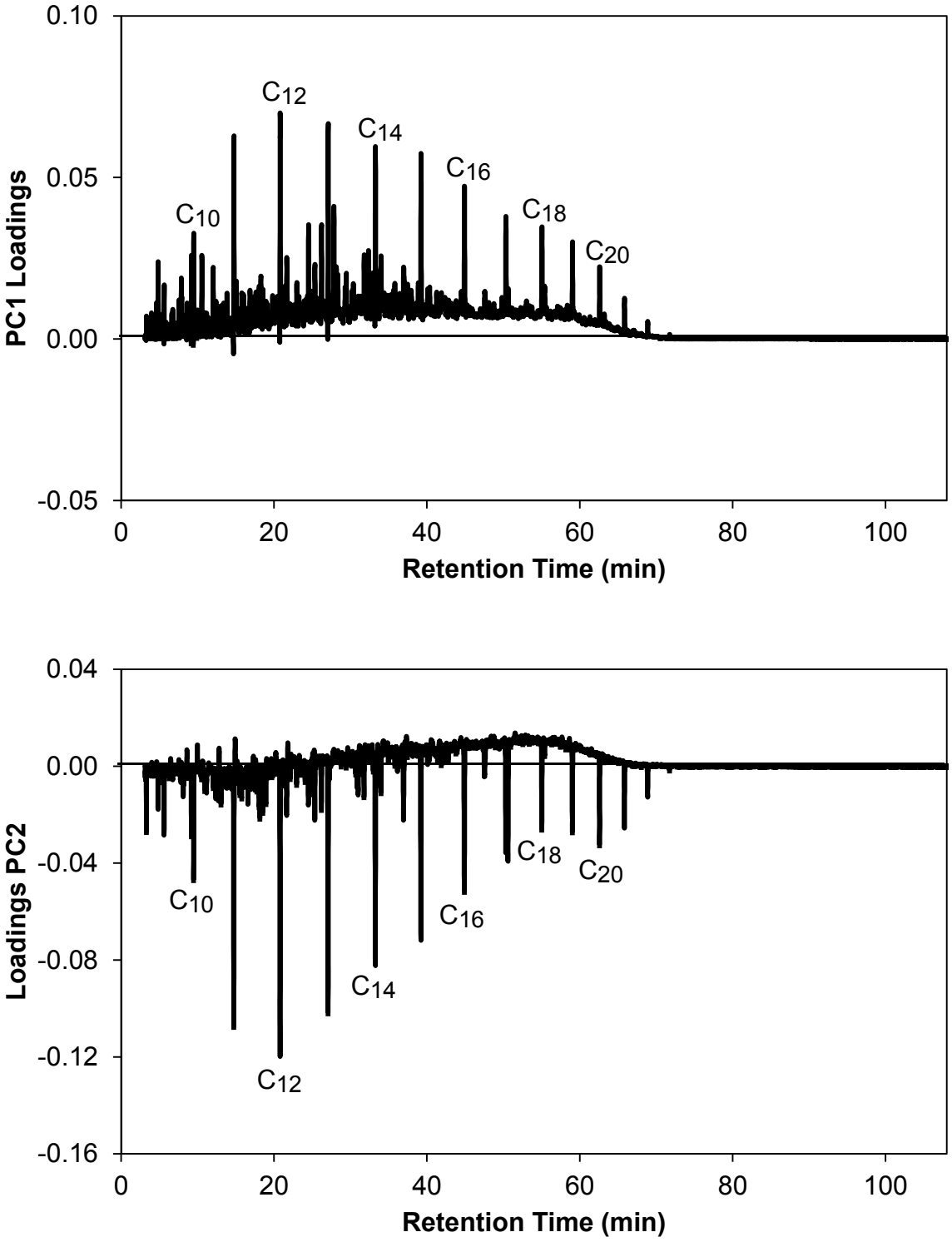


Figure 5-6. Loadings plot for PC1 (a) and PC2 (b) after PCA smoothing.

There are also only slight differences in the positioning of samples on the scores plot (Figure 5-7) after baseline correction and normalization (a) and after baseline correction, smoothing, and normalization (b). The variance accounted for by PC1 and PC2 increased from 79.3% to 81% with the inclusion of normalization as a pretreatment. Additionally, the loadings plots after baseline correction, smoothing, and normalization (Figure 5-8) for PC1 (a) and PC2 (b), appear similar to the loadings plots after only baseline correction and normalization (Figure 4-9). This again demonstrates that the noise is not a major contribution to the variance in this dataset.

5.4.2. Quantitative Assessment

To quantify the clustering of replicates on the scores plot, the percent change in the clustering of replicates (PCC) was again employed (Table 5-2). After baseline correction using the fitting of the extracted ion profiles (EIP fit) and smoothing using the FFT with 2 points, only a very small improvement in replicate clustering was observed (0.5%), further confirming that noise was not a major source of variation in this dataset. When baseline correction, smoothing, and normalization were applied, there was an increase in the PCC over normalization alone and over baseline correction and normalization (Table 5-2). The improvements that are observed are greater than the simple sum of the smoothing affect alone and are likely due to the decreased variation in the noise at the end of the chromatogram. While these improvements were small, this highlights the power of combining data pretreatment procedures to enhance discrimination of highly similar samples.

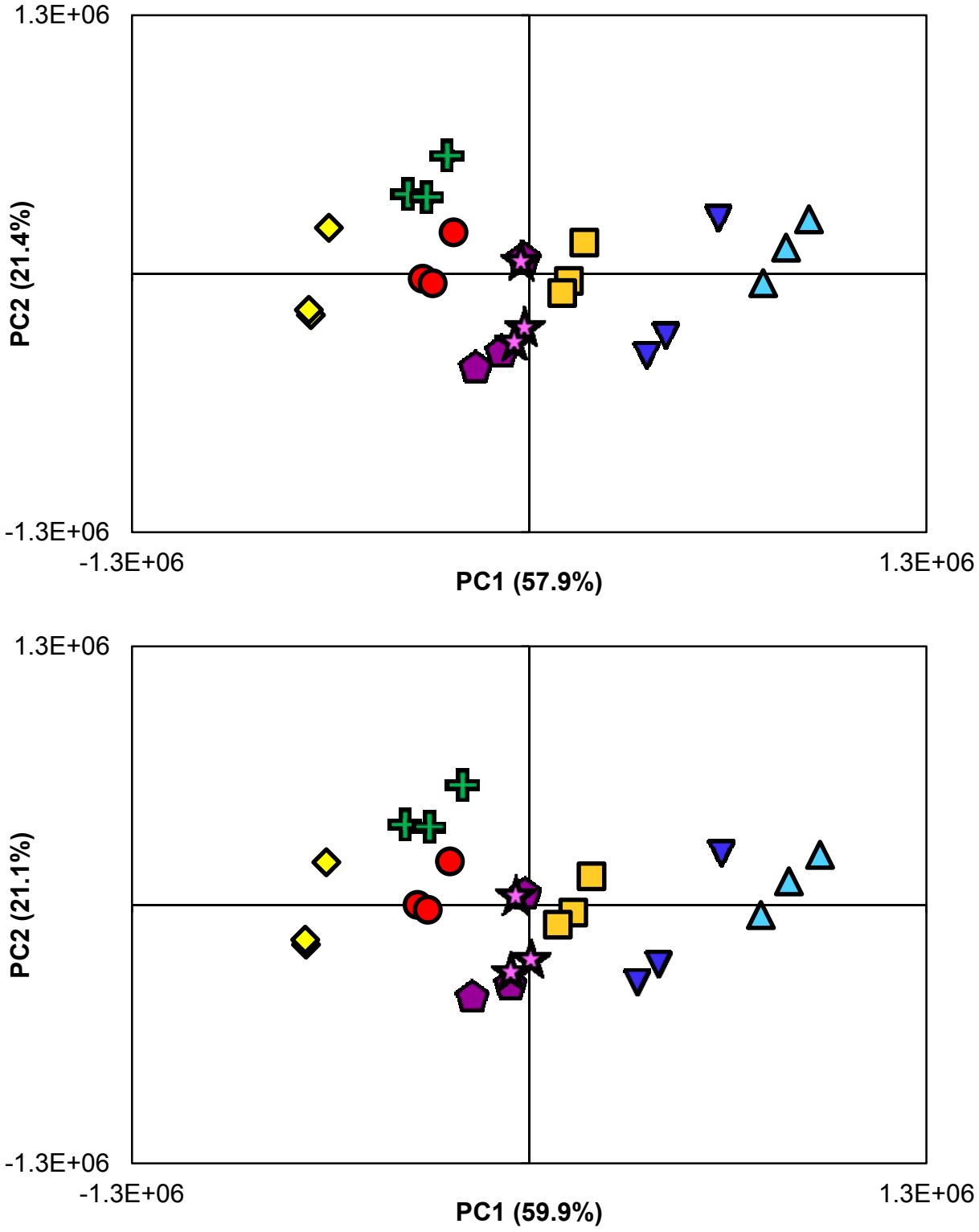
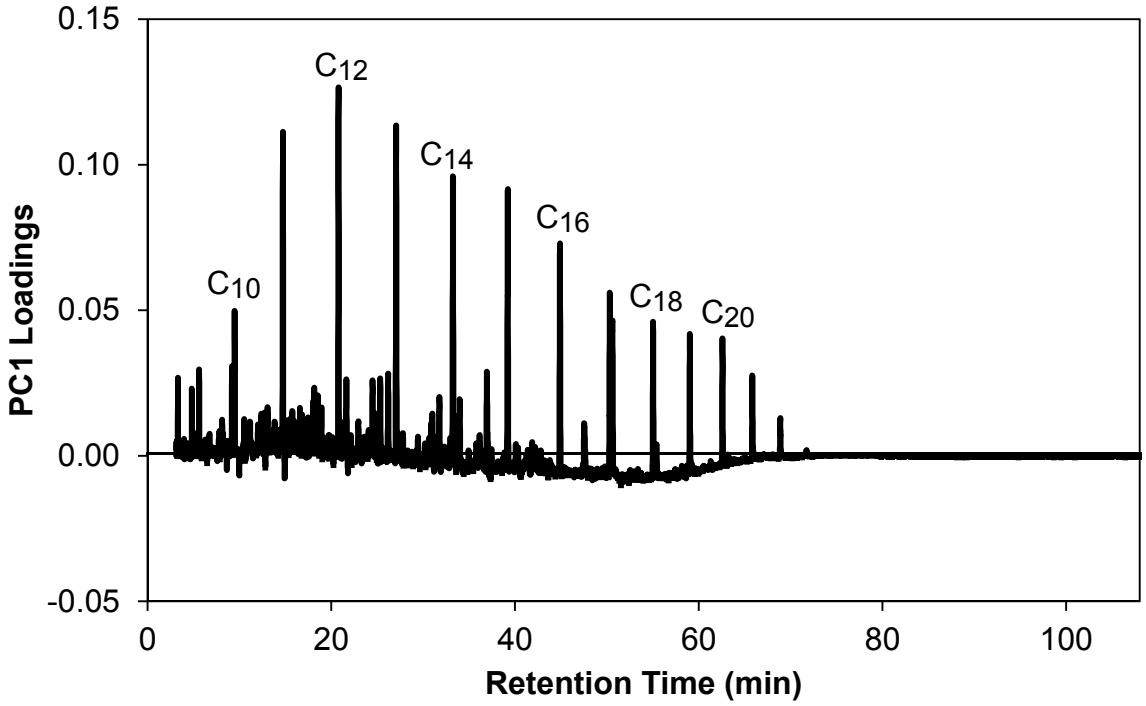


Figure 5-7. PCA scores plot of eight diesel chromatograms in triplicate after baseline correction and normalization (a) and after baseline correction, smoothing, and normalization (b).



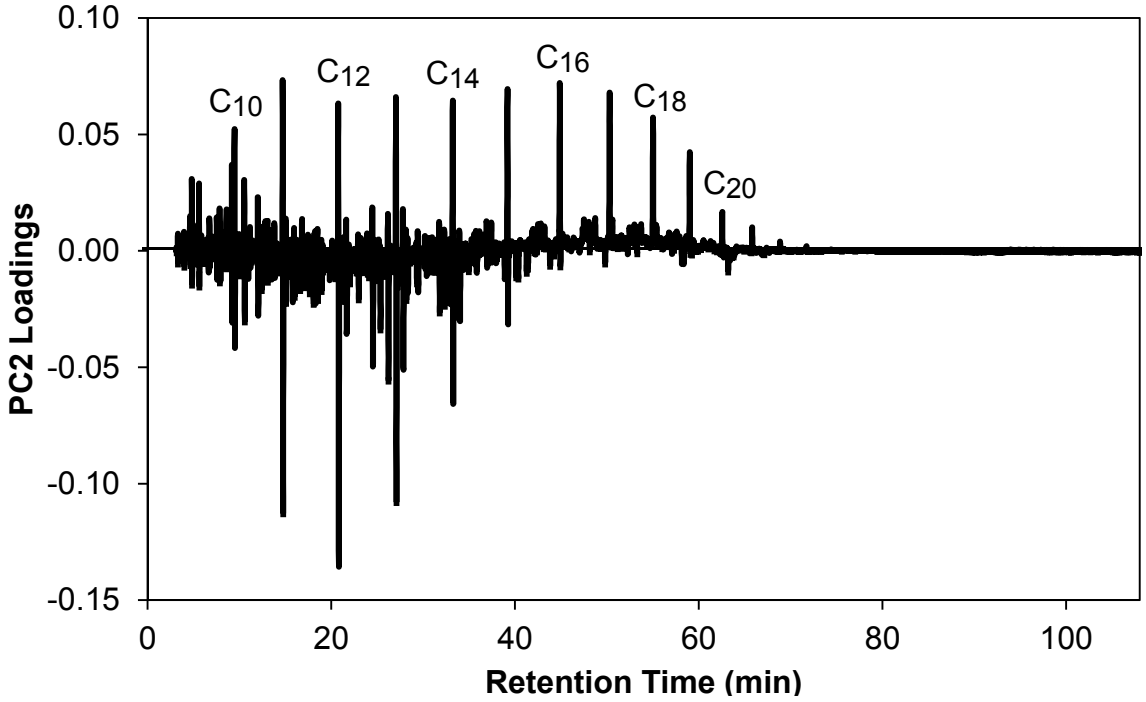


Figure 5-8. Loadings plot for PC1 (a) and PC2 (b) after PCA smoothing and normalization.

Table 5-2. The average percent change in the clustering (PCC) of replicates after the listed pretreatment procedures including baseline correction using the extracted ion profiles (EIP fit), smoothing using fast Fourier transform smooth with 2 points (FFT2) and normalization using total area (Area) and single peak (Peak) normalization methods.

Baseline Correction	Smoothing	Normalization	PCC
-	-	Area	45.1
-	-	Peak	59.0
EIP fit	-	-	0.0
EIP fit	-	Area	44.6
EIP fit	-	Peak	58.7
EIP fit	***FFT 2***	***-***	***0.5***
EIP fit	***FFT 2***	***Area***	***46.5***
EIP fit	***FFT 2***	***Peak***	***62.1***

5.5. Summary

There are a wide array of smoothing methods, each with many parameters that can be applied, depending on the enhancement that is required. However, care must be taken not to introduce peak distortion, particularly the negative-going peaks on the edges of peaks, which, when present, can be identified as a major source of variation between samples. All smoothing parameter groupings led to a reduction in noise; however, different smoothing parameters resulted in different degrees of peak distortion. Peak distortion was first identified when there was a greater than 35% reduction in the noise.

In this work, there was little difference in the chromatogram or scores plot after applying a moderate degree of smoothing. For this dataset, the contribution of the noise was small compared to the signal. Therefore, noise generally had a minimal effect on statistical comparisons. For datasets where signal-to-noise ratio is lower, application of smoothing would be more critical. The enhanced discrimination after applying baseline correction, smoothing, and normalization demonstrate the improved enhancement that can be achieved by applying several pretreatment procedures.

REFERENCES

REFERENCES

[1] K.M. Pierce, J.S. Nadeau, R.E. Synovec, in: C.F. Poole (Ed.), Gas Chromatography, Elsevier, Waltham, MA, 2012.

[2] Y.-z.L. Foo-tim Chau, Junbin Gao, Xue-guang Shao (Ed.), Chemometrics From Basics to Wavelet Transform, John Wiley & Sons, Inc. , Hoboken, NJ, 2004.

[3] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, Inc., New York, NY, 1998.

[4] S.L. Morgan, E.G. Bartick, in: R.D. Blackledge (Ed.), Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis, John Wiley & Sons, Inc., Hoboken, NJ, 2007.

[5] A. Savitzky, M.J.E. Golay, Anal. Chem., 36 (1964) 1627.

[6] P. Barak, Anal. Chem., 67 (1995) 2758.

[7] M.J. Adams, Chemometrics in Analytical Spectroscopy, Royal Society of Chemistry, Victoria, Australia, 1995.

[8] J.P. Foley, J.G. Dorsey, Anal. Chem., 55 (1983) 730.

[9] D.W. Morton, C.L. Young, J. Chromatogr. Sci., 33 (1995) 514.

[10] J.C. Giddings, Unified Separation Science, John Wiley & Sons, Inc., New York, 1991.

[11] V.L. McGuffin, in: E. Heftmann (Ed.), Chromatography Elsevier, New York, NY, 2004.

CHAPTER 6: ALIGNMENT

6.1. Introduction

After appropriate minimization of the baseline and the noise in a chromatogram, ideally, only the analytical signal remains. However, drift in the retention time of peaks in the chromatogram between analyses can remain, particularly in datasets that were collected over a relatively long time period (usually months or longer). This drift can arise from variation in injection mode, fluctuation in mobile phase pressure and flow rates, degradation of the stationary phase, variation in the oven of the gas chromatograph, among other sources. All of these sources of drift effect how the analytes move through the column. When performing principal component analysis (PCA) on these data, each retention time serves as a variable. Therefore, peaks from the same compounds in different chromatograms must be well aligned so that the variables rise and maximize at the same retention time. Any retention time misalignments will be identified as sources of variation between samples, which will be highlighted in the statistical analysis [1-4]. Many factors can affect the severity of the misalignments. Chromatograms analyzed on the same instrument immediately after one another will have smaller misalignments than chromatograms analyzed months apart or from different instruments [4]. Also, variation in alignment can be reduced by optimizing the injection parameters and by utilizing an auto-sampler.

In order to correct misalignments, retention time alignment algorithms are employed. All alignment algorithms utilize interpolation or extrapolation of points in the chromatogram to shift the peak in a sample chromatogram to the corresponding peak in

a target chromatogram. The target chromatogram is considered to have the true retention times, and all sample chromatograms are then aligned so that the peaks from the same compounds in different chromatograms maximize at the same retention time. Alignment algorithms can be generally classified into four types according to their mode of operation: scalar shifts, selected peak alignment, local alignment, and global optimized alignment algorithms [1, 5]. Considerations for choosing a target for alignment will be discussed later in this chapter.

Alignment algorithms based on scalar shift apply a shift to the entire sample chromatogram, to maximize the similarity between the sample and target chromatogram. This simplistic alignment allows for a fast, but crude, alignment [1] and shift all of the peaks in the sample chromatogram in one direction and by the same number of points. This type of coarse alignment is sometimes performed to correct for large shifts in retention times prior to more robust alignment methods [6, 7].

Selected peak alignment is performed by assigning a specific value for the retention time of known peaks in the chromatogram. The retention times of the other peaks are then scaled between the known peaks. This method is similar to scaling using the Kovat's retention index [8]. However, in complex samples it is difficult for algorithms to identify known peaks and manual intervention is often necessary [1]. A target chromatogram is not needed for this type of alignment and specific peaks that are present in all samples serve as the bases for alignment.

Local alignment algorithms are applied iteratively to regions of the sample chromatogram to maximize similarity to the target within each region [1]. Generally, this

method requires peak detection or other method of defining the local regions of interest in both the target and sample chromatograms [3, 5, 9]. This method requires no prior knowledge about the sample; however, it only aligns small regions of the chromatogram, such as selected peaks, rather than maximizing similarity between all peaks in the chromatogram [1, 10].

The global optimized alignment algorithms are the most dynamic and robust as they maximize a local as well as a global measure of similarity [1-3, 7, 11-14]. This method allows for alignment of chromatograms with different numbers of data points and with severe shifts in retention time. These methods are often computationally intensive and require optimization of several parameters [15].

6.2. Methods Tested and Evaluation Metrics

The performance of two common retention time alignment algorithms was compared: a local alignment algorithm or peak-matching (PM) algorithm [10] and a global alignment algorithm, or correlation-optimized warping (COW) algorithm [11]. These two algorithms were selected as they are more robust and require less manual intervention than the other types of algorithms. Many commercially available chemometric software packages include a COW alignment algorithm. The PM algorithm was applied in Matlab (version 7.12 R2011a, MathWorks, Natick, MA) and the COW alignment was applied in LineUp (version 3.5, Infometrix, Inc., Bothwell, WA). The performance of the alignment algorithms was evaluated using Diesels 1 - 3, which were analyzed in triplicate, after baseline correction and smoothing.

6.2.1. Peak-Matching Alignment Algorithm

The PM algorithm identifies and matches individual peaks in a sample chromatogram to peaks found in a target chromatogram [10]. Peaks are detected in each chromatogram by identifying zero-crossings after an estimation of the first derivative of each chromatogram. The algorithm considers points starting at the beginning of the chromatogram and moves to the end. The leading edge of a peak is identified when the point-to-point difference exceeds five times the standard deviation of the baseline. When this threshold is met, the first zero-crossing that is encountered is considered the peak maximum. The time point closest to the zero-crossing is then added to a list for each chromatogram (the target and all sample chromatograms). The algorithm continues to locate peaks until the end of the chromatogram is reached, creating a list of time points closest to each zero-crossing identified. The peaks found in each sample chromatogram are then compared to those found in the target chromatogram. If a peak is present in both the target and sample chromatograms, within a user-defined window, then the peaks are considered a match. The retention time axis is interpolated, so that the point closest to the zero-crossing in the sample chromatogram occurs at the same retention time as the point closest to the zero-crossing in the target chromatogram [10].

In this work, the algorithm was used as described by Johnson *et al*. [10], except that the baseline subtraction step was omitted as baseline correction was performed as a separate pretreatment method, prior to alignment. The threshold was calculated as five times the standard deviation of the noise, which was defined as the region in the chromatograms between 79.5 and 80.5 minutes. This particular region was selected as there were no peaks present and the region was only minimally affected by baseline

correction. The window size is the only user-defined parameter in the algorithm and window sizes ranging from 2 to 20 data points were evaluated.

6.2.2. Correlation Optimized Warping Algorithm

The COW algorithm optimizes the correlation coefficient (Equation 1-1) between a sample and target chromatogram [12]. As with the PM algorithm, each sample chromatogram is compared to a target chromatogram. In order to align the chromatograms, both the target and sample chromatograms are divided into segments, based on a user-defined parameter of segment size. The segment size is the number of data points in each segment. Beginning at the end of the chromatogram and moving towards the beginning, each segment of the sample chromatogram is stretched or compressed by adding or removing points, using interpolation, in order to better align the peaks in the sample chromatogram to those in the target. The maximum number of points added or removed is determined from the warp, which is also a user-defined parameter. The Pearson product-moment correlation (PPMC) coefficient is used to assess the similarity between data points in the segment of the sample chromatogram and the corresponding points in the segment of the target chromatogram. The PPMC is calculated for each permutation of adding or removing up to the number of data points specified by the warp. This process is repeated for each segment. The alignment is based on the highest global correlation coefficient for all segments [11].

In this work, the COW algorithm was tested using varying warps (1 - 4 data points) and segment sizes (25 - 120 data points). For the COW alignment algorithm, the initial starting point recommended for the segment size is the approximate number of points

across a peak, with the warp typically being just a few points. Peaks in this dataset were approximately 45 points across and generally peaks were shifted less than two points from the target chromatogram. Hence, the starting point for the alignment was chosen to be a segment size of 45 and a warp of 2. During investigation of warp and segment size, one parameter at a time was varied while the other was held constant.

6.2.3. Target Selection

As discussed above, each alignment algorithm compares sample chromatograms to a target chromatogram. Therefore, selection of the target is a critical and often a challenging aspect of alignment. The ideal target chromatogram has well resolved peaks and is representative of the sample chromatograms [6, 10]. There are generally three targets that can be selected: one of the sample chromatograms, an average target, and a consensus target [16-20]. Generally, if a sample chromatogram is used as the target, the chromatogram is chosen at random from the dataset. Using a sample chromatogram can be problematic if all of the compounds in the dataset are not present in the selected sample chromatogram. An average target is generated mathematically from all chromatograms in the dataset. To do this, the abundance at each retention time is added from each of the sample chromatograms, then divided by the total number of chromatograms to yield the average. The average target is advantageous because it includes peaks that may not be present in every sample. However, averaging leads to peak broadening and a reduction in signal, which make alignment more challenging. A consensus target is a separate sample that contains a mixture of all compounds of interest that are in the sample chromatograms. This type of target is challenging to create

because trial and error is often required to create a mixture with all of the compounds at the correct abundance.

In this work, a random target was used to investigate the alignment parameters for the COW and PM alignment algorithms using replicates (n=3) of Diesels 1 - 3. The second replicate from Diesel 2 was randomly selected (using a random number generator in Excel) to serve as the target chromatogram. The optimized parameters from this preliminary study were then used to align the dataset to evaluate different target chromatograms. Each diesel chromatogram (from samples 1 - 10, analyzed in triplicate) and the average chromatogram were used to determine the most appropriate type of target chromatogram for alignment of these data. The success of alignment was evaluated for each target.

6.2.4. Evaluation Metrics

Two metrics were used to quantitatively evaluate retention time alignment. These metrics were the percent change in the average of the standard deviation of the retention time of selected peak maxima (PC-SDRT) and the sum of the percent change in the PPMC coefficient for each chromatogram before and after alignment (PC-PPMC). Calculation of the PC-SDRT is performed using peak maxima selected using a peak-finding algorithm, based on the peak-matching algorithm described previously [10]. In general, there were 100 - 200 peaks identified per chromatogram using this algorithm. The standard deviation in the retention time was calculated for each selected peak across all chromatograms in the dataset, and then averaged across all selected peaks, both before ($\overline{s_U}$) and after ($\overline{s_A}$) alignment.

$$PC - SDRT = \frac{\left(\overline{s_A} - \overline{s_U}\right)}{\overline{s_U}} * 100$$ Equation 6-1

To determine the PC-PPMC, PPMC coefficients were first calculated in Excel, in a pair-wise fashion, between all chromatograms, both before and after alignment. The percent change in the PPMC coefficient before ($PPMC_U$) and after ($PPMC_A$) alignment was calculated and summed.

$$PC - PPMC = \sum \left[\frac{\left(PPMC_A - PPMC_U\right)}{PPMC_U} * 100 \right]$$ Equation 6-2

After proper retention time alignment, chromatograms become more similar, which results in a lower standard deviation of peak maxima and higher PPMC coefficient. Each metric is similar to one of the methods used for alignment: the PC-SDRT is based on the peak finding algorithm while the PC-PPMC utilizes PPMC coefficients, similar to the COW algorithm.

6.3. Effect of Retention Time Alignment on Chromatographic Data

6.3.1. Visual Assessment

Visual assessment of alignment is challenging because multiple chromatograms must be overlaid and compared. In this work, misalignments were considered small, often within 5 points (generally ± 0.02 min), and were difficult to visualize. Figure 6-1 shows the chromatograms of three diesel samples, each analyzed in triplicate, overlaid. In this example, the peak maxima, as well as the leading and tailing edges of the peak, can vary

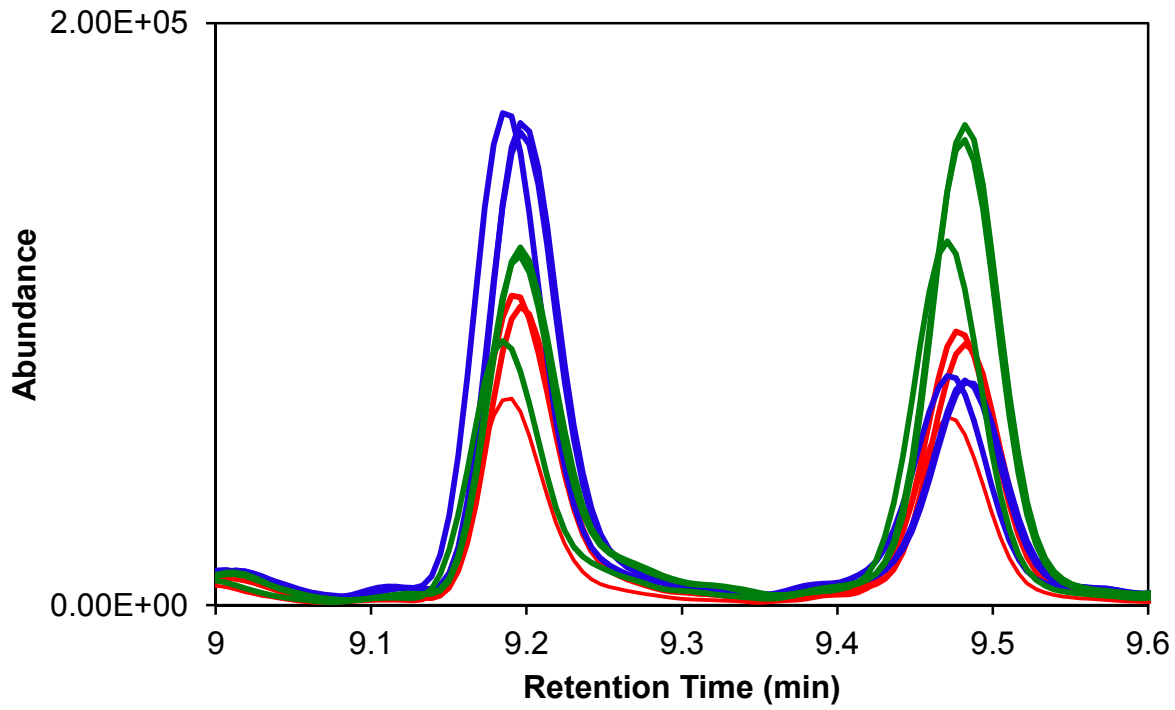


Figure 6-1. An expanded region of chromatograms of three diesel samples analyzed in triplicate, each represented by a different color, before alignment. The peaks correspond to 1, 3, 5-trimethyl-benzene (9.20 min) and decane (9.48 min).

slightly, even among replicates. When multivariate statistical procedures are applied, these differences can be identified as sources of variation and hence, it is important to minimize or eliminate such differences.

Selection of an appropriate window size (for the PM algorithm) or warp (for the COW algorithm) is facilitated by careful inspection of the chromatograms, as shown in Figure 6-2. When the chromatograms are overlaid, the peaks resulting from the same compound in all chromatograms should be inspected. In Figure 6-2, the peak shown corresponds to 1, 3, 5-trimethyl-benzene from Diesels 1 - 3, each analyzed in triplicate. The minimum window size or warp is the number of points that a peak would need to be shifted to align with peaks from the same compound in the other chromatograms. In this example, the window size or warp would need to be at least 2. The maximum window size or warp is the number of points a peak could be shifted before being aligned to the peak from another compound. If the window size or warp is too small, the peaks cannot be aligned; if it is too large, then peaks from different compounds could be aligned.

Figure 6-3 shows the same region as Figure 6-1, after alignment using the PM (a) and COW (b) alignment algorithms. In order to compare the alignment algorithms, all diesel samples were aligned. The PM alignment was performed using a window size of 5 and the COW alignment was performed using a warp of 2 and a segment size of 75. In both cases, an average chromatogram was utilized as the target. After alignment, peak maxima and edges are generally more similar across all chromatograms, generally only varying by 1 or 2 points (approximately ± 0.005 min). However, there are anomalies that

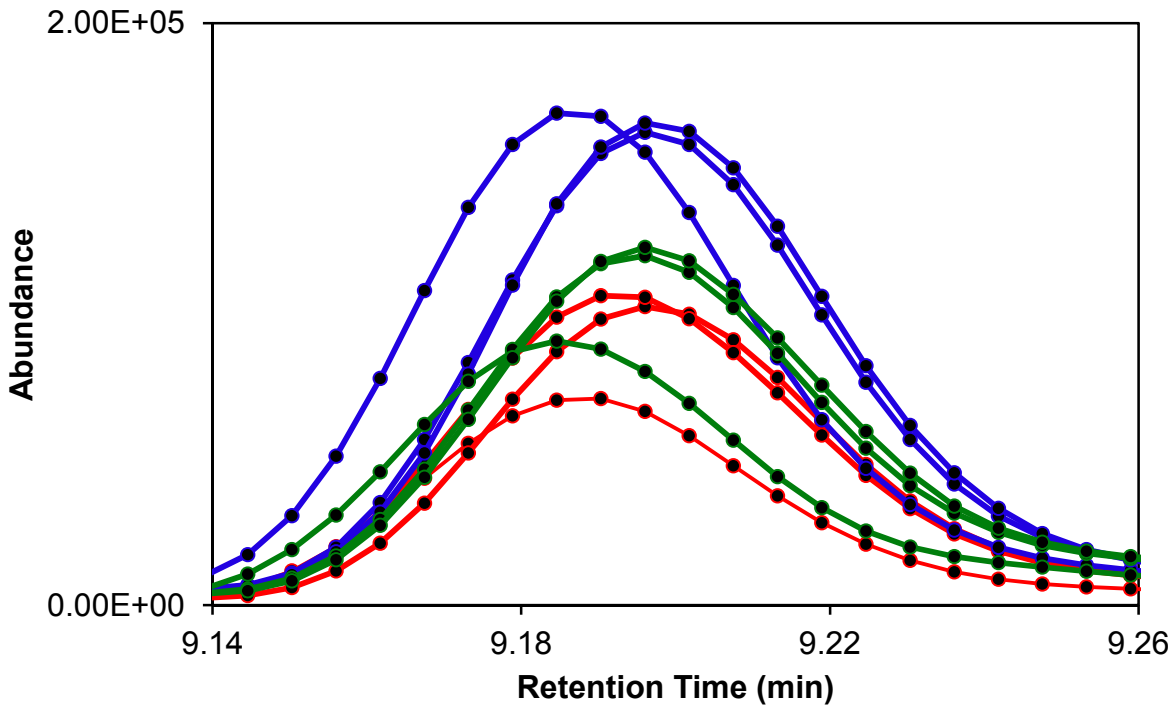


Figure 6-2. An expanded region of the 1, 3, 5-trimethyl-benzene peak in chromatograms of three diesel samples analyzed in triplicate, each represented by a different color, before alignment. The individual data points are shown as black circles. In this example, peak maxima are shifted by approximately three data points.

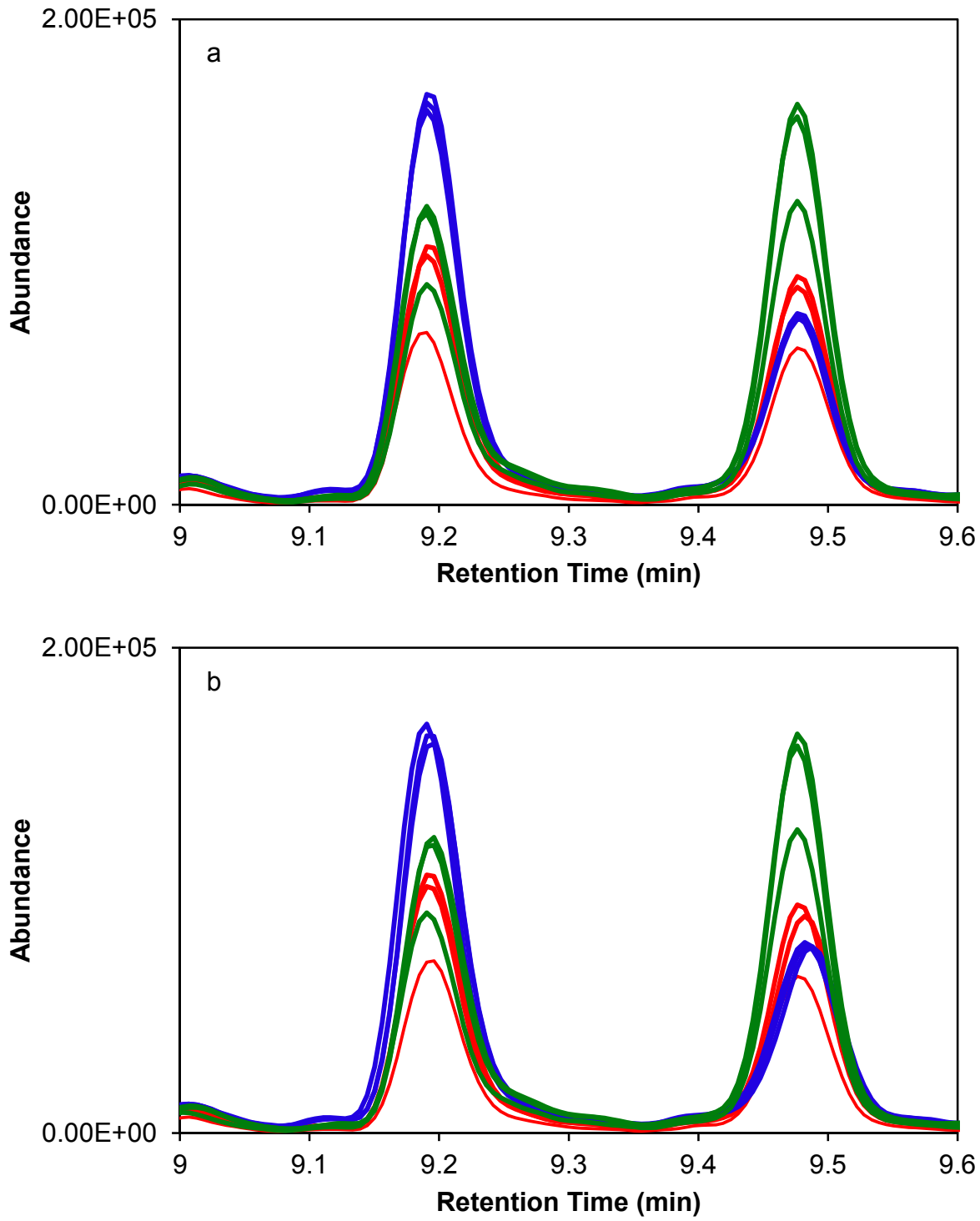


Figure 6-3. The same expanded region of chromatograms of three diesel samples from Figure 6-1, each represented by a different color, after alignment using the peak-matching algorithm (a) and the correlation-optimized warping algorithm (b).

are observed after some alignments. Using the COW algorithm, the most commonly encountered anomaly is that peaks are often aligned to one of the edges of the peak, rather than to the peak maxima. This can be seen in both peaks in Figure 6-3b. Peaks that are approximately the same height and width are well aligned. However, when comparing peaks across several chromatograms, most peaks are aligned to either the leading or tailing edge of the peak. The first peak in all chromatograms (1, 3, 5-trimethyl-benzene) is aligned to the tailing edge of the peak. The leading edge of the peak and the peak maxima are not well aligned. In the second peak (decane), most of the replicates are also aligned to the tailing edge of the peak. However, one of the replicates shown in red is aligned to the leading edge of a peak in another diesel sample. Even though these variations are often only 1 or 2 points, they can still be identified as a non-chemical source of variation in PCA.

The resulting alignment of the peak edges using the COW alignment algorithm is not surprising because optimization is based on maximizing correlation coefficients between the sample and target chromatograms. The correlation is higher when the abundance from point to point increases and decreases at the same retention times across all chromatograms. When the peaks being aligned are different width or height, alignment to the front or tail of the peak is common, so that correlation is optimized.

In the PM algorithm, peak maxima are identified and aligned (Figure 6-3a). Therefore, differences in peak size do not affect the alignment. However, only peaks that have been identified by the algorithm in both the sample and target chromatogram are aligned. Therefore, some low abundance or co-eluting peaks are often not aligned. Additionally, this can result in alignment of peaks that do not correspond to the same

compound, often making alignment worse. The sensitivity of the algorithm to identify peaks is the major drawback of this method. Figure 6-4 shows the phytane peak and another co-eluting peak in three diesel chromatograms, each analyzed in triplicate, before alignment (a) and after alignment using the PM algorithm with a window size of 10 (b). Using a large window size results in peak maxima of one replicate from each sample to shift, creating retention time misalignments. This problem can be minimized by selecting an appropriate window size, which requires manual optimization.

6.3.1. Quantitative Assessment

The PC-SDRT and the PC-PPMC were utilized as metrics to evaluate the alignment. A decrease in the PC-SDRT or an increase in the PC-PPMC indicates an improvement in alignment. The PM algorithm resulted in improved alignment for most of the window sizes that were investigated (Table 6-1). The PC-SDRT ranged from -62 to 49% and the PC-PPMC ranged from -2.8% to 9.3%. Similar improvement in the quality of alignment was observed for window sizes 3 - 7 using both the PC-SDRT (-61% to -62%) and the PC-PPMC (9.2% to 9.3%). This indicates that windows of 3 - 7 resulted in similar alignment. Window sizes of less than 3 were not able to shift peaks far enough to align them. When the window size was greater than 7, peaks were shifted too far, resulting in improper alignment.

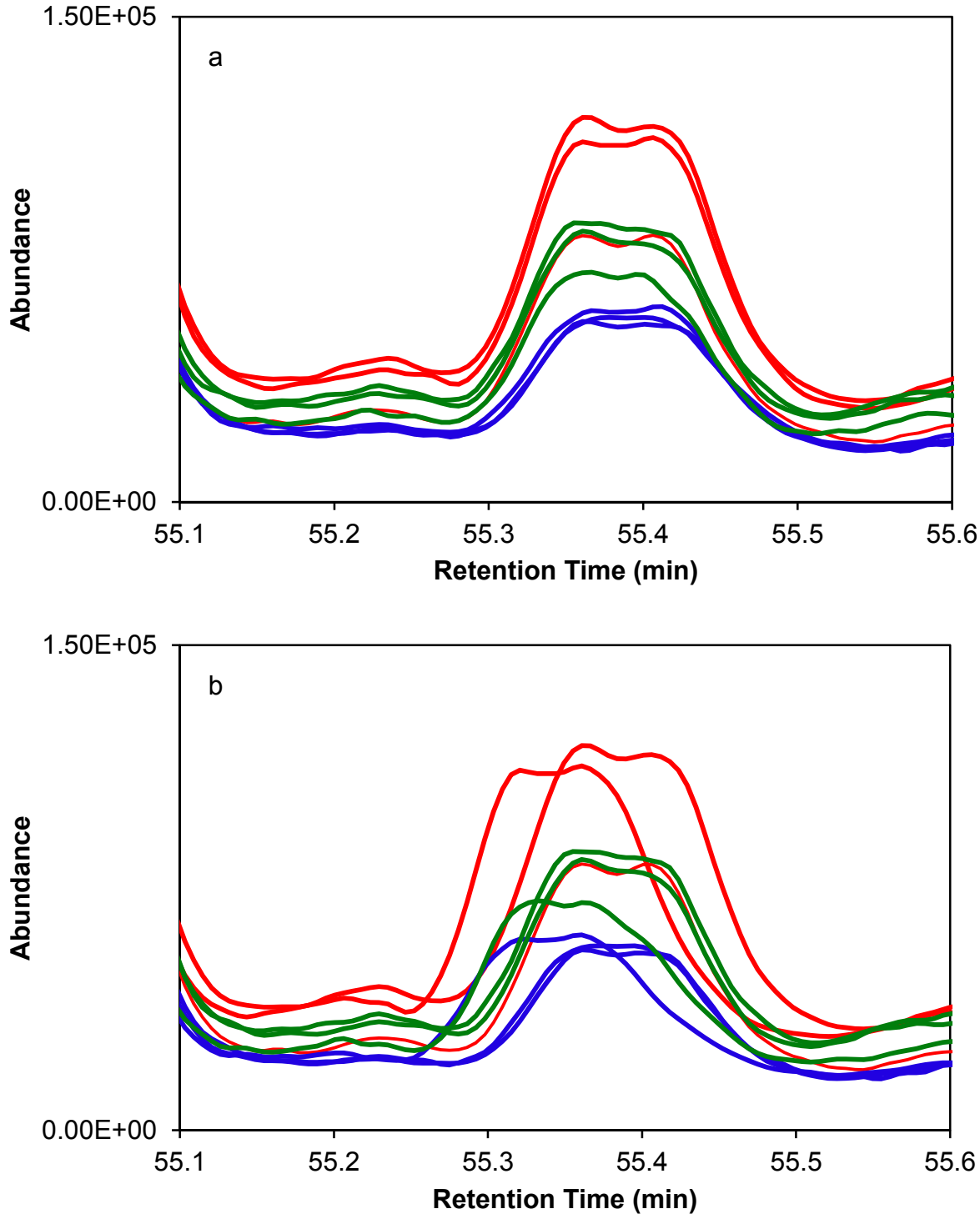


Figure 6-4. An expanded region of the phytane peak in chromatograms of three diesel samples analyzed in triplicate, each represented by a different color, before alignment (a) and after alignment using the peak-matching algorithm with a window size of 10 (b).

Table 6-1. Percent change in the standard deviation of the peak maxima of selected peaks (PC-SDRT) and the sum of the percent change in the PPMC coefficients (PC-PPMC) for different window sizes using the peak-matching alignment algorithm. A decrease in the PC-SDRT of the retention time or an increase in the sum of the PC-PPMC indicates an improvement in alignment.

Window Size	PC-SDRT	PC-PPMC
2	-55	9.2
3	-61	9.2
4	-62	9.2
5	-62	9.2
6	-62	9.3
7	-61	9.2
8	-61	8.4
10	-46	8.5
15	-19	5.8
20	49	-2.8

There was a decrease in the quality of the alignment based on the PC-SDRT metric for window sizes below 3 and above 8 and a decrease based on the PC-PPMC using a window size greater than 7. This reduction in quality of alignment is likely a result of a mixture of improved alignment for some peaks and a worsening in alignment for other peaks in the same chromatogram. At a window size of 20, there is a positive PC-SDRT and a negative PC-PPMC, indicating that this window size actually resulted in worse alignment than prior to application of the alignment algorithms. This is likely a result of additional misalignments caused by aligning peaks in the sample chromatogram to the improper corresponding peak in the target chromatogram.

To compare the COW alignment algorithm, a range of segment sizes (20 – 120) and warps (1 – 4) were investigated. The recommended segment size for this algorithm is the average number of points across a peak [6]. In this work, the number of points across a peak ranged from approximately 25 - 45 points. The largest peaks were expected to be the most problematic, so an initial segment size of 45 data points was selected. A warp of 2 was used to compare the various segment sizes. This warp was chosen based on visual assessment of the unaligned chromatograms, which showed that most peaks were only misaligned by 1 or 2 data points.

The COW alignment algorithm resulted in improved alignment over the range of segment sizes (20 – 120) that were investigated (Table 6-2). The PC-SDRT ranged from -21 to -30% and the PC-PPMC ranged from 13 to 21%. Both the decrease in the standard deviation in retention time and the increase in PPMC coefficient indicates an increase in

Table 6-2. Percent change in the standard deviation of the peak maxima of selected peaks (PC-SDRT) and the sum of the percent change in the PPMC coefficients (PC-PPMC) for varying warp and segment sizes using the COW alignment algorithm. A decrease in the PC-SDRT of the retention time or an increase in the sum of the PC-PPMC indicates an improvement in alignment.

Segment	Warp	PC-SDRT	PC-PPMC
20	2	-26	14
25	2	-21	13
37	2	-27	19
45	1	-29	21
45	2	-28	21
45	3	-24	20
45	4	-23	20
50	2	-23	20
60	2	-25	20
75	2	-28	21
90	2	-30	21
100	2	-26	21
120	2	-29	21

the quality of the alignment for all combinations. All segment sizes of 45 and larger resulted in a 20% to 21% increase in the PC-PPMC while segment sizes 45, 75, 90 and 120 showed the highest decrease in the PC-SDRT (28% to 30%).

Using the segment size of 45, the warp was varied between 1 and 4 data points. The greatest improvements in alignment were observed using a warp of 1 or 2 points. There was a 29% and 28% reduction in the PC-SDRT for warp sizes of 1 and 2, respectively. There was also 21% increase in the PC-PPMC for both a warp 1 and 2 points. Using a warp of 3 and 4 points, there was a 24% and 23% reduction in the PC-SDRT and a 20% increase in the PC-PPMC. The smaller warps resulted in better alignment, as the chromatograms were collected over a short period of time, and only small differences in alignment were observed. When larger warps were applied, chromatograms were shifted more, resulting in poorer alignment, due to more peaks aligning to the leading or tailing edge of the peak.

Rather than selecting the ideal alignment parameters (which can only be obtained through optimization), adequate alignment parameters were selected for further analysis. For COW, a warp of 2 data points and a segment size of 75 data points were selected. Most peaks were misaligned by 1 or 2 points, making 2 a reasonable choice for warp. A segment size of 75 corresponds to 1.5 to 2 times the number of points across a peak. For the PM algorithm, a window size of 5 was selected, to allow for slightly larger shifts that might be present when applying the pretreatment to the larger dataset.

Each metric is based on the method used to align the chromatograms resulting in a potential bias when trying to comparison the alignment algorithms. When comparing

the PM parameters (window size 5) and COW parameters (warp of 2, segment size 75), the PC-SDRT indicates greater improvement for the PM algorithm, while the PC-PPMC indicates greater improvement for the COW alignment algorithm. Because each evaluation method is similar to the alignment algorithms, the evaluation favors the alignment algorithm from which the metric is derived. In PCA, variation between samples must be minimized. Therefore, the change in PPMC coefficients would be a more reliable indicator as coefficients account for more of a global change, rather than just selected peaks. Additionally, when a peak is not correctly aligned using the COW algorithm, the misalignment is generally less severe than with the PM algorithm. Lastly, the COW algorithm is widely available in a number of commercial software packages. Therefore, the COW alignment algorithm with a warp of 2 data points and a segment size of 75 data points was utilized for the rest of this work.

6.3.2. Target Selection

After choosing the alignment algorithm and parameters, a method for choosing a target chromatogram was investigated. Each chromatogram contained all of the compounds, making each chromatogram a suitable target for consideration. In addition, the average chromatogram was also utilized as the target. Both metrics were again applied to evaluate the selection of a target.

The PC-SDRT using each of the selected targets ranged from 12.1 to -14.2% (Table 6-3), where negative values indicate an improvement in the alignment and positive values indicate a worsening in the alignment. The greatest improvement in alignment was observed when the average chromatogram was used as the target. Additionally,

most of the possible target chromatograms resulted in an improvement in the quality of the alignment. However, three chromatograms, when used as the target, resulted in a worsening in the quality of the alignment. Upon visual inspection, this decrease in alignment quality was due to misalignments of several low abundance, co-eluting peaks.

Using the PC-PPMC, all chromatograms when used as a target resulted in an improvement in the quality of alignment, ranging from 211% to 221% (Table 6-4). This small range indicates that all chromatograms were more similar after alignment, regardless of which sample was selected as the target. In addition, the similarity of the PC-PPMC demonstrates the insensitivity of this metric for when evaluating the COW alignment algorithm. It is not clear why diesel samples 3 and 8 resulted in the best alignment. However, the average chromatogram was still among the best choices for a target. The use of the average chromatogram as the target is advantageous because it has been shown to result in good alignment, without requiring testing of all possible chromatograms.

Table 6-3. Percent change in the standard deviation of the peak maxima of selected peaks (PC-SDRT) using the correlation optimized warping alignment algorithm with a warp of 2 and a segment size of 75, with each sample chromatogram as well as the average chromatogram serving as the target. A decrease in the PC-SDRT of the retention time indicates an improvement in alignment.

Target Chromatogram	PC-SDRT
Average	-14.2
D5B	-11.8
D6A	-11.0
D10A	-10.9
D9B	-9.7
D8B	-9.4
D3B	-7.2
D4C	-6.8
D7B	-6.7
D10C	-6.7
D9C	-6.4
D9A	-6.2
D8C	-5.9
D6B	-5.7
D5A	-5.5
D3A	-4.7
D8A	-4.0
D3C	-4.0
D4A	-3.2
D10B	-2.2
D7A	-0.1
D4B	-0.0
D5C	0.7
D6C	5.8
D7C	12.1

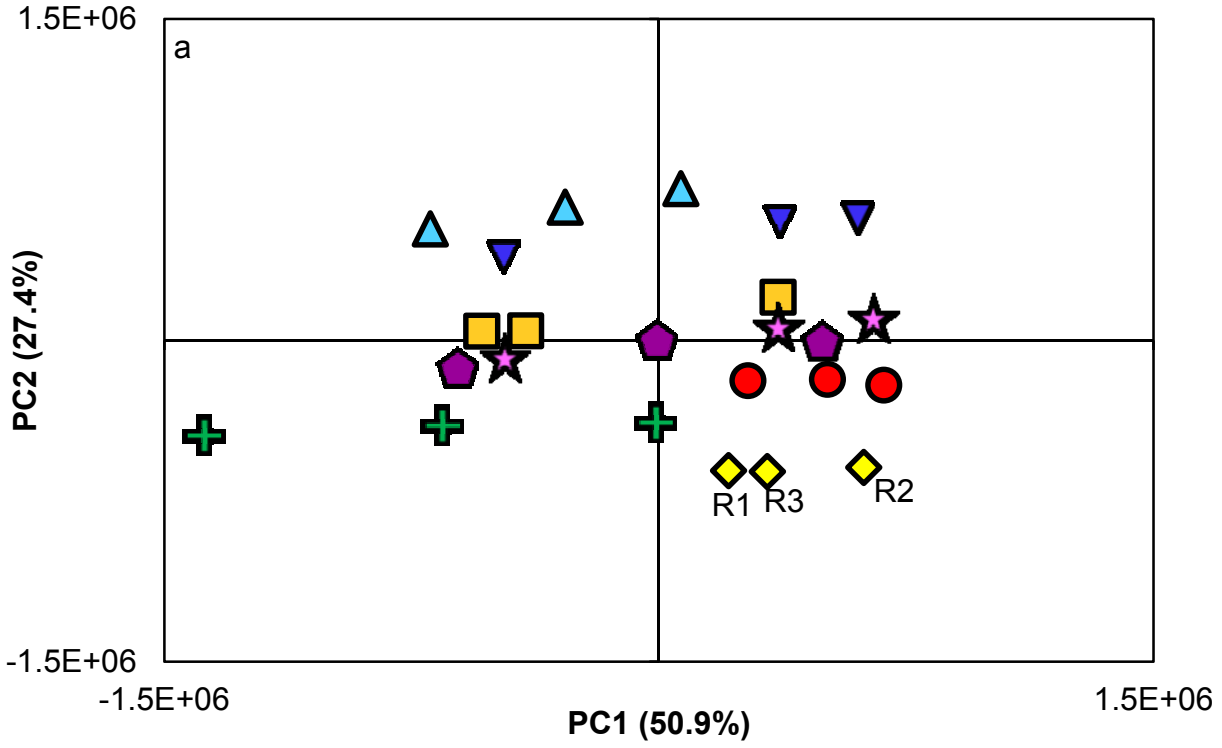
Table 6-4. The sum of the percent change of the Pearson product moment correlation coefficients (PC-PPMC) using the correlation optimized warping alignment algorithm with a warp of 2 and a segment size of 75, with each sample chromatogram as well as the average chromatogram serving as the target. An increase in the sum of the PC-PPMC indicates an improvement in alignment.

Target Chromatogram	PC-PPMC
D3B	221
D8A	221
D8B	220
D3C	220
D3A	219
Average	219
D4A	219
D6A	219
D7A	219
D6C	218
D4C	218
D9A	217
D5A	217
D7B	216
D4B	216
D6B	216
D7C	215
D5C	215
D10C	215
D5B	214
D10B	214
D9B	214
D10A	214
D9C	212
D8C	211

6.4. Effect of Retention Time Alignment on PCA Scores Plot

6.4.1. Visual Assessment

PCA was performed after baseline correction, smoothing, and alignment. In comparing the scores plot (Figure 6-5) with baseline correction and smoothing (a) and with baseline correction, smoothing, and alignment (b), only a small enhancement in clustering of replicates is observed, specifically in replicates of Diesel 4 (orange squares) and Diesel 5 (yellow diamonds). The total variance accounted for in PC1 and PC2 increased from 78.3% to 84.9% after alignment. Similarly, there were only small changes observed in the loadings plots (Figure 6-6) of PC1 (a) and PC2 (b). Prior to alignment (Figure 5-6 and Figure 5-8), the loadings plot for PC1 contained derivative-shape peaks, indicative of misalignments [18] for C_{10}-C_{14}. After alignment, the negative portions are no longer observed, indicating that there is no longer misalignment of these peaks. The loadings plot for PC2 remained largely unchanged after alignment.



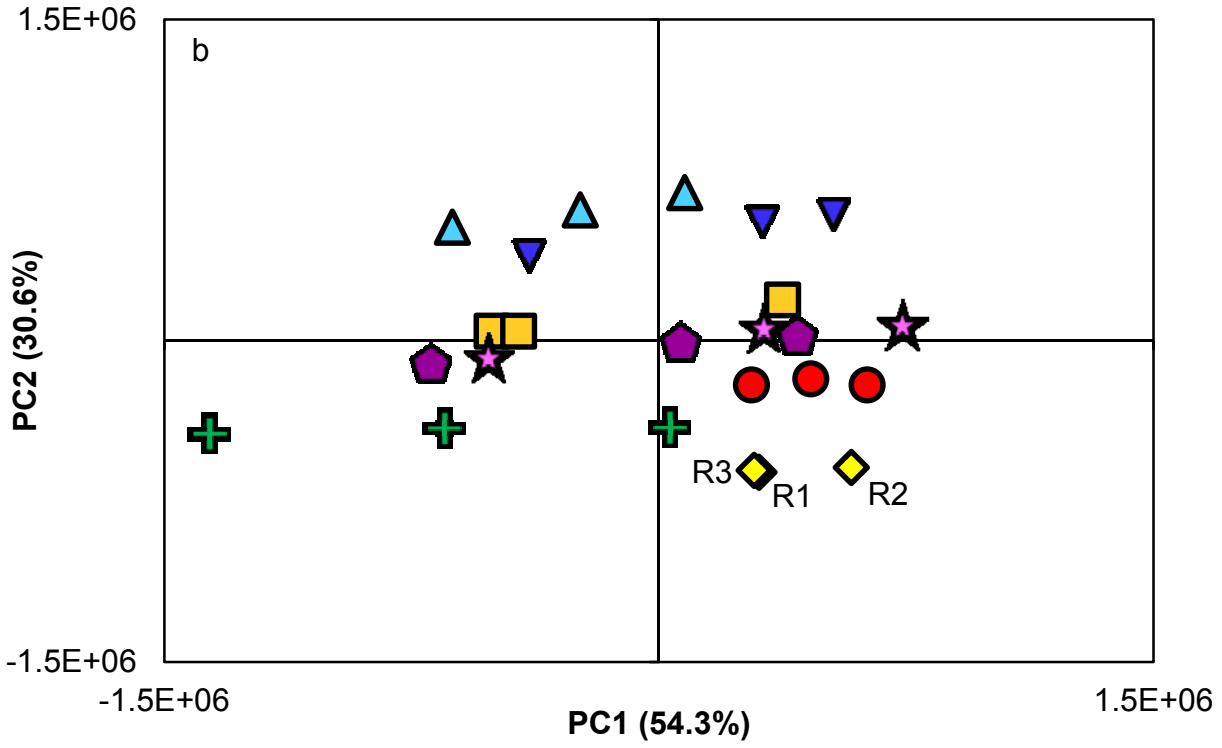


Figure 6-5. PCA scores plot of eight diesel chromatograms in triplicate after baseline correction and smoothing (a) and after baseline correction, smoothing, and alignment (b).

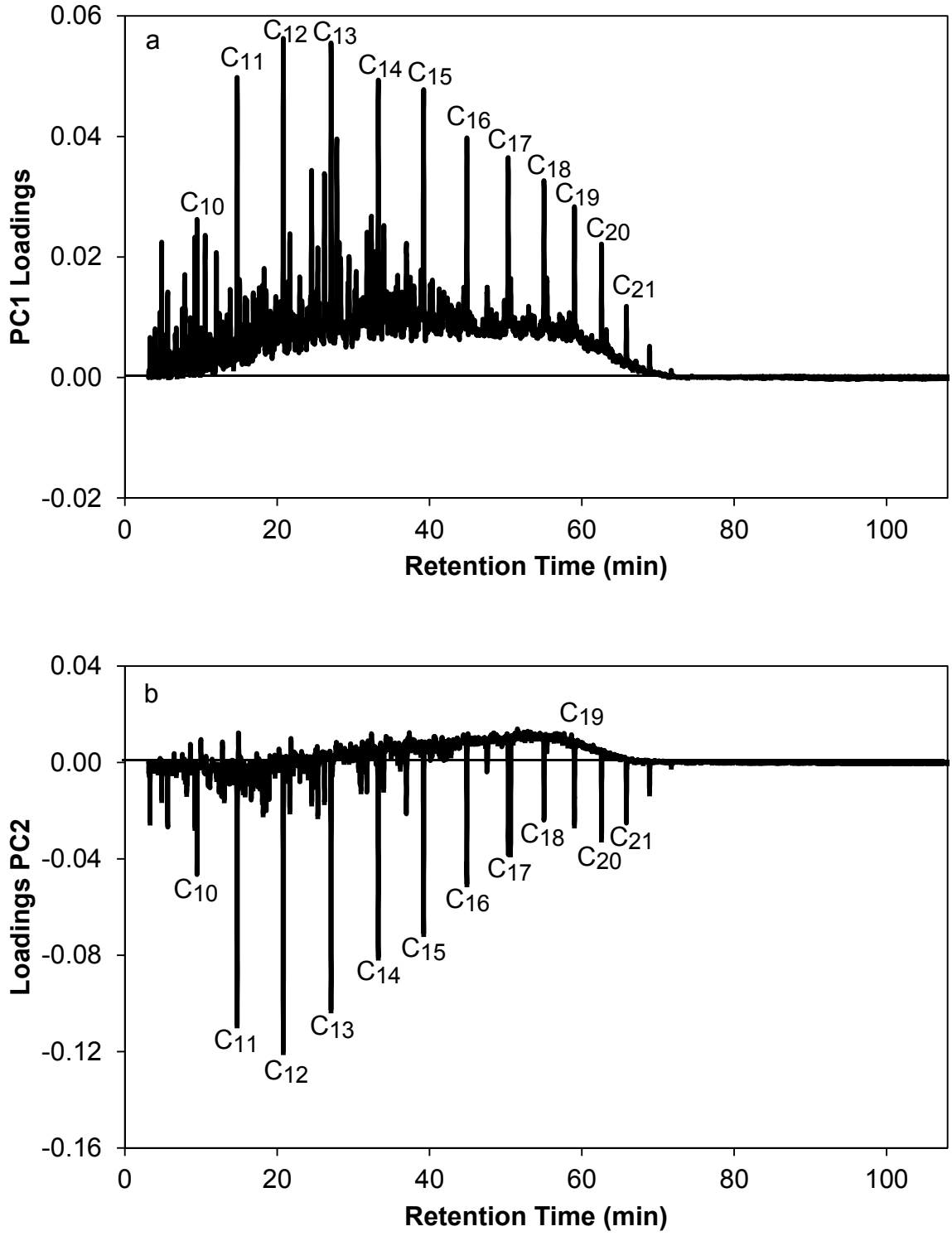


Figure 6-6. Loadings plot for PC1 (a) and PC2 (b) after baseline correction, smoothing, and alignment.

Diesel 5, represented by the yellow diamonds was previously discussed in Chapter 3 and will be again used to highlight the applied data pretreatment and resulting positioning on the scores plot. Prior to alignment, the replicates of Diesel 5 are spread along PC 1 (Figure 6-5a). After alignment using the COW algorithm with a window size of 2 and a segment size of 75, replicates 1 and 3 are closely clustered while replicate 2 is not. The source of this change in clustering of replicates 1 and 3 is due in large part to misalignments of several normal alkanes. In Figure 6-7a, the dodecane peak in replicate 1 (red) of Diesel 5 is misaligned from replicates 2 (blue) and 3 (green). After alignment (Figure 6-5b), the three replicates are well aligned. However, replicate 2 is still at a higher abundance than replicates 1 and 3. This is reflected in the scores plot (Figure 6-5b) which shows that after alignment, replicates 1 and 3 are clustered together, while replicate 2 is still not clustered.

Figure 6-8 shows the scores plot after baseline correction, smoothing, and total area normalization (a) and after baseline correction, smoothing, alignment, and total area normalization (b). The total percent variance accounted for on the first 2 PCs increased from 77.4% with no data pretreatment (Figure 2-6) to 88.7% after baseline correction, smoothing, alignment, and normalization. After normalization, replicates are clustered and several groupings of the samples can now be observed. Diesels 3 and 7 are positioned closely, as are Diesels 4, 9, and 10, as well as Diesels 6 and 8. Diesel 5,

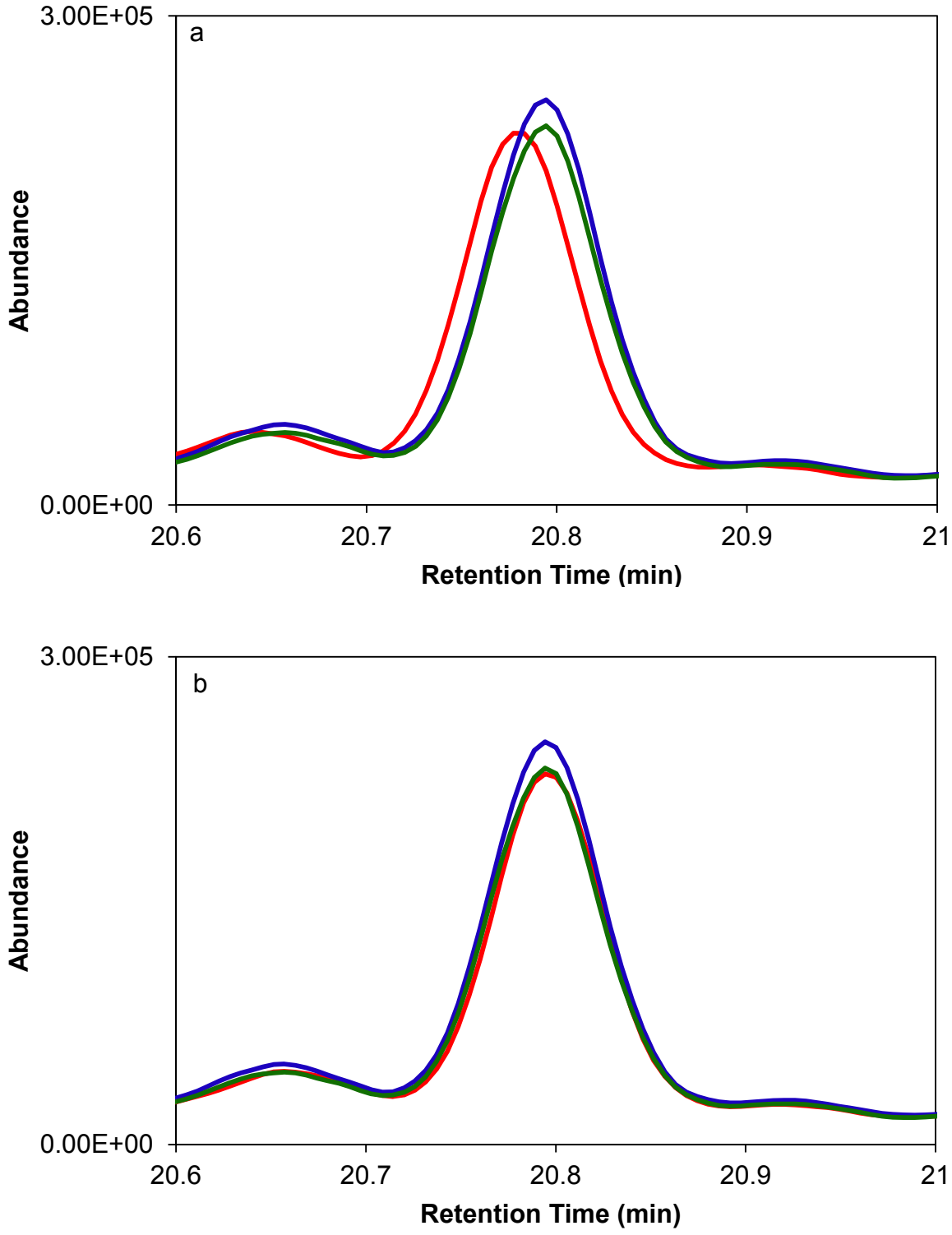
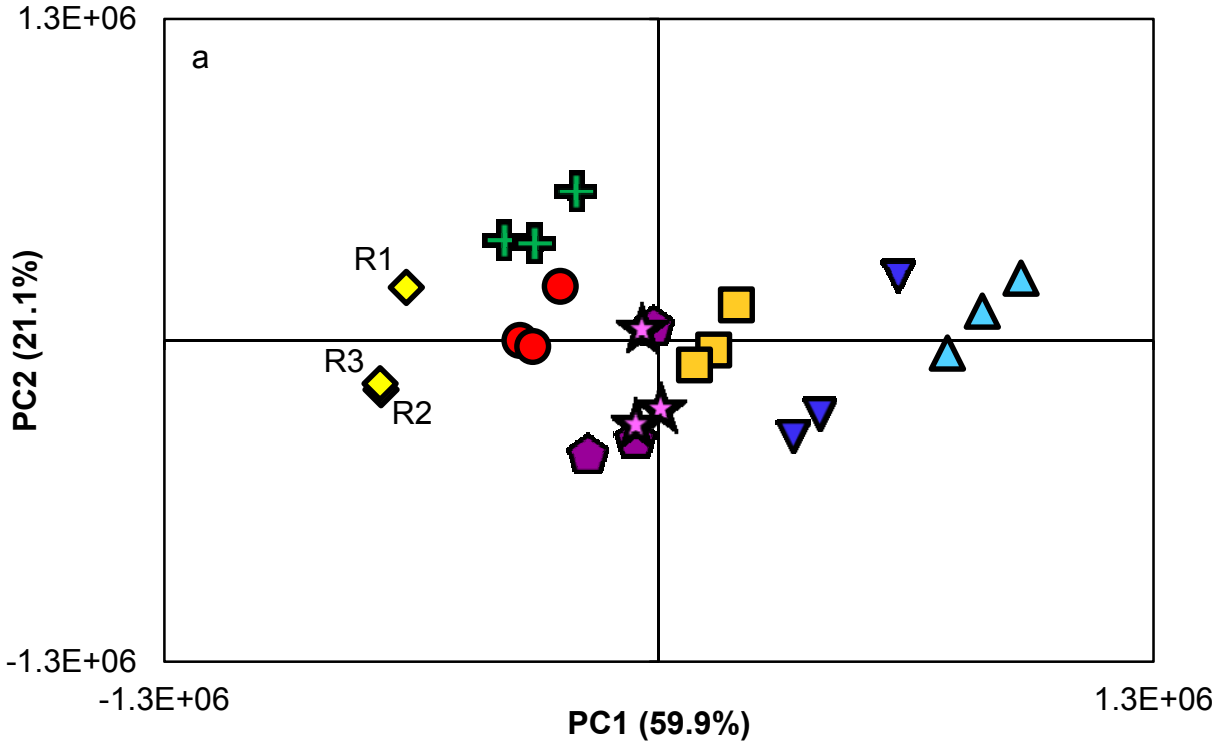


Figure 6-7. An expanded region of dodecane in three replicate chromatograms of Diesel 5 before (a) and after (b) alignment. Each replicate is indicated by a different color (replicate 1: red, replicate 2: blue, replicate 3: green).



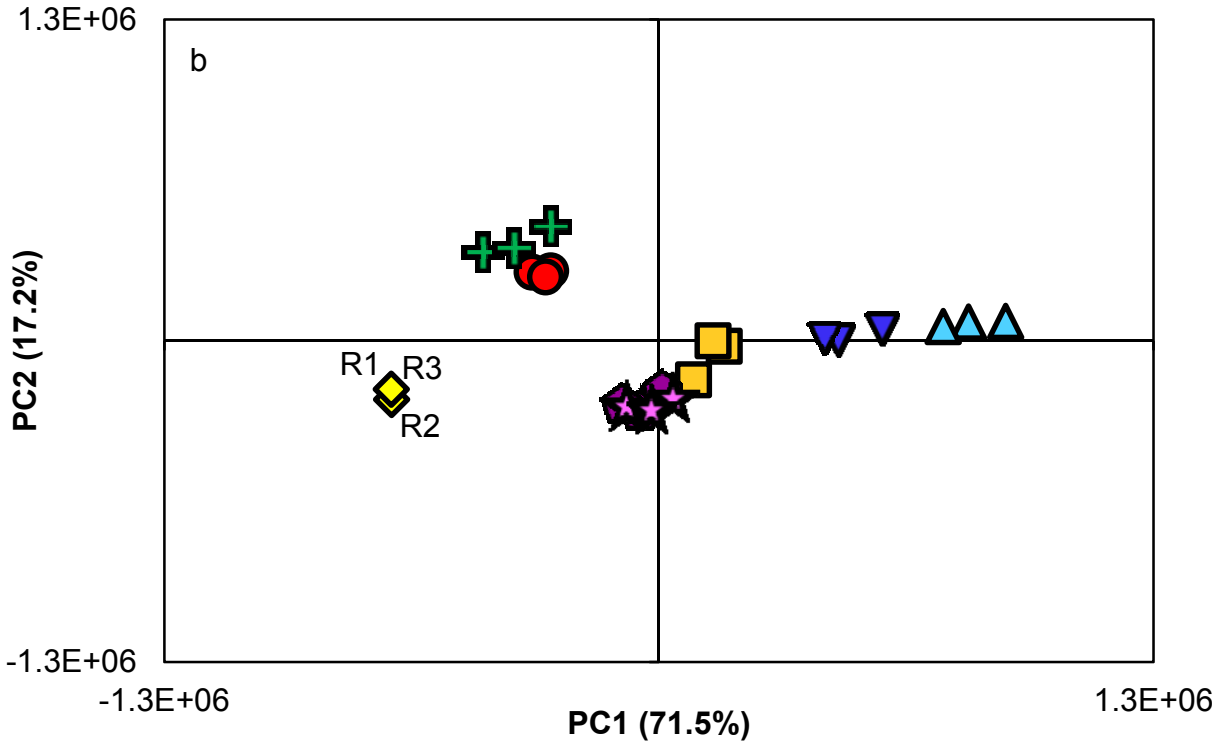


Figure 6-8. PCA scores plot of eight diesel chromatograms in triplicate after baseline correction, smoothing, and normalization (a) and after baseline correction, smoothing, alignment and normalization (b).

however, is discriminated from the other diesels. Based on the loadings plots (Figure 6-9), the positioning of samples on PC1 is based mostly on the abundance of the normal alkanes, particularly the short-chain normal alkanes, which load positively on PC1 (Figure 6-9a). The variation in the short chain alkanes could arise from differences in the distillation of the fuel or from the mixing of different summer and winter diesel blends in the storage tanks at each service station. Also on PC1, a portion of the baseline between 40 and 60 minutes is loading negatively. This is the retention time region that shows an increase in the baseline for Diesel 5 (Figure 2-1). This corresponds to the negative positioning of Diesel 5 on the PCA scores plot (Figure 6-8a). The positioning of samples on PC2 is influenced most by the long-chain normal alkanes loading positively as well as a few of the most volatile compounds. Most of the short chain normal alkanes as well as some branched alkane and aromatic compounds are loading negatively on PC2 (Figure 6-9b). PC2 is differentiating compounds using the unimodal versus bimodal distribution.

Using the chromatograms (Figure 2-1), the scores plot (Figure 6-8b) and the loadings plots (Figure 6-9) the positioning of each diesel sample can be explained. As mentioned previously, Diesels 3 and 7 have a unimodal distribution of the normal alkanes (rather than the bimodal distribution observed in the other diesel samples) and should be positioned close together. Diesels 3 and 7 are closely associated on the scores plot and are positioned negatively on PC1 and positively on PC2. Diesels 3 and 7 have a lower abundance of short-chain alkanes (which load positively on PC1), and a higher abundance of long-chain normal alkanes (which load positively on PC2). Diesels 6 and 8 are positioned positively on PC1, while Diesel 5 is positioned negatively on PC1. For

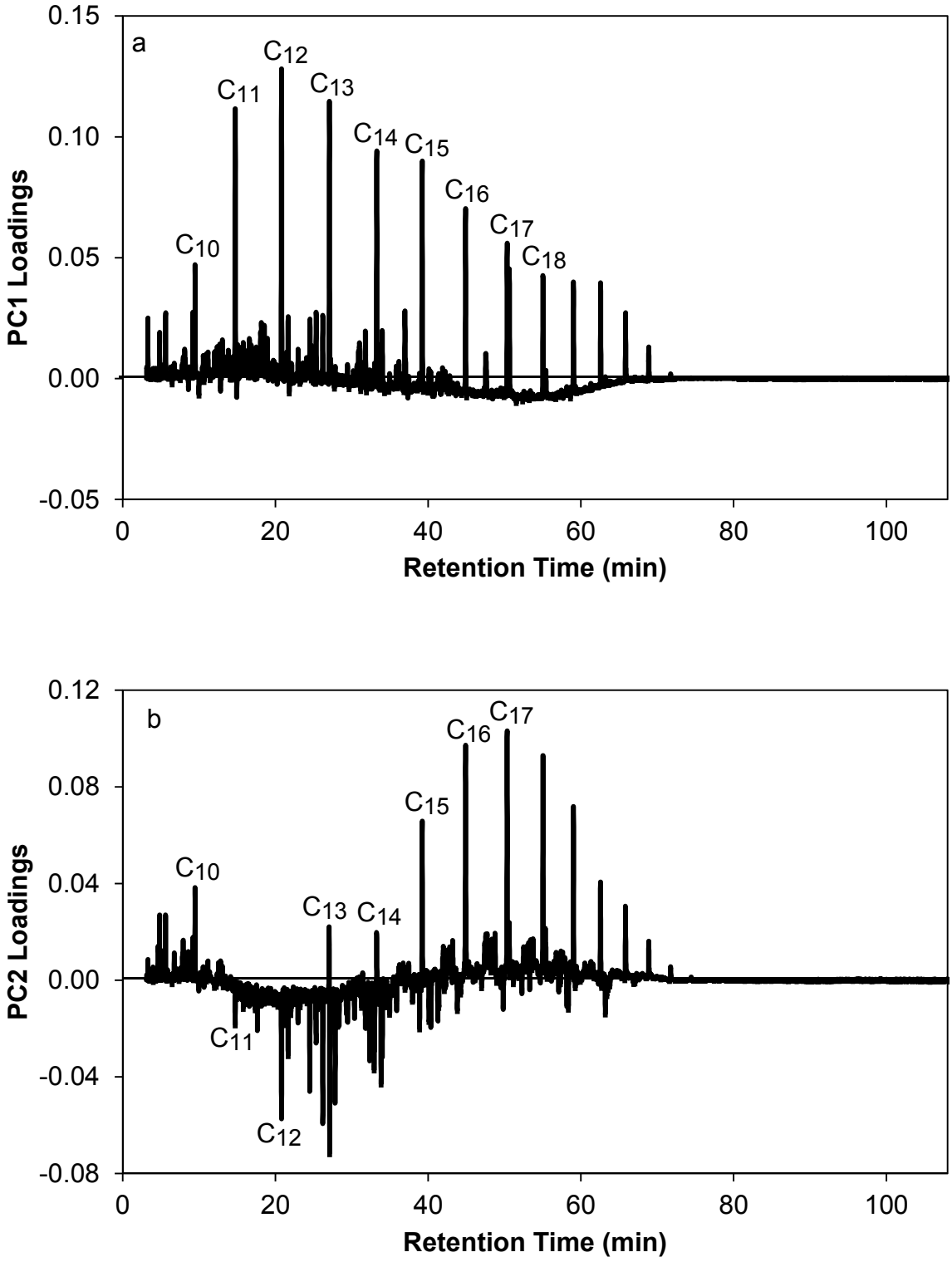


Figure 6-9. Loadings plot for PC1 (a) and PC2 (b) after baseline correction, smoothing, alignment, and normalization.

these samples, the distribution of peaks appears similar, however, Diesels 6 and 8 have an overall higher abundance while Diesel 5 has a lower abundance than other samples. Diesels 4, 9, and 10 are all positioned near the origin, which shows that they are not being well differentiated using PC1 and PC2 and are not well described by the variance on PC1 or PC2.

As discussed at the beginning of this section, alignment resulted in replicates 1 and 3 of Diesel 5 becoming more closely clustered. Also, after applying normalization (Chapter 3), replicates 2 and 3 became more closely clustered. When alignment and normalization are both applied, all three replicates become closely clustered. After alignment and normalization, non-chemical variation from shifts in retention time and differences in abundance have been minimized (Figure 6-10), resulting in replicates that are more similar and therefore clustered closer together.

6.4.2. Quantitative Assessment

The percent change in the clustering of replicates (PCC) on PC1 and PC2 was again used to assess the effect of data pretreatment on the samples in the PCA scores plot. After baseline correction, smoothing, and alignment, there was a 5.9% increase in the clustering (Table 6-5), while there was only a 0.5% increase when smoothing and baseline correction were applied and no change in the clustering when baseline correction alone was applied. The largest increase in the PCC was observed after normalization was also applied. The PCC increased 85.1% when total area normalization was applied and 71.8% when single peak normalization was applied. This shows that after normalization, alignment is the next most important data pretreatment.

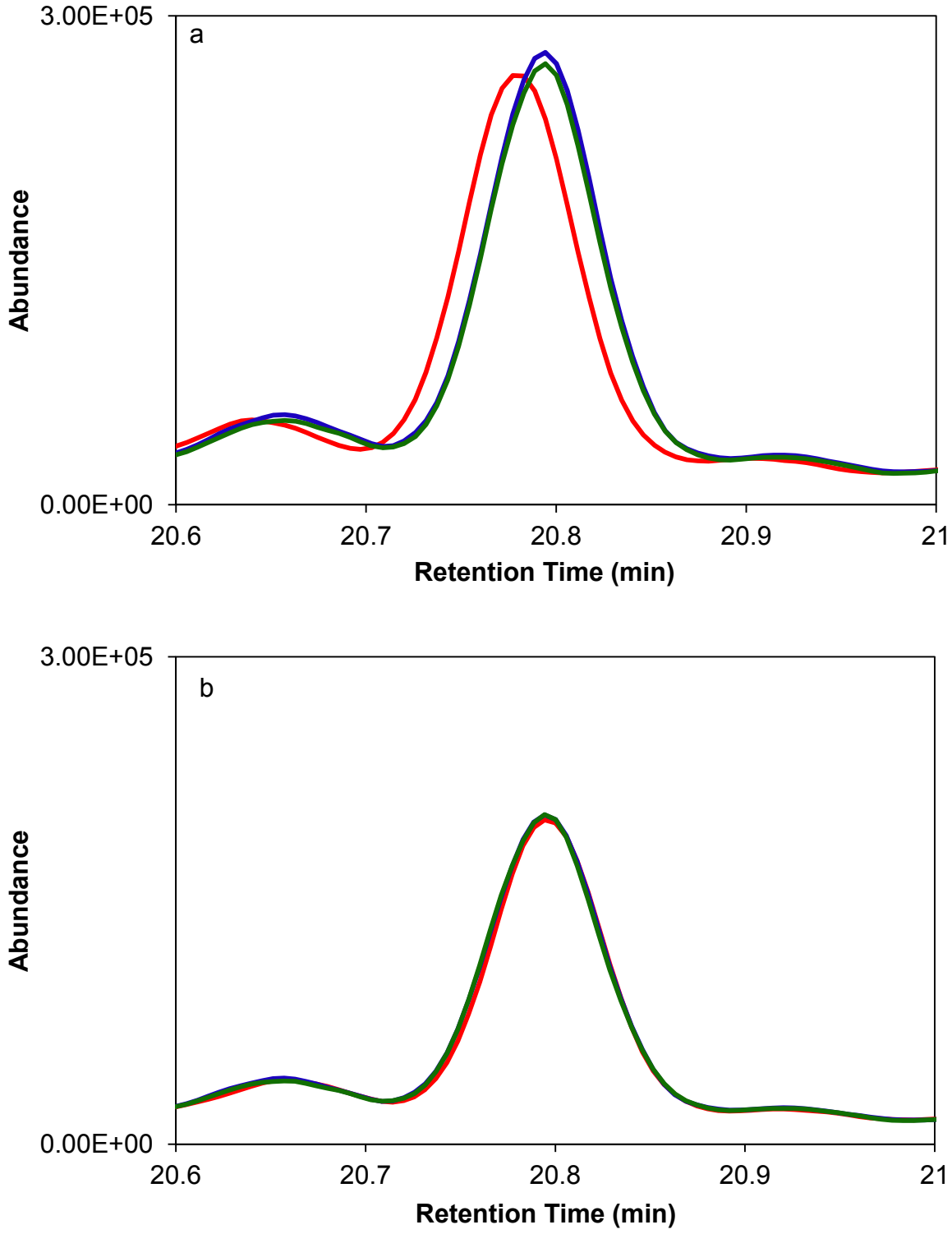


Figure 6-10. An expanded region of dodecane in three replicate chromatograms of diesel 5 after baseline correction, smoothing, and alignment (a) and after baseline correction, smoothing, alignment, and normalization (b). Each replicate is indicated by a different color (R1: red, R2: blue, R3: green).

Table 6-5. The average percent change in the clustering of replicates (PCC) after the listed pretreatment procedures including baseline correction using the extracted ion profiles (EIP fit), smoothing using fast Fourier transform smooth with 2 points (FFT2), alignment using the correlation optimized warping algorithm with a warp of 2 and a segment of 75 (COW 2, 75) and normalization using total area (Area) and single peak (Peak) normalization methods.

Baseline Correction	Smoothing	Alignment	Normalization	PCC
-	-	-	Area	45.1
-	-	-	Peak	59.0
EIP fit	-	-	-	0.0
EIP fit	-	-	Area	44.6
EIP fit	-	-	Peak	58.7
EIP fit	FFT 2	-	-	0.5
EIP fit	FFT 2	-	Area	46.5
EIP fit	FFT 2	-	Peak	62.1
EIP fit	***FFT 2***	***COW 2, 75***	***-***	***5.9***
EIP fit	***FFT 2***	***COW 2, 75***	***Area***	***85.1***
EIP fit	***FFT 2***	***COW 2, 75***	***Peak***	***71.8***

6.5. Summary

Retention time misalignments were observed in overlaid chromatograms of diesel samples that were analyzed over the course of approximately two weeks. After processing each chromatogram with a PM or a COW retention time alignment algorithm, these misalignments were reduced. Both alignment algorithms resulted in a minimization of the non-chemical sources of variation when appropriate parameters were selected. For the PM algorithm, window sizes of 3 - 8 resulted in a similar quality alignment. For the COW algorithm, many combinations of the warp and segment size resulted in improved alignment. For both alignment algorithms, the selection of an appropriate target is critical. Rather than testing every possible chromatogram as the target, the average target proved to be a fast and effective choice for the target. This demonstrates that different alignment algorithms can result in a similar quality of alignment, even over a range of different parameters.

These results indicate that the optimization of alignment is not necessary, at least for this work. Because samples were collected over a short period of time using temperature-programmed GC, there is a reduced need for retention time alignment. However, when samples are collected over a long period of time or collected using other thermal modes, such as isothermal GC, there would likely be an increased need for optimization of the retention time alignment.

REFERENCES

REFERENCES

[1] K.M. Pierce, J.S. Nadeau, R.E. Synovec, in: C.F. Poole (Ed.), Gas Chromatography, Elsevier, Waltham, MA, 2012.

[2] M. Daszykowski, B. Walczak, Trac-Trends Anal. Chem., 25 (2006) 1081.

[3] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y.V. Heyden, J. Chromatogr. A, 1118 (2006) 199.

[4] G. Malmquist, R. Danielsson, J. Chromatogr. A, 687 (1994) 71.

[5] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A, 1096 (2005) 101.

[6] Infometrix Inc., LineUp: Chromatographic Alignment Tool, Version 3.0, 2010.

[7] J.S. Nadeau, B.W. Wright, R.E. Synovec, Talanta, 81 (2010) 120.

[8] B.K. Lavine, D. Brzozowski, A.J. Moores, C.E. Davidson, H.T. Mayfield, Anal. Chim. Acta, 437 (2001) 233.

[9] R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson, J. Chemometr., 17 (2003) 573.

[10] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A, 996 (2003) 141.

[11] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A, 805 (1998) 17.

[12] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr., 18 (2004) 231.

[13] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta, 456 (2002) 77.

[14] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemometr., 20 (2006) 484.

[15] S. Peters, E. van Velzen, H.G. Janssen, Anal Bioanal Chem, 394 (2009) 1273.

[16] K. Prather, Using Multivariate Statistical Procedures to Identify Ignitable Liquid Residues in the Presence of Interferences, Michigan State University, Ann Arbor, 2011.

[17] J.M. Baerncopf, Association and Discrimination of Ignitable Liquids from Matrix Interferences using Chemometric Procedures, Michigan State University, Ann Arbor, 2009.

[18] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith, Anal. Bioanal. Chem., *394* (2009) 2049.

[19] L.J. Marshall, Association and Discrimination of Diesel Fuels using Chemometric Procedures for Forensic Arson Investigations (Masters Thesis), Michigan State University, Ann Arbor, MI, 2008.

[20] A.M. Hupp, L.J. Marshall, D.I. Campbell, R.W. Smith, V.L. McGuffin, Anal. Chim. Acta, 606 (2008) 159.

CHAPTER 7: CONCLUSIONS AND FUTURE WORK

7.1. Conclusions

Multivariate statistical analyses are being applied to complex data in a growing number of fields, including forensic science. Discriminating complex profiles is critical in many areas of research from arson investigation to proteomics. As shown from this work, complex and similar samples are not well discriminated using principal component analysis (PCA). Identifying small differences in complex samples is often complicated by non-chemical sources of variation, such as differences in abundance, shifts in retention time, noise, and signals from background compounds, which can often be identified as the greatest sources of variance among samples. This limitation can be overcome by utilizing appropriate data pretreatment methods to minimize the non-chemical sources of variation in a dataset. Data pretreatment procedures cannot be treated with a "black box" approach. Therefore, visual examination and metrics to monitor the application of data pretreatment are required. The analyst must take care to ensure that proper pretreatment procedures are being applied and the assumptions that are made prior to applying the pretreatments are valid.

In this work, eight different diesel samples were each analyzed in triplicate by gas chromatography-mass spectrometry (GC-MS). Four data pretreatment procedures (*i.e.*, baseline correction, smoothing, retention time alignment, and normalization) were applied to the chromatograms to minimize the non-chemical variation. For each type of data pretreatment applied, several different procedures were tested. For baseline correction, the background subtracted baseline function, the subtraction of an extracted ion profile,

and the subtraction of the baseline using a modeled function were compared. The Savitzky-Golay and the fast Fourier transform smoothing algorithms were compared for their ability to reduce noise in the chromatogram. Misalignments were corrected using a correlation-optimized warping algorithm and a peak-matching algorithm. Normalization was compared using a single peak normalization and a total area normalization. After each pretreatment, chromatograms were compared using the developed metrics and PCA was performed.

The metrics that were developed provide a rapid method for evaluating the effect of each pretreatment on the chromatograms. Each metric was designed to evaluate the increase in similarity obtained by minimizing the non-chemical sources of variation among the replicates. These metrics also allow for a comparison and optimization of parameters associated with each data pretreatment procedure. The evaluations that were utilized include a visual examination of the chromatogram, a metric measuring the change in the chromatogram after pretreatment, a visual examination of the PCA scores plot, and the percent change in the clustering of replicate samples in the PCA scores plot.

The minimization of non-chemical sources of variation improves the multivariate statistical analysis in two ways. First, after data pretreatment, replicate chromatograms are more similar, therefore replicates will cluster more closely using PCA. Second, and more importantly, when PCA is applied, more chemical (rather than non-chemical) differences will be identified as the greatest sources of variance. Therefore, the loadings plots will contain variables that reflect chemical differences between samples, rather than non-chemical differences due to instrumental variation. As replicate chromatograms are

chemically identical, replicates should cluster better and samples should be well discriminated from one another following appropriate pretreatment.

In baseline subtraction, the background subtracted baseline, the subtraction of the extracted ion profiles, and the subtraction of the baseline using a modeled function all resulted in a reduction in the baseline. The subtraction of the modeled function allowed for a reduction in the baseline, without any reduction in signal from the peaks in the chromatogram, ensuring that chemical information was not inadvertently removed from the chromatogram. Therefore, it was selected as the most appropriate option for baseline correction. The baseline of the chromatograms in this work was not a major source of variation, because the chromatograms were generated over a relatively short period of time. However, when chromatograms that have been generated over a long period of time are compared, the difference in baseline may become more significant. Additionally, when the GC is operated at high temperature or the column is old, stationary phase degradation is more prevalent and baseline correction will become more critical.

The Savitzky-Golay and fast Fourier transform smoothing algorithms both resulted in noise reduction in the chromatogram. By changing the number of points in the smooth and the polynomial order (only for the Savitzky-Golay smoothing algorithm) the degree of smoothing in each chromatogram was similar. The similar reduction in noise will result in similar clustering of replicates on the scores plot after PCA. Therefore, optimization and careful selection of the smoothing algorithm is not necessary; however, care must be taken to ensure that peak distortion does not occur. Based on this work, significant peak distortion was not observed until there was more than a 45% reduction in noise. For this work, the fast Fourier transform smoothing algorithm with 2 points was selected but this

algorithm has similar performance to several combinations of the Savitzky-Golay smoothing algorithm.

A peak-matching and a correlation optimized warping algorithm were compared for peak alignment. Nearly all combinations of window size (for the peak-matching algorithm) or warp and segment size (for the correlation optimized warping algorithm) resulted in an improvement in alignment. Most parameters resulted in similar quality alignment. The number of points the peaks could be shifted (called the window for the peak-match algorithm and the warp for the correlation optimized warping algorithm) is an important consideration. Ideally, the chromatograms should be overlaid and visually inspected. The window or warp should then be selected based on the number of points that each peak needs to shift in order to be aligned. If the window or warp is too small, the peaks cannot be aligned; if it is too large, peaks may be shifted too far and aligned to the wrong compound. Target selection can also influence the alignment. The target must include all of the compounds that require alignment. For this work, the average target was selected and resulted in good alignment, without the need for optimization. The correlation optimized warping algorithm with a warp of 2 and segment size of 75 was selected.

Two different normalizations were compared, a total area normalization and a single peak normalization. Both normalization procedures resulted in the largest improvement of clustering of replicates compared to the other pretreatment procedures, but each normalization is based on a different assumption. The total area normalization assumes that the total area is the same for all chromatograms. The single peak normalization assumes that there is a single compound within each chromatogram that is

at the same concentration. When choosing a normalization procedure, an analyst must decide which of these assumptions are correct for the specific application and data. In this work, the total area normalization was selected. Because each diesel sample is from a different source, there is no reason to assume that any single compound has the same concentration across all of the samples. However, as there are so many different compounds, in this case, the total area would likely be equivalent.

After applying data pretreatment procedures, replicates on the PCA scores plot were shown to cluster more closely. This is due to the removal of non-chemical sources of variation. The scores plot shown prior to data pretreatment (Figure 2-6) shows that the replicates are spread along PC1 and the samples are separated along PC2. Replicates are chemically the same, so any spread along PC1 is due to non-chemical sources of variation introduced during the analysis. This means that the loadings plot of PC1 prior to applying data pretreatment (Figure 2-7a) contains only non-chemical sources of variation. Because the samples were separated along PC2 this indicates chemical differences were identified in PC2, prior to application of data pretreatment. Therefore, the loadings plot of PC2 showed chemical differences between samples (Figure 2-7b and Figure 7-1a). After applying data pretreatment procedures, the non-chemical sources of variation were minimized, which was reflected in the PCA scores plot (Figure 6-8). Replicates were positioned close together and samples were differentiated from one another, indicating that PC1 and PC2 contain chemical differences. The loadings plot for PC1 after applying the data pretreatment procedures contained the same

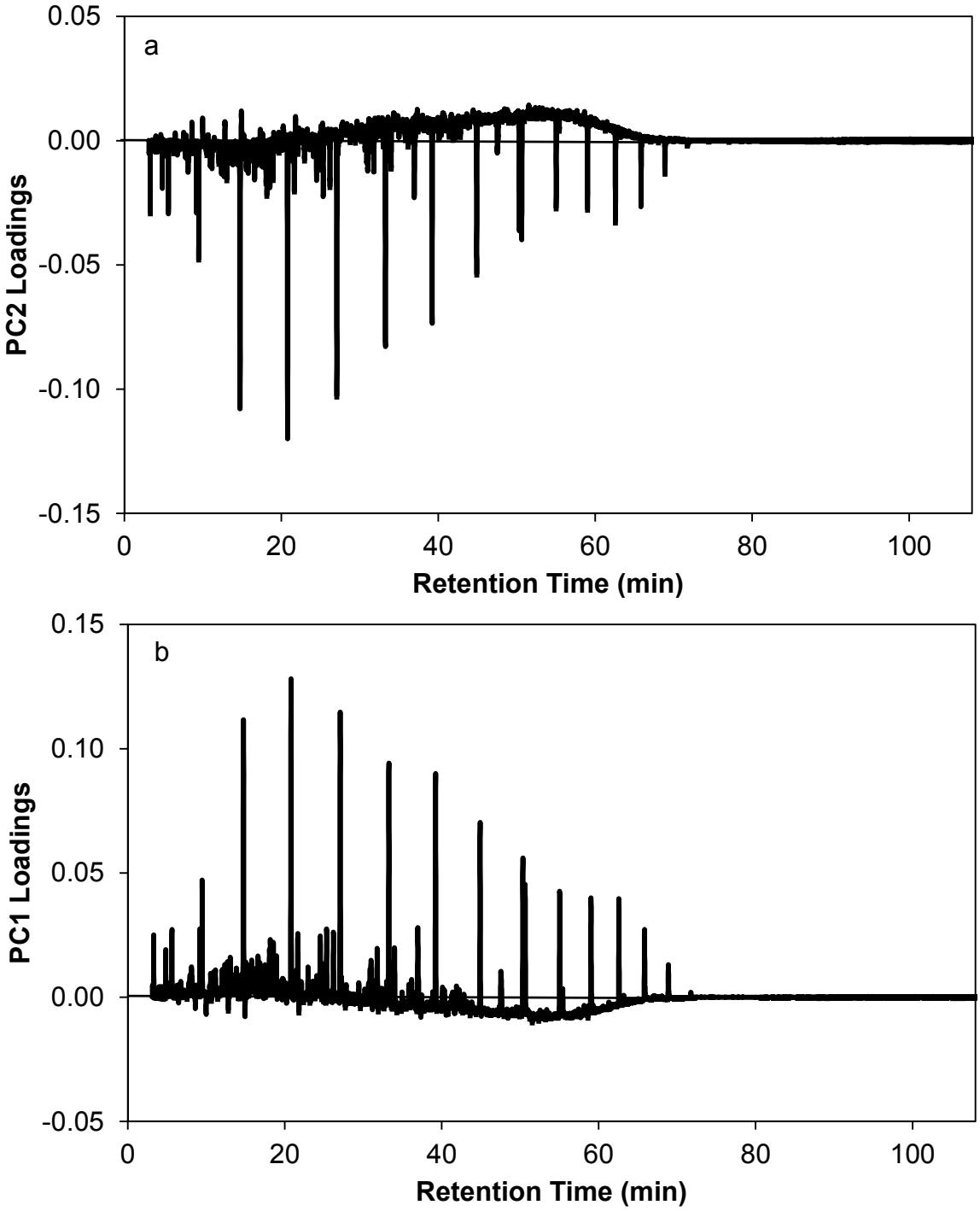


Figure 7-1. Loadings plot for PC2 prior to applying data pretreatment (a) and for PC 1 after applying baseline correction, smoothing, alignment, and normalization (b).

chemical differences that were identified on PC2, prior to data pretreatment (Figure 6-9a and Figure 7-1b). This demonstrates that prior to the application of data pretreatment, PC1, which accounts for the most variance, was only accounting for non-chemical sources of variation. After applying data pretreatment, the non-chemical sources of variation had been minimized and PC1 accounted for chemical differences between samples.

Overall, this work has demonstrated that application of data pretreatment procedures can significantly enhance the discriminatory ability of PCA. For this work, normalization was shown to provide the largest improvement in the clustering of replicates followed by retention time alignment. Smoothing and baseline correction had relatively little effect on the clustering of replicates. Overall, there was an 85% improvement in the clustering of replicates after applying all of the data pretreatment procedures.

Additionally, this work has shown that when multiple pretreatments are applied to the same chromatogram, there is a larger increase in clustering than with a single pretreatment. For example, normalization alone resulted in a 45% increase in clustering, however when baseline correction, smoothing, alignment and normalization were applied, there was an 85% increase in the clustering of replicates.

This work has also shown that optimization of the data pretreatment procedures is not necessary to obtain enhanced clustering of replicates. Most of the parameters tested resulted in a reduction of non-chemical sources of variation, based on the metrics used to evaluate each pretreatment procedure. As previously discussed, not all non-chemical sources of variation were equally prevalent (*e.g.* normalization had a larger effect than

baseline correction, indicating that differences between injections were more variable than the baseline). Therefore, pretreatment selection is more important for non-chemical sources of variation that are more prevalent in the chromatograms.

The application of data pretreatment procedures results in an enhancement of the discrimination of complex and chemically similar mixtures by minimizing the non-chemical sources of variation. In forensic science and many other fields, the comparison of complex samples is becoming more common. Prior to applying multivariate statistical procedures, data pretreatment is commonly utilized. However, for the data pretreatment procedures to be permitted into court, the procedures would have to be shown to not alter the chemical information contained in the chromatograms. In addition, it is critical to demonstrate that samples can be differentiated using chemical information once the non-chemical differences have been minimized. This work provides methodologies for comparing and selecting appropriate pretreatment procedures. It is critical to ensure that the chemical information is not being altered by the pretreatment procedures and to understand the effect of each pretreatment on the chromatographic data.

7.2. Future Work

There are several areas presented in this research that could be further expanded. First, additional research could focus on novel data pretreatment procedures. As shown in this work, each of the data pretreatment procedures did not remove all of the non-chemical sources of variation. Retention time alignment and normalization were shown to result in the largest reductions in non-chemical variation, and would benefit the most for additional investigation. For alignment, developing an algorithm that is capable of

finding co-eluting and low abundance peaks would result in better alignment because more peaks would be identified and aligned. Developing a normalization that is able to normalize the baseline and peak maxima would allow for a more complete normalization.

As part of this work, the precision between replicates was evaluated using each metric. An additional metric could be developed that is capable of demonstrating that chemical compounds between replicates are not changing as a result of the data pretreatment procedures. The metric could be based on an average chromatogram created from multiple replicates. If chemical information is not being lost, the average chromatogram should remain relatively unchanged after applying each data pretreatment procedure. If the average remains unchanged, this will demonstrate that chemical information is remaining, even after applying data pretreatment.

Another area in which this work could be expanded would be to evaluate the effect of data pretreatment on the output from other types of samples (*i.e.* drugs, paints, *etc*.) and instrumentation (*i.e*. infrared spectroscopy, scanning electron microscopy with electron dispersive spectroscopy). Demonstrating that non-chemical variation can be minimized without altering the underlying chemical information would be a critical step in the use of multivariate statistic for forensic cases.

The ultimate goal is to apply multivariate statistics to forensic data. However, work is still required to develop statistical methods and ways to apply those methods to forensic data. This will help to limit bias and assign statistical confidence to forensic comparisons, addressing concerns outlined in by National Academies of Science.