Executive Summary

Description of the Problem

Gasoline and kerosene are the two most common ignitable liquids used as accelerants in arson cases. These products are readily available and can be bought in quantity without arousing suspicion. According to the American Society for Testing and Materials (ASTM) E1618 method, products classified as middle petroleum distillates (MPD) are also commonly used ignitable liquids. MPD are readily available in the form of charcoal lighters, paint thinners, and industrial solvents making them other readily available choices for arsonists.

Fire debris analysis (FDA) is the science that involves examination of fire debris samples performed to detect and identify ignitable liquid residues (ILR). Using the E1618 classification scheme can readily identify the ignitable liquid residue as a gasoline, MPD, or kerosene; however, comparing two samples (i.e. residue from fire debris to a liquid sample in the suspect's possession) to determine if they are from the same source is much more problematic. The field of fire investigation is very subjective because of the high degree of visual interpretation needed. Recently a number of different statistical based methods have been published. No single method for the statistical comparison of ignitable liquids has proven to be the best in all cases involving fire debris analysis. Although classification of the type of liquid used to create a fire is important, distinguishing within a specific class is more necessary to the world of fire investigation. Therefore, it is important, however complicated it is, to be able to compare two or more samples in a case, to determine if the ignitable liquid residues share a common source. In addition, determination of the precision and error rates for the comparison of IL samples is also important for the evidence to have legal standing in many Daubert states.

The recent National Academy of Science report recommends that pattern recognition techniques (of which ignitable liquid residue analysis is one) have established error rates to meet Daubert rules of evidence in court.

Purpose, Goals and Objectives:

In this study, high resolution GCMS and target compound analysis was used to develop a valid method in order to statistically differentiate between different kerosene and MPD samples. Kerosene and MPD, unlike gasoline, are simple distillation products from crude oil and should be strongly related to the petroleum from which it was distilled. The relative concentrations of the key components in kerosene and MPD from a refinery will often change daily because of the variety of sources that distribute crude oil. This variation in concentration of the key compounds can provide sufficient variability for comparison between individual samples. Kerosene was evaluated from a single refinery to first establish a target compound list. Applying target compound ratio analysis to all of the kerosene samples established a method for differentiation. For MPD, a selection of commercial products (paint thinners and charcoal lighters) was used for establishing the corresponding peak ratios along with samples from the Ignitable Liquid Collection maintained by the Technical Working Group for Fire and Explosives (TWGFEX). The key objectives can be summarized as:

1. Collecting a large set of kerosene and MPD samples and analyzing them by high resolution GCMS.
2. Selecting a number of candidate target compounds common to most samples which elute in pairs relatively close in elution time and have repeatable peak area ratios (ideally within 5% RSD) within samples, but variable between samples of different origin.
3. Using multivariate statistical methods, determine those candidate peak ratios which are best for discriminating between samples for kerosene and MPD.

4. Determine the “false positive rate” when comparing all the samples in the data set.
5. Determine the robustness of the method for evaporated and simulated burned samples.

Research Design and Methods

Kerosene samples were obtained from the Marathon-Ashland refinery in Catlettsburg, Kentucky near Marshall University. The quality control laboratory at the refinery receives samples from all Marathon refineries and distribution terminals in a nine state region. Additional samples were purchased locally from service stations as well as home improvement stores. The original goal of obtaining samples from a wider region proved impractical due to Department of Transportation restrictions requiring ‘certified’ shippers in each state. For MPD samples, a number of commercial products which frequently contain MPD (charcoal lighters, paint thinners and solvents) were purchased locally and supplemented by samples from the Ignitable Liquid Reference Collection administered by TWGFEX.

GCMS analysis was performed using a 60m polymethylsiloxane column which had been shown to separate gasoline individualization studies. Usually a 30m column is recommended for routine fire debris analysis for ignitable liquid residues for classification by the ASTM E1618 method. The longer column used in this study allows for better separation of closely eluting compounds at the expense of longer analysis time. Integration of peak areas utilized standard methods included in the instrument software as well as routines developed using MATLAB.

Statistical methods used to establish error rates included a number of different forms of multivariate statistics including principal components analysis (PCA), projected difference resolution (PDR) mapping, fuzzy-rule building expert systems (FuRESs), Pearson correlation coefficient, Spearman’s-rho rank correlation coefficient and Kendall’s-tau coefficient.

Results:

A total of 76 kerosene samples were analyzed at least three times each. Of that data set 44 were selected for developing the kerosene model. Thirty-six target compounds were selected and 35 peak area ratios determined. For MPD, 44 commercial samples of charcoal lighters, paint thinners and other solvents were determined by E1618 method to be classified as MPD. These samples along will 111 samples obtained from the ILRC which were classified as MPD were analyzed. Forty-one target compounds were selected resulting in 33 peak area ratios for the MPD samples.

For the kerosene analyses, the peak integration routines in the instrument software did not give consistent results as the MATLAB routines processing the raw data files. Differences in background correction for the increasing baseline from column bleed at higher temperatures may have contributed to this observation. For MPD, which elute well before significant column bleed, like previous studies with gasoline, were not affected and did not need special background correction routines.

For kerosene, the FuRES models yielded correct classification rates greater than 90% for discriminating between samples. PDR mapping, a new method for characterizing complex data sets was consistent with the FuRES classification result. For MPD, the Kendall's-tau metric for 'association/no-association' was utilized to demonstrate 'association/no-association' true positive rates in excess of 95% with false positive rates ~5%.

Evaporation studies, although limited, gave results as expected. Lower boiling point components were lost and those peak area ratios containing those components were affected. Some preliminary studies on the effects of substrate were inconclusive and need to be expanded. Such studies are on-going related to substrate effects on E1618 classification.

Conclusions:

This method shows promise for comparison of ignitable liquids in these two classes as well as gasoline, which has been previously reported. Although the false positive rate (less that 5%) will never approach that of DNA or some other tests, it is more likely that two samples can be excluded as being significantly different. Additional work, especially with a larger data set of kerosene samples, optimization of software routines to analyze data from a variety of vendors' instruments and continued work on the effects of evaporation and substrate are needed.

As with any determination of the presence of an ignitable liquid in fire debris, this does not necessarily mean that the ignitable liquid was used as an accelerant, but may be incidental to the scene.

*Implications for policy and practice***.**

Application of high resolution GCMS analyses for comparison between neat ignitable liquids is possible using pattern matching techniques described in this report with some reasonable degree of statistical validity. A reference database of target compound ratio analyses of gasoline, kerosene and MPD samples such as from the ILRC should allow individual laboratories to utilize these techniques without having to invest considerable time, manpower and expense to recreate the data. Key to the adoption of such a database will be additional studies incorporating different instrumentation and laboratories utilizing a 'round robin' approach.

Implications for further research.

Initial studies in our laboratory have shown that there may be some effects due to substrate and/or evaporation, thus additional studies are needed to test the 'robustness' of this method.

I. Introduction:

Statement of the problem: This research applied target compound analysis that was utilized for gasoline individualization and applied it to medium petroleum distillates (MPD) and kerosene which may be used as accelerants in arson cases. A major goal of this project was to establish analytical tools that provide a statistical evaluation for forensic comparison between ignitable liquid residues in fire debris and ignitable liquids in the possession of a suspect. Establishment of the statistically based error rates is essential in order to meet current and future legal challenges.

Statement of rationale for the research: Gasoline and kerosene are the two most common ignitable liquids used as accelerants in arson cases (1). Ignitable liquids are petroleum based or related products that have certain flammable or combustible properties (2). Products classified as middle petroleum distillates (MPD) according to the American Society for Testing and Materials (ASTM) E1618 method are also commonly used ignitable liquids (3). MPD are readily available in the form of charcoal lighters, paint thinners, and solvents making them another good choice for arson crimes. Although 'arson' is a legal term, a commonly acceptable definition of the word is "a criminal act of deliberately setting fire to a property" which calls for the need of fire investigation (1).

Fire debris analysis (FDA) is the science related to the examination of fire debris samples to detect and identify ignitable liquid residues (ILR) (1). Classification of ignitable liquids according to the E1618 classification scheme can readily identify the ignitable liquid residue as a gasoline, MPD, or kerosene; however, comparing two samples (i.e. residue from fire debris to liquid in the suspect's possession) to determine if they are from the same source is much more problematic. The field of fire investigation is very subjective because of the high degree of visual interpretation needed. No single method for the statistical comparison of ignitable liquids has proven to be the best in all cases involving fire debris analysis. Although classification of

the type of liquid used to create a fire is important, distinguishing within a specific class is often necessary to the world of fire investigation (1). Therefore, however complicated it is to compare two or more samples in a case, it is important to determine if the ignitable liquid residues share a common source. In addition, determination of the precision and error rates for the comparison of IL samples is also important for the evidence to have legal standing in many Daubert states. The recent National Academy of Science report recommends that pattern recognition techniques (of which ignitable liquid residue analysis is one) have established error rates to meet Daubert rules of evidence in court(4,5).

Gas Chromatography Mass Spectrometry (GCMS) is the standard analytical technique used in fire debris analysis. A separation process must occur in order to analyze ignitable liquids. GCMS can be used to produce qualitative and quantitative results because of the composition of ignitable liquids which are composed of many different components based on their production process. When analyzing ignitable liquids produced or distilled in the same manner and classified in the same category, differentiation can be challenging. In order to find differences in similarly classified ignitable liquids, target compound analysis can be used.

Target compound analysis allows the data system to search for specified retention time windows for the mass spectra of specific compounds expected to elute within each window (2). As the target compounds are identified, the data system provides quantitative data containing the peak area of the key components found. Comparison of these key components is done by peak area ratios. Early work by Mann demonstrated the value of using GCMS and calculating peak-to-peak ratios for comparing one gasoline to another (6). Following work done by Mann, Keto (7) and later Dolan and co-workers (8, 9) applied peak ratio analysis to sequential peaks in order to establish a unique identifying profile for each evaluated gasoline sample. The subtle variations in peak area cause variations in peak area ratios resulting in a bigger variation between similarly

classified samples. Although these variations are unique, comparing neat samples to fire debris analysis is not as straightforward.

Fire debris samples are not as easy to analyze as neat samples because of their level of contamination. Contamination, usually the result of pyrolysis of organic materials (wood, carpet, padding, etc.) at the fire scene, adds unwanted peaks to the gas chromatographic pattern (1). These unwanted peaks could cause incorrect visual interpretation. The mass spectrometer's data system is used to "filter out" contaminating species in the chromatogram and produce data that is petroleum distillate related based on its target ions (8).

In this study, GCMS and target compound analysis was used to develop a method like Dolan in order to statistically differentiate between kerosene and MPD samples. Kerosene and MPD, unlike gasoline, are simple distillation products from crude oil and should be strongly related to the petroleum from which it was distilled. The relative concentrations of the key components in kerosene and MPD from a refinery will often change daily because of the variety of sources that distribute crude oil. This variation in concentration of the key compounds can provide sufficient differences for comparison between individual samples. Kerosene was evaluated from a single refinery to first establish a target compound list. Applying target compound ratio analysis to all of the kerosene samples established a method for differentiation. For MPD, a selection of commercial products (paint thinners and charcoal lighters) was used for establishing the corresponding peak ratios.

II. Methods

Sample Source and collection

Kerosene

The kerosene used for this research was collected from the Marathon Ashland Refinery and from commercial sources in Huntington, West Virginia (Figure 1). Samples from the Quality

Assurance lab at Marathon Ashland were received from Marathon refineries and distribution terminals primarily from the Midwestern states. The samples were received over several years representing both seasonal and spatial variation of kerosene. A complete list of samples with collection dates is given in Appendix 1. The kerosene samples from Marathon were received in one-liter glass bottles and were immediately transferred into 125 mL bottles with Teflon®-lined lids to prevent evaporation. Sub-samples of each of the ignitable liquids were refrigerated in 20 mL glass vials. After samples were taken from the refinery bottles, excess kerosene was mixed in a 5 gallon Scepter® gasoline storage cans to be used as a QA/QC kerosene composite sample.

Figure 1: Geographical distribution of 50 kerosene samples collected from 9 states. Refineries or distribution terminals located in some states may distribute to neighboring states.

Middle petroleum distillates

The medium petroleum distillate (MPD) samples used for this experiment were sampled from two primary sets of known liquids. First, 44 MPD samples from our lab's in-house collection of various ignitable liquids were analyzed. These samples consisted primarily of commercial paint thinners and charcoal lighters purchased from home improvement stores in the Huntington, West Virginia area over a period of several years. Identification of these liquid samples was made according to the E1618 guideline. Second, 111 MPD samples from the Ignitable Liquid Reference Collection (ILRC) developed by the ILRC Committee of the Technical Working Group for Fire and Explosives (TWGFEX) were extracted and analyzed (10). These samples were classified as MPD by the ILRC committee. A complete list of samples is given in Appendix 2.

Samples analyzed from the in-house MPD collection were run as neat samples. A 1.5 ml GC vial was filled with the respective MPD and capped. Samples from the ILRC collection were received adsorbed onto small charcoal pieces. Five to seven of these charcoal pieces were placed into a 1.5 ml GC vial and the ignitable liquid residues were desorbed off with pentane. The pentane solvent and extracted ignitable liquid residues were then transferred to a new clean 1.5 ml GC vial and capped.

GCMS Instrument parameters

All analyses were performed on an Agilent 6890N Network GC System coupled to an Agilent 5973 Network Mass Selective Detector. A Varian 60 m DB-1 column with an internal diameter of 250 μm and a 1 μm film thickness was used. The carrier gas was ultra-purity helium for all analyses.

Kerosene: For neat kerosene samples the injection was 0.1 μl with a split flow of 50:1. For diluted samples (E1412 activated charcoal strip passive adsorbent method eluted with carbon

disulfide) the injection was 1 µl with split flow of 30:1. The injector temperature was 250 °C. Oven temperature program: Initial 100°C with 1.0 minute hold time. Linear temperature ramp of 5°C/min to 275°C with a 5.0 minute final hold time.

Middle Petroleum Distillates: For the neat injections of the in-house samples, a split ratio of 50:1 was used. For the ILRC samples and E1412 samples from burn studies, a split ratio of 20:1 was used. The injector temperature was 250 °C. Oven temperature program: Initial 125°C for 1 minute followed by a temperature ramp of 5°C/min to 250°C, hold for 5 minutes.

Comparisons

Target compound analysis was used to identify the key components in both MPD and kerosene using ChemStation software as the output tool for the GCMS data. Parameters were entered into the software program to select the peaks of interest based on retention time and the presence of target ions. A spreadsheet template was developed, similar to that of Dolan and Ritacco (7) for gasoline, and calculated sequential ratios from the averages of the three GCMS injections using Excel©. A relative standard deviation (RSD) less than 5% was used as the criterion for acceptable repeatability among 3 individual injections of each sample (7). Where possible, retention times and mass spectra were confirmed using pure hydrocarbon standards. However, in some cases, no pure standards were available for particular isomers. In those cases, the compounds were identified as "A trimethyl –benzene" or where less certain "Compound A". This follows the practice of Dolan and co-workers (7,8) for gasoline.

Statistical methods

Principal components analysis (PCA) is a multivariate statistical technique which seeks to reduce the dimensionality of large data sets that result from "hyphenated methods" such as GCMS. In such data sets there is often a high degree of correlation between variables (i.e. total

ion intensities at adjacent time slices across a chromatographic peak). PCA fits a sequence of multilinear equations to the variables in the data set to explain the observed variance. Each succeeding equation is fit to the residual variance not explained by the previous one. Ideally a number of strongly correlated variables are collapsed into a few "latent variables". In our data these may be related to actual key compounds or ratios of compounds in the ignitable liquid which vary significantly between samples. Examining the "loadings" (essentially the coefficients of the latent variables) can be useful in selecting key chemical components which can differentiate the ignitable liquids in the data set. (10) The distance in the multidimensional factor space between any two samples is a measure of their chemical difference. Two ignitable liquids which came from a common source (i.e. same station on the same day) should be indistinguishable within some measure of statistical confidence. There are several methods of measuring this difference including Projected Difference Resolution (PDR)

The PDR metric is an analog to chromatographic resolution (12). PDR is applied to measure the separation of pairs of samples quantitatively in a multivariate data space. Given a GCMS data set that comprises two classes, the number of variables (i.e., number of data points, which is calculated by the number of retention time measurements times the number of mass-to-charge ratio measurements) is n, the numbers of objects (i.e., amount of GCMS spectra) in classes a and b are respectively m1 and m2. The data matrices Xa and Xb respectively have sizes of m1 × n and m2 × n, in which each row is a two-way GCMS data. The PDR measure of class separation Rs(a, b) is a scalar calculated by

$$\mathrm{R_s(a,b)} = \frac{|\bar{\mathbf{t}}_\mathbf{a} - \bar{\mathbf{t}}_\mathbf{b}|}{2(s_\mathrm{a} + s_\mathrm{b})} \qquad (1)$$

for which ta and tb are the scores for the two classes obtained by projecting the objects onto the difference vector $\overline{\mathbf{X}_\mathbf{a}} - \overline{\mathbf{X}_\mathbf{b}}$ of the class averages, given by

$$\mathbf{t}_\mathbf{a} = \mathbf{X}_\mathbf{a}(\overline{\mathbf{X}_\mathbf{a}} - \overline{\mathbf{X}_\mathbf{b}})^\mathrm{T} \qquad (2)$$

$$\mathbf{t_b} = \mathbf{X_b}(\overline{\mathbf{X_a}} - \overline{\mathbf{X_b}})^{\mathrm{T}} \quad (3)$$

for which $\overline{\mathbf{X_a}}$ and $\overline{\mathbf{X_b}}$ are the average class vectors that have a length of n. The column vectors ta and tb have lengths of m1 and m2, respectively. From the projections the averages $\overline{t}_a$ and $\overline{t}_b$ and their corresponding standard deviations s_a and s_b are calculated.

PDR is proposed as a straightforward multivariate measure for rapidly quantifying the separation of multivariate data objects for a pair of classes. The smaller the PDR, the harder to predict two classes by multivariate pattern recognition methods. Generally, a well-resolved separation of two classes has a PDR value greater than 1.5, which is comparable to the minimum resolution for baseline resolution between a pair of chromatographic peaks. When the data set contains more than two classes, the PDR metric for each pair of classes is systematically calculated for all combinations of pairs. The PDR matrix can be viewed as a triangle that measures the separation of each pair of classes.

Column bleeding was observed in the measurement of kerosene samples because of the higher column temperatures that are required for elution, which may bias the pattern classification of the two-way profiles. As a result, the baseline was corrected by a procedure based on the PCA result. For a two-way GCMS data matrix X, mass spectrometry scans are stored by rows and extracted ion chromatograms are stored by columns. First, the spectrum segment of the final one minute retention time window was selected to acquire background mass spectral scans. The PCA is performed on this background matrix of mass spectra for each sample. By performing the classification based on background subtraction using 1–10 largest principal components, it is concluded that the loading of the first principal component v1 characterize the mass spectrum of column bleeding impurities. Therefore, background correction is obtained by

$$\mathbf{X_c} = \mathbf{X} - \mathbf{X} \cdot \mathbf{v_1} \cdot \mathbf{v_1}' \quad (4)$$

for which Xc is the baseline corrected spectrum.

After baseline correction a third order polynomial was used to adjust the retention times of each chromatogram to the reference chromatogram using a two-way layout in the full retention time times the mass spectral image is used in the alignment process. The nonlinear simplex algorithm was used to maximize the correlation coefficient between the reference and each GCMS measurement. One sample (K10) was used as the reference chromatogram for retention time (RT) correction, however, several tests performed using other reference samples gave little difference in the final result. The RT correction was applied only to choose the correct target compounds in the AMDIS peak list report in the target compound ratio method. Because in the two-way profile method, peak binning reduces the effect of retention time drift, the RT correction did not improve peak identification.

Two approaches were applied to determine the number of latent variables in the partial least squares-discriminant analysis (PLS-DA) (12) method. In the optimal partial least squares-discriminant analysis (oPLS-DA), the number of latent variables is determined by achieving highest prediction accuracy for the prediction set. As a result, oPLS-DA is positively biased, which is applied as a reference method.

The other PLS-DA method is unbiased because the prediction set is not used to determine the number of latent variables in the PLS-DA model. The procedure for unbiased PLS-DA training is similar to the BLP method (13). First, the training set is split into two subsets using Latin partitions. Then, each partitioned subset is used once for prediction and once for model-building. The procedure is bootstrapped 10 times. Lastly, the prediction accuracies are averaged across the 10 bootstraps, the number of latent variables is determined by achieving highest average prediction accuracy.

FuRES is based on the decision tree algorithm and fuzzy logic theory, where each branch of the decision tree model is a multivariate fuzzy rule (14). FuRES has been successfully applied

in forensic researches to analyze two-way GCMS and gas chromatography–differential mobility spectrometry (GC–DMS) data (15,16).

OPLS-DA, PLS-DA, and FuRES were validated by the BLP method (18). The PDRs of each training set is calculated for comparison as well. Bootstrapping is a re-sampling method. Latin partition is a block cross-validation, in which the class distributions are maintained between the training set and the prediction set. BLP provides validation results with confidence intervals by running the evaluation repeatedly with different training and prediction set partitions. The MATLAB scripts for the kerosene analysis are given in Appendix 3.

For MDP data, peak area ratio data was analyzed using statistical routines (18) written in the open source statistical programming language R (R Development Core Team. "R: A Language and Environment for Statistical Computing". R Foundation for Statistical Computing (Vienna, Austria); http://www.R-project.org, 2012) (19). The R scripts utilized are given in Appendix 4.

RT drift was checked by running an E1618 standard mix of normal hydrocarbons with each set of samples. The retention times of the standard mix were compared with the normal hydrocarbons in representative samples from each set of samples Because the same column was used throughout the many months of data collection without removal/replacement, negligible changes in RT were observed. Where drift was observed, appropriate changes in the retention time windows were made. A QC sample was routinely analyzed to also check for drift.

Evaporation study

The QA/QC composite kerosene sample was evaporated to 25%, 50%, 75%, and 90% by volume under a stream of dry nitrogen gas at low heat on a hot plate.

Burn Tests

1.0 mL kerosene was dispensed on ~2x2" (5x5 cm) squares of wood, carpet, or carpet pad, ignited with a butane lighter, and allowed to self-extinguish. Kerosene residue was tested using the passive adsorption (ACS) ASTM Method E1412 with carbon disulfide as the extraction solvent.

Kimwipes® tissues spiked with 20µL of neat kerosene samples, K005 and K006, were tested using the ASTM Method E1412 for adsorption or desorption issues. Also, unburned and uncontaminated pieces of wood, carpet, and carpet pad were tested using the ASTM Method E1412 to see if overlapping data would occur in the kerosene elution region.

III. Results:

Kerosene Analysis

Forty-four kerosene samples were analyzed using target compound ratio analysis and 36 target compounds were identified, corresponding to 35 sequential ratios in each kerosene sample, as seen in Table 1.

Table 1. Target compound list, estimated retention time, and corresponding ratios identified in each kerosene sample. Specific isomers were not able to be identified in several cases due to lack of authentic standards. Qion was the ion used for peak area determination, given presence of the three qualifier ions and within the preset RT window.

					IONS (m/z)			
Peak #	Compound	Nominal RT	Ratio	Peak Ratio	Qion	Qual1	Qual2	Qual3
1)	Toluene	7.979	2/1	1	91	65	57	281
2)	p-Xylene	10.203	3/2	2	91	106	105	77
3)	Nonane	10.746	4/3	3	57	85	71	56
4)	1-ethyl-2-methyl-Benzene	12.575	5/4	4	105	120	91	77
5)	ethyl-methyl-benzene	12.683	6/5	5	57	105	71	120
6)	a-trimethyl-benzene	12.763	7/6	6	105	120	77	91
7)	ethyl-methyl-Benzene	13.18	8/7	7	105	120	91	77
8)	Decane	13.415	9/8	8	57	71	85	55
9)	1,2,4-trimethyl-Benzene	13.575	10/9	9	105	120	77	119
10)	A trimethyl-Benzene	14.495	11/10	10	105	120	77	91
11)	A diethyl-Benzene	15.255	12/11	11	57	119	105	71
12)	3-methyl-Decane	15.478	13/12	12	57	71	85	126
13)	Undecane	16.255	14/13	13	57	71	85	56
14)	A methyl-trans-Decalin	17.667	15/14	14	81	67	95	152
15)	Compound a	17.998	16/15	15	71	105	106	57
16)	2-methyl-Undecane	18.113	17/16	16	57	71	85	127
17)	Compound b	18.204	18/17	17	105	152	91	95
18)	methyl-ethyl-Benzene	18.358	19/18	18	119	134	57	85
19)	1,2,3,4-tetrahydro-Napthalene	18.816	20/19	19	104	132	91	115
20)	Dodecane	19.119	21/20	20	57	71	85	56
21)	Compound c	19.239	22/21	21	97	55	69	131
22)	2,6-dimethyl-Undecane	19.553	23/22	22	57	71	98	56
23)	Compound d	20.359	24/23	23	104	146	91	131
24)	Compound e	20.805	25/24	24	105	162	71	91
25)	A dihydro-dimethyl-1H-Indene	21.21	26/25	25	131	146	115	91
26)	1,2,3,4-tetrahydro-6-methyl-Napthalene	21.748	27/26	26	131	118	146	117
27)	Tridecane	21.896	28/27	27	57	71	85	55
28)	A tetrahydro-dimethyl-Napthalene	23.199	29/28	28	118	145	160	105
29)	Tetradecane	24.548	30/29	29	57	71	85	55
30)	1,7-dimethyl-Napthalene	26.086	31/30	30	156	141	57	71
31)	2,6,10,14-tetramethyl-hexadecane	26.2	32/31	31	57	71	85	141
32)	3-methyl-tetradecane	26.394	33/32	32	57	71	85	183
33)	Pentadecane	27.069	34/33	33	57	71	85	55
34)	Hexadecane	29.452	35/34	34	57	71	85	55
35)	Heptadecane	31.709	36/35	35	57	71	85	207
36)	Octadecane	33.847			57	71	207	85

Baseline correction

The TIC chromatograms of a kerosene sample are given in Figure 2. The effect of baseline correction is demonstrated. In kerosene samples, the baseline goes upwards in the uncorrected TIC profile. Compared to the uncorrected spectra, the baseline of the corrected spectra is improved. The pattern classification and PDR metric of the spectra before baseline correction is

performed. Comparisons were made between baseline corrected spectra and original spectra. The results are given in Table 2 below.

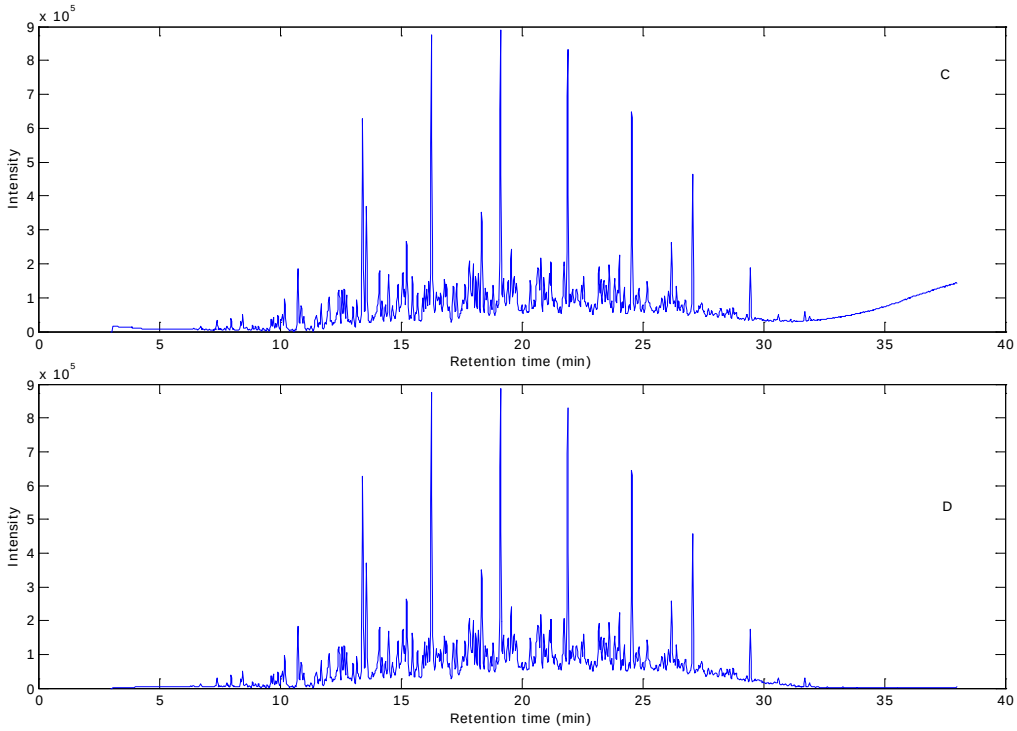

Figure 2. TIC chromatograms of a kerosene sample (C): before baseline correction, (D) after baseline correction.

PDR mapping

The PDR mapping of kerosene samples by the two-way profile method is given in Figure 3. The geometric mean of PDRs, plotted in grayscale, are measured repeatedly by removing one replicate from each class, a total of nine combinations of subsets for a pair of classes. The darkness of the box indicates the PDR value. All PDR values that are greater than or equal to 5 are plotted in white. The numbers printed in the box are the number of times out of a total of 60 times that an object was misclassified between the pair of classes during the BLP validation by FuRES. Most of the misclassifications of the classes are located in gray boxes, indicating that

the PDR metric effectively measures the predictive ability of the classifiers. It can be concluded the lower the PDR between two classes, the more likely misclassification will occur.

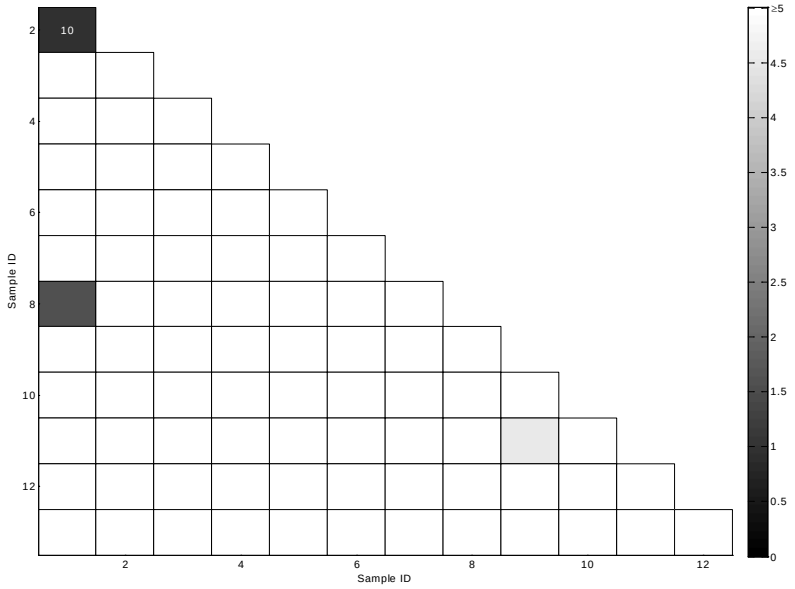

Figure 3. The PDR mapping of kerosene samples by the two-way profile method. The PDR values and the FuRES prediction use different bootstrap approaches. The PDR values are encoded by grey scale, which is the geometric mean of all possible subsets of Latin partitions. All PDR values that are greater than or equal to 5 are plotted in white. In a pair of classes that comprised of six objects, the subsets that comprised of four objects were obtained by removing one out of three objects in each class, which results nine possible combinations of subsets. The numbers in the box are the numbers of misclassifications between the corresponding pair of samples out of a total of 60 times by the BLP validation of the FuRES model.

Pattern classification

The BLP validation of PDR metric, oPLS-DA, PLS-DA, and FuRES are given in Table 2. The effect of baseline correction is evaluated. Although the prediction accuracy is not improved for two-way profiles after baseline correction, the PDRs were improved significantly in the kerosene spectra, indicating that the separation between each pair of classes was generally improved.

Table 2. PDRs and prediction accuracies of oPLS-DA, PLS-DA and FuRES with 95% confidence intervals by BLP validation. Both full two-way profile and component ratio methods are reported.

	Kerosene
Total number of objects	39
Two-way profile, original spectra	
Geometric mean PDR	17 ± 8
oPLS-DA (%)	100 ± 0
PLS-DA (%)	83 ± 6
FuRES (%)	97 ± 0
Two-way profile with baseline correction	
Geometric mean PDR	41 ± 15
oPLS-DA (%)	100 ± 0
PLS-DA (%)	92 ± 5
FuRES (%)	97 ± 0
Component ratio	
Geometric mean PDR	9 ± 2
oPLS-DA (%)	81 ± 7
PLS-DA (%)	62 ± 5
FuRES (%)	91 ± 6

Both the two-way profile and component ratio methods achieved prediction accuracies greater than 90% using the FuRES classifier. For the gasoline data set, the two-way profile method and the component ratio method performed equally well. The two-way profile method achieved higher prediction accuracies than the component ratio method for the kerosene data set because the two-way profiles retain more chemical information. The loss of peak information manifests itself in lower PDR values. The PDRs of the component ratio method is lower than the two-way profile method for both gasoline and kerosene data. It is essentially a differential transformation so there is a loss in signal-to-noise ratio, which can be expected with any differential transformation.

The two-way spectra contain noise. As a positively biased method, oPLS-DA achieved higher prediction accuracies for the data because of overfitting the data. FuRES is a soft classifier and is inherently resistant to overfitting. However, for the component ratio method overfitting is mostly avoided because the training data set is overdetermined (i.e., fewer variables

than objects). As a result, the FuRES method achieved better predictions for the component ratio data than the biased oPLS-DA method. The unbiased PLS-DA method achieved marginally better prediction accuracies for the two-way gasoline data that are statistically insignificant (15). Unbiased PLS-DA performs worse than the FuRES method for the rest of the data sets, especially in the classification of component ratio data. The performance demonstrated that FuRES is a powerful classifier for samples measured by GCMS.

Middle Petroleum Distillates

High resolution GCMS analyses for 155 MPD samples were collected in triplicate. Forty-one compounds were identified resulting in 33 peak area ratios. Not all compounds were found in all samples. Where a compound was not detected, its corresponding ratio was set to zero if it was in the denominator (normally would result in "divide by zero" error). Some ratios had %RSD exceeding the 5% nominal cut-off. The list of compounds and peak area ratios is given in Table 3.

Table 3. MDP target compounds, nominal retention times, peak ratios and monitored ions. Specific isomers were not able to be identified in several cases due to lack of authentic standards. Qion was the ion used for peak area determination, given presence of the three qualifier ions and within the preset RT window.

					Ions (m/z)			
Peak #	Compound	Nominal RT	Ratio	Peak Ratio	Qion	Qual1	Qual2	Qual3
1)	Octane	6.530	1	2/1	85	57	71	56
2)	1,3,5-trimethyl-cyclohexane.	7.251	2	3/2	111	69	55	126
3)	ethyl-cyclohexane	7.361	3	4/3	83	55	82	112
4)	2-methyl-octane	7.448	4	5/4	57	71	85	84
5)	3-methyl-octane	7.579	5	6/5	57	56	98	99
6)	1α,2β,4β-trimethyl- cyclohexane	7.645	6	7/6	111	69	55	126
7)	p-Xylene	7.711	7	17/7	91	106	105	77
8)	Nonane	7.995	8	10/8	57	85	71	56
9)	1,2,3-trimethyl- cyclohexane	8.082	9	14/9	69	111	55	126
10)	1-ethyl-4-methylcyclohexane.	8.191	10	11/10	97	55	126	69
11)	2,5-dimethyl-octane	8.497	11	12/11	57	71	85	70
12)	2,6-dimethyl-octane	8.672	12	13/12	57	71	105	56
13)	2,3-dimethyl-octane	8.913	13	14/13	57	98	55	56
14)	propyl-cyclohexane	9.000	14	17/14	83	55	82	126
15)	4-methyl-nonane	9.153	15	15/14	57	70	71	98
16)	3-methyl-nonane	9.372	16	16/15	57	105	71	56
17)	propyl-benzene	9.263	17	18/17	91	120	65	92
18)	1-ethyl-2-methyl-benzene	9.372	18	19/18	105	57	120	71
19)	1,2,3-trimethyl-benzene	9.503			105	120	119	77
20)	Decane	9.897	19	30/20	57	71	85	56
21)	1,2,4-trimethyl-benzene	10.115	20	23/21	105	97	120	55
22)	Unknown alkyl aromatic	10.421	21	22/20	71	57	70	55
23)	1,3,5-trimethyl-benzene	10.837			105	120	57	71
24)	butyl-cyclohexane	11.099	22	25/24	83	55	82	67
25)	5-methyl-decane	11.186	23	26/25	57	85	98	71
26)	2-methyl-decane	11.317	24	27/26	57	71	85	55
27)	3-methyl-decane	11.514			57	71	85	56
28)	1,2-diethyl-benzene	11.623			105	119	134	97
29)	1-methyl-2-propyl-benzene	11.776	25	30/29	105	134	0	0
30)	Undecane	12.126	26	37/30	57	71	85	56
31)	1,2,3,4-tetramethyl-benzene.	13.001	27	32/31	119	134	105	91
32)	1,2,3,5-tetramethyl- benzene	13.110			119	134	120	91
33)	4-methyl-undecane	13.613	28	34/33	71	57	70	85
34)	2-methyl-undecane	13.700	29	33/32	57	71	85	55
35)	3-methyl-undecane	13.919			57	71	85	56
36)	1,2,3,4-tetrahydro-napthalene.	14.465	30	37/36	104	132	91	57
37)	Dodecane	14.596	31	38/37	57	71	85	55
38)	2,6-dimethyl-undecane	14.968			57	71	98	56
39)	2-methyl-dodecane	16.214			57	71	85	99
40)	Tridecane	17.110	32	40/37	57	71	85	55
41)	Tetradecane	19.624	33	41/40	57	71	85	55

Chemometric Analysis of MPD target compound peak area ratios.

Three metrics of association between sets of MPD peak ratios (referred to as MPD chromatograms for brevity) were examined to test the efficacy of "matching" the chromatogram of an unknown MPD sample to known MPD. A (machine) association is the numerically based