Source Identification of Underground Fuel Spills by Solid-Phase Microextraction/High-Resolution Gas Chromatography/Genetic Algorithms

B. K. Lavine,*† J. Ritter,† A. J. Moores,† M. Wilson,† A. Faruque,‡ and H. T. Mayfield‡

Department of Chemistry, Clarkson University, Potsdam, New York 13699-5810, and AL/EQ, 139 Barnes Drive, Suite 2, Tyndall Air Force Base, Florida 32403-5323

Solid-phase microextraction (SPME), capillary column gas chromatography, and pattern recognition methods were used to develop a potential method for typing jet fuels so a spill sample in the environment can be traced to its source. The test data consisted of gas chromatograms from 180 neat jet fuel samples representing common aviation turbine fuels found in the United States (JP-4, Jet-A, JP-7, JPTS, JP-5, JP-8). SPME sampling of the fuel’s headspace afforded well-resolved reproducible profiles, which were standardized using special peak-matching software. The peak-matching procedure yielded 84 standardized retention time windows, though not all peaks were present in all gas chromatograms. A genetic algorithm (GA) was employed to identify features (in the standardized chromatograms of the neat jet fuels) suitable for pattern recognition analysis. The GA selected peaks, whose two largest principal components showed clustering of the chromatograms on the basis of fuel type. The principal component analysis routine within the fitness function of the GA acted as an information filter, significantly reducing the size of the search space, since it restricted the search to feature subsets whose variance is primarily about differences between the various fuel types in the training set. In addition, the GA focused on those classes and/or samples that were difficult to classify as it trained using a form of boosting. Samples that consistently classify correctly were not as heavily weighted as samples that were difficult to classify. Over time, the GA learned its optimal parameters in a manner similar to a perceptron. The pattern recognition GA integrated aspects of strong and weak learning to yield a “smart” one-pass procedure for feature selection.

Groundwater is the last remaining source of potable water for many households and communities in the southeastern United States.1 Its possible contamination by fuels stored in leaking underground tanks and pipelines has become a serious environmental problem, prompting both federal and state regulatory agencies to fund the development of new methods for the identification of fuel materials recovered from subsurface environments. Burgeoning interest in techniques that can determine the type of fuel responsible for an underground spill is motivated in part by the cleanup costs, legal fees, and fines incurred by the polluter.

The potential of gas chromatography and pattern recognition (GC/PR) methods for tracing underground fuel spills to suspected sources is recognized by many workers in the field of hazardous waste management. GC/PR is rapid, sample preparation is minimal, and the instrumentation and software required for the analysis is inexpensive and readily available. The chromatographic reproducibility and the discriminatory power necessary for tracing an underground fuel spill to its source can be achieved using conventional gas chromatography.2–7 Although spectroscopic techniques such as infrared and fluorescence spectroscopy have been used to characterize fuel spills,8–12 they do not possess sufficient identification power because the chemical composition

---

1 Clarkison University.

2 Tyndall Air Force Base.
of a jet (middle distillate) fuel is primarily \( \text{C}_9 - \text{C}_{13} \) alkanes and alkenes. (Aromatics are only minor constituents.) In addition, the contamination of a spill sample by salt, sand, water, or organic material can pose a serious problem when optical techniques are used since there is no prior separation step. Nevertheless, these techniques are attractive because they can be implemented in the field using fiber-optic probes in the ultraviolet, visible, or infrared range, thereby eliminating errors associated with transporting samples.

GC/PR can also be performed in the field by taking advantage of a sampling technique called solid-phase microextraction (SPME). SPME is a microscale extraction technique designed for use with gas chromatography. SPME utilizes a fused-silica fiber coated with a polymer (e.g., poly(dimethylsiloxane)) to extract organic compounds from aqueous samples. The fiber is attached to a modified syringe, and the constituents of the fuel are extracted by introducing the syringe into the headspace of the sample and depressing the plunger for a specified time period. (When the plunger is depressed, the polymer-coated fiber is in contact with the sample.) Following extraction, the plunger is retracted, and the fiber is then inserted directly into the injection port of the gas chromatograph. The plunger is depressed, exposing the polymeric coating to the high temperature of the injector port of the GC ensuring thermal desorption of the adsorbed molecules. The adsorbed molecules are immediately released into the carrier gas stream, where they are swept onto the GC column for separation and subsequent identification.

Using an SPME sampling device, it is possible to directly sample a layer of floating fuel in a water sample brought to the surface by a drilling rig. Hence, fuel spill identification can be performed on-site via a portable gas chromatograph equipped with suitable pattern recognition software. On-site chemical analysis of fuel spills is attractive for a variety of reasons. The problems associated with collecting and storing samples to the laboratory are eliminated, and the cost of characterizing a fuel sample can be substantially reduced by on-site chemical analysis.

In this study, headspace SPME/GC has been investigated as a possible alternative to conventional gas chromatography for fuel spill identification. Headspace SPME/GC has been used to analyze groundwater for benzene and toluene and has been used for analyses involving food and beverage components. The headspace technique is advantageous when either the volatility of the analyte permits a headspace determination or undesirable components in the bulk sample make direct SPME sampling undesirable. Headspace sampling prolongs fiber life and is faster than direct sampling. Finally, recovered fuels cannot be analyzed by direct SPME because of irreversible fiber damage caused by excessive swelling of the poly(dimethylsiloxane) coating, which is a direct result of the fiber being in intimate contact with the fuel layer.

To test the suitability of SPME for the proposed application, a database consisting of gas chromatograms from 180 neat jet fuel samples representing common aviation fuels found in the United States (JP-4, Jet-A, JP-7, JP-8, JP-9) was assembled. SPME sampling of the fuel’s headspace afforded well-resolved reproducible profiles; microcomputer-controlled instrumentation enabled transmission of the gas chromatograms to a host facility for processing via specially developed software. Since the statistical procedures used in pattern recognition require the same feature (GC peak) to have the same identifier (retention time) in all gas chromatograms, a computer program for peak matching the gas chromatograms to a cumulative reference standard was used. After peak matching, a total of 84 peaks were present in the reference, although not all peaks were present in all samples.

Employing pattern recognition techniques, the gas chromatograms of jet fuels that had undergone weathering in a subsurface environment were correctly identified by type using discriminants developed from the gas chromatograms of the neat jet fuels. To identify peaks in the gas chromatograms of the neat jet fuels suitable for pattern recognition analysis, it was necessary to employ a genetic algorithm. A genetic algorithm (GA) is a search method that simulates natural evolution by employing a population of potential solutions and various nature-inspired operations on these solutions or points to evolve toward better solutions or better regions of the search space. The pattern recognition GA employed in this study chose peaks that optimized the separation of the fuel classes in a plot of the two largest principal components of the data. Because the two largest principal components often capture the bulk of the variance in the data, the peaks selected by the GA contain information primarily about differences between the classes in the data set. (Since variance is synonymous with information, feature selection via this criterion can cause information about class differences to emerge as the dominant source of variation in the data.)

The principal component analysis routine used in the fitness function of the GA served as an information filter, significantly reducing the size of the search space, since it restricted the search to feature sets whose principal component plots show clustering on the basis of fuel type. In addition, the GA focused on those classes and/or samples that were difficult to classify as it trained using a form of boosting. Samples that consistently classify correctly were not as heavily weighted as samples that were difficult to classify. Over time, the GA learned its optimal parameters in a manner similar to a perceptron. The pattern recognition GA integrated aspects of strong and weak learning to yield a “smart” one-pass procedure for feature selection.

**Table 1. Neat Jet Fuels**

<table>
<thead>
<tr>
<th>no. of samples</th>
<th>fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>JP-4 (fuel used by USAF fighters)</td>
</tr>
<tr>
<td>48</td>
<td>Jet-A (fuel used by civilian airliners)</td>
</tr>
<tr>
<td>12</td>
<td>JP-7 (fuel used by SR-71 reconnaissance plane)</td>
</tr>
<tr>
<td>15</td>
<td>JP-5 (fuel used by TR-1 and U-2 aircraft)</td>
</tr>
<tr>
<td>19</td>
<td>JP-5 (fuel used by Navy jets)</td>
</tr>
<tr>
<td>42</td>
<td>JP-8 (fuel used by USAF fighters in NATO)</td>
</tr>
</tbody>
</table>

*JP-8 will replace JP-4 as the principal United States Air Force fuel.*

**EXPERIMENTAL SECTION**

A total of 180 jet fuel samples representing JP-4, Jet-A, JP-7, JP-5, and JP-8 were analyzed by headspace SPME/GC (see Table 1). The fuel samples were obtained from Wright Patterson Air Force Base or Mukilteo Energy Management Laboratory. They were splits from quality control standards used by these laboratories to verify the authenticity of manufacturer's claims. These control standards constituted a representative sampling of the fuels. A total of 13 fuel samples (7 JP-4, 4 JP-5, and 2 diesel) recovered from the subsurface environment near Air Force and Navy airfields were also analyzed by headspace SPME/GC.

The following experimental protocol was used to obtain GC profiles of a jet fuel's volatiles components. Four milliliters of a neat jet fuel were placed in a 40-mL VOA bottle. A microstirring bar was also placed in the VOA bottle prior to the introduction of the fuel sample to permit stirring of the sample during the SPME sampling period, which was 10 min. In a previous study, we found that 10 min was sufficient to obtain a representative sampling of the volatile compounds present in a jet fuel. SPME procedures were performed manually using a commercial SPME holder (Supelco). A single extraction fiber (100-μm poly(dimethylsiloxane), Supelco) was used to sample the headspace of the fuels in the training set.

Gas chromatograms of the fuels were obtained using a high-efficiency fused-silica capillary column (J&W Scientific), 10 m in length with an internal diameter of 0.10 mm. The stationary phase was a 0.34μm bonded and cross-linked 5% phenyl-substituted polyethylene glycol phase. The GC column was temperature programmed using a HP-5890 gas chromatograph equipped with a flame ionization detector. The column oven was initially set at 40 °C for 3 min, and the column was then temperature programmed at 10°/min to a final temperature of 250 °C and a final hold time of 0 min. Gas chromatograms representative of the six fuel types in this study are shown in Figure 1.

The GC was operated in the split mode (30:1 split ratio) because of the high solubility of the fuel components in the polymer coating of the SPME. Helium was used as the carrier gas with the column head pressure maintained at 30 lb/in.². Detector gas flow rates for hydrogen and air were 25 and 250 mL/min, respectively.

**PATTERN RECOGNITION ANALYSIS**

Pattern recognition analyses of the data were based on four underlying principles. First, every fuel class has a unique chemical composition. Second, each class has a characteristic GC profile distinguishable from other sources of hydrocarbons. Third, information from the gas chromatograms can be encoded in such a way as to be resistant to chemical noise and/or analytical error. Fourth, discriminants can be developed that are relatively insensitive to changes in the overall GC profile of the original fuels due to weathering, contamination, or analytical error. Pattern recognition analysis of the GC profile data was performed in five distinct steps (see Figure 2): peak matching, outlier analysis, feature selection, classification, and prediction, which are summarized below.

**Peak Matching.** The GC data were stored on a HP-3357 laboratory automation system implemented on a HP-1000-F minicomputer. A FORTRAN program translated the integration reports into ASCII files formatted for entry into SETUP, a computer program developed for peak-matching gas chromatograms. SETUP analyzed the GC data in four distinct steps. First, Kovat's retention indexes were computed for each peak in every gas chromatogram, which was a simple matter because n-alkane peaks are the most prominent features present in the gas chromatograms of these fuels. Second, a template of unique peaks for the GC data was constructed, with the peaks arranged according to their Kovat's retention index. Third, a preliminary data vector was generated for each GC profile by matching it against the template. If the peak was present, its area from the integration report was assigned a value of zero. For peak matching, the user was required to specify a tolerance window for acceptable retention time differences. Thus, peaks would be matched provided that differences in adjusted retention times for a given peak pair fell within the specified tolerance window. SETUP also computed the frequency of each feature. In other words, the number of times a particular peak was found to have a nonzero occurrence was computed. Features below a user-specified number of nonzero occurrences, which was set equal to 10% of the total number of samples in the training set, were deleted from the data set. The aforementioned peak-matching procedure yielded a full cumulative reference file containing 84 features, though not all peaks were present in all gas chromatograms. For pattern recognition analysis, each gas chromatogram was initially represented by an 84-dimensional data vector, $x = (x_0, x_1, x_2, ..., x_{83}, x_{84})$, where $x_i$ is the area of the jth peak. Each data vector was normalized to constant sum using the total integrated peak area.

**Outlier Analysis.** Outliers can adversely influence the performance of pattern recognition techniques. Hence, outlier analysis was performed on each fuel class in the training set (see Table 1) prior to pattern recognition analysis using the generalized distance test, which was implemented via SCOUT. A total of 24 chromatograms were identified as outliers and were removed from the training set. The outliers included three JP-4 chromatograms, seven Jet-A chromatograms, three JP-7 chromatograms, and others.

---

Figure 1. Headspace SPME/GC profiles of neat jet fuels.

Figure 2. Five-step procedure for pattern recognition

426 Analytical Chemistry, Vol. 72, No. 2, January 15, 2000
three JPTS chromatograms, four JP-5 chromatograms, and four JP-8 chromatograms. The large number of outliers in the data set was surprising and is probably due to some difficulties associated with using the SPME sampling device. (In a previous study involving the same neat jet fuel samples, there were only a few outliers, ~3% of the samples in the training set.) The fuels in the previous study had been diluted with methylene chloride prior to their injection onto the same GC column.

**Feature Selection.** Feature selection is a crucial step in all pattern recognition studies. By identifying the appropriate features, a hard problem can be reduced to a simple one. The feature selection method employed should be multivariate to ensure that crucial features are not discarded. In this study, a GA was employed to select peaks for pattern recognition analysis (see Figure 3). First, the GA built a population of binary string (or chromosomes) of length 84. Each string represented a unique subset of the 84 GC peaks. A particular GC peak was present in a string (or feature subset) only if the corresponding bit in the string was set to 1. If the bit was set to 0, the peak was not included in the feature subset. The number of chromosomes (or potential solutions) in the initial population was set to 100. The chromosomes comprising the initial population of strings (i.e., solutions) were generated at random to minimize potential bias.

With each generation, the pattern recognition GA computed the class and sample weights. These weights are an integral part of the fitness function, since classes and/or fuel classes that consistently classify correctly are not as heavily weighted as samples that are difficult to classify. Class and sample weights were adjusted during each generation to preserve the following property: the sum of the class weights equals 100, and the sum of the sample weights in a class equals the class weight. This facilitates the tracking and scoring of the chromosomes (sets of GC peaks) between generations. The normalization functions are given in eqs 1 and 2. CW(c) is the weight of class c and SW(c)

$$CW(c) = \frac{100 \cdot CW(c)}{\sum_c CW(c)} \quad (1)$$

$$SW_c(s) = CW(c) \frac{SW_c(s)}{\sum_s SW_c(s)} \quad (2)$$

is the weight of sample s in class c. Prior to the first generation, the user initializes the class weights, with the sample weights being uniformly distributed in a class.

Each binary string (or set of GC peaks) was evaluated using a fitness function, which scored the GC peaks according to their ability to optimize the separation of the different sample types in a plot of the two largest principal components of the feature subset. (In other words, the clustering of the samples by fuel type was evaluated in a plot of the two largest principal components of each subset of the 84 GC peaks selected by the GA.) Scoring was performed on each principal component plot, which was generated for each chromosome, after the subset of features coded in the chromosome has been extracted. The k-nearest neighbor (K-NN) classification algorithm was used to score the plots. For a given sample point in the principal component plot, Euclidean distances are computed between it and every other point in the plot. These distances are arranged from smallest to largest, and a poll is taken of the point’s k-nearest neighbors, where k equals the number of samples with the same class label as the data point. For the most rigorous classification, k equals the number of samples in the class to which the sample belongs. The number of like nearest neighbors for a sample is called the sample hit count (SHC) and is multiplied by the sample weight for each point in the principal component map to yield the fitness function; see eq 3 (0 ≤ SHC ≤ K).

$$\sum \frac{1}{s} \times SHC(s) \times SW(s) \quad (3)$$

The sample hit rate (SHR), which is used to drive the boosting routines of the GA; see eq 4, where φ is the size of the population. (In this study, φ = 100.)

$$SHR(s) = \frac{1}{\phi} \sum_{c=1}^{K} SHC_c(s) \quad (4)$$

To understand scoring by the fitness function, consider a data set with two classes, which have been assigned equal weights by the GA. Class 1 has 25 samples, and class 2 has 50 samples. For uniformly distributed sample weights, class 1 samples have a weight of 2 and class 2 samples have a weight of 1. If a sample in class 1 has nine class 1 samples as its nearest neighbors in a principal component plot developed from a particular feature subset, then $SHC(c)/K_c = 9/10$, and the contribution of this sample to the fitness function is 0.9 × 2 or 1.8. Multiplying SHC/K by SW(s) for each sample and summing up the corresponding product for the 75 points in the principal component plot yields the value of the fitness function for this particular set of features.

During each generation, a reproduction operator was applied to the chromosomes. In this study, selection and crossover were implemented by ordering the population of strings, i.e., potential solutions, from best to worst, while simultaneously generating a copy of the same population and randomizing the order of the strings in this copy with respect to their fitness. A fraction of the population was then selected as per the selection pressure, which

---


was set at 0.5. The top half of the ordered (first) population was mated with strings from the top half of the random (second) population, guaranteeing the best 50% are selected for reproduction, while every string in the randomized copy has a uniform chance of being selected. This is due to the randomized selection criterion imposed on strings from the second population. If we used a purely biased selection criterion to select strings, only a small region of the search space would be explored. Within a few generations, the population would consist of only copies of the best strings in the initial population.

For each pair of strings selected for mating, two new strings were generated using a variation of three-point crossover (see Figure 4). As in the case of simple three-point crossover, the length of each new string is the same as that of its parents. However, our crossover operator is not compelled to preserve order among the exchanged string fragments, which safeguards the loss of information or features in the population. It becomes less likely for the population variability to fall below a critical value due to the additional degree of freedom provided by the reordering.

In the last step of reproduction, a mutation operator was applied to the new strings. The mutation probability of the operator was set at 0.01%. In other words, 1% of the feature subsets were selected at random for mutation. A chromosome marked for mutation has a single bit flipped, which allows the GA to explore adjacent regions of the parameter space. If the GA finds a better point, the bits from this point can invade the population, with optimization continuing in a new direction.

The resulting population of strings, both parents and children, was sorted by fitness, and the top φ strings were retained for the next generation. Since the selection criterion used for reproduction exhibits bias for the higher ranking strings, the new population is expected to perform better on average than its predecessor. However, the reproduction operators also ensure a significant degree of diversity in the new population, since every string in the randomized copy of the population has an equal chance of being selected and the reordering of exchanged string fragments of each chromosome pair and crossover points is selected at random.

The pattern recognition GA was able to concentrate its efforts on classes and/or samples, which were more difficult to classify, because it boosted their weights. There were two stages in boosting. In the first stage, class weights were adjusted relative to each other to achieve an optimum configuration. The class hit rate or CHR (see eq 5), which is the average sample hit rate for samples in a particular class, was computed. Those classes with lower class hit rates were weighted more heavily than classes, which scored well.

The change in the class weights, which is computed using a perceptron (see eq 6), was monitored throughout the run. P, which was assigned a value by the user, is the momentum parameter for the perceptron.

If the average change in the class weights is greater than some tolerance value, the GA is said to be learning its class weights. Once the change is less than some tolerance, the class weights become fixed. (In other words, the pattern recognition GA turns off the perceptron algorithm for the class weights.) The sample weights in each class can then be reset, so they are uniformly distributed in the class. This initiates the second stage. P or the momentum, which controls the rate at which the sample and class weights are changed (see eqs 6 and 7), is initially assigned a value of 0.8 while the GA is learning, but P is adjusted to 0.4, once the class weights are fixed. These P values have been chosen in part because they facilitate learning but do not cause a particular sample or class to dominate the calculation, which would result in the other samples or classes not contributing to the fitness function of the GA.

During each generation, class and sample weights were updated using the class and sample hit rates from the previous generation. (g + 1 is the current generation, whereas g is the previous generation.) A block diagram of boosting is shown in Figure 5. Evaluation, reproduction, and boosting of the potential solutions were carried out until a specified number of generations had been executed or a feasible solution (a set of GC peaks, which can be used for discriminant analysis) was found.

Classification. SIMCA pattern recognition was used to classify the gas chromatograms of the neat jet fuels. The method involves fitting a principal component model to each fuel class in the training set (see Table 2). Confidence envelopes around each class model are constructed for the purpose of containing the data points. The parameters of the confidence envelope are derived from the principal component analysis. Chromatograms to be classified are fitted to the class models and assigned a probability for each class. If the probability of belonging to a class is high, then an assignment is made. If all the probabilities are low, the sample is either an outlier or a member of a new class.

Prediction. To test the predictive ability of the GC peaks selected by the GA and the discriminants associated with them, it was necessary to use a prediction set (see Table 3). For this
study, the prediction set consisted of 13 weathered jet fuel samples and 20 neat jet fuel gas chromatograms. The 20 neat jet fuel gas chromatograms were obtained under conditions different from the chromatograms in the training set. For the neat jet fuel samples, we employed a different poly(dimethylsiloxane) SPME fiber or varied the sampling time of the fiber over the headspace of the fuel, to assess the robustness of the SPME sampling technique. The 13 weathered jet fuel samples were obtained from sampling wells as a neat oily phase found floating on top of the water. The sampling wells were located at Tyndall or Keywest Naval Air Station in Florida. The approach to fuel spill identification used here, classifying the gas chromatograms of weathered jet fuels with discriminants developed from the gas chromatograms of neat jet fuels, was taken because the physical and chemical interactions of jet fuel components with a subsurface environment are not yet fully understood.

RESULTS AND DISCUSSION
The first step in this study was to apply principal component analysis\(^\text{33}\) to the data. Principal component analysis attempts to reduce the dimensionality of the data while simultaneously preserving the information present in the data. Dimensionality reduction is possible with principal component analysis because chromatographic data sets are often redundant. That is, the measurement variables are correlated. High collinearity between

\[ \text{Analytical Chemistry, Vol. 72, No. 2, January 15, 2000 429} \]
measurement variables is a strong indication that a new set of basis vectors can be found that are better at conveying the information content present in the data than axes defined by the original measurement variables. This new basis set, which is linked to variation in the data, can be used to develop a new coordinate system for displaying the data. The principal components define the variance-based axes of this new coordinate system. Each principal component is a linear combination of the original measurement variables. Often, only two or three principal components are necessary to explain all of the information present in a data set.

Figure 6 shows a principal component map of the 84 GC peaks obtained from the 156 neat jet fuel gas chromatograms. The two largest principal components explain 65% of the total cumulative variance. Only JP-4 is linearly separable from the other fuels in the map. The gas chromatograms of the JP-4 fuel samples are linearly separable from the gas chromatograms of the other jet fuels in the principal component map. However, the gas chromatograms of the Jet-A, JP-7, JPTS, JP-5, and JP-8 gas chromatograms overlap, which is not surprising because these fuels are kerosene based, similar to Jet-A, the fuel used by civilian airliners. Mayfield and Henley (34) reported that gas chromatograms of kerosene-based fuels are more difficult to classify than gas chromatograms of other types of jet fuels, due to the similarity in their overall hydrocarbon profile.

Clearly, feature selection will be crucial for increasing the signal-to-noise ratio of the data because measurements on chemical components that are not characteristic of the source profile of the fuels represented in the data set are discarded. In this study, we used a genetic algorithm to uncover features characteristic of the source profile of each fuel type. The genetic algorithm identified features by sampling key feature subsets and scoring their principal component plots. As it trained, the GA focused on those classes and/ or samples that were difficult to classify. The boosting routine used this information to steer the population quickly to an optimal solution. After 100 generations, the genetic algorithm identified 22 peaks. The principal component map of the 22 chromatographic peaks shows clustering of the jet fuel samples according to fuel type (see Figure 7). Since the data points are well separated by fuel type in the principal component map, it follows that a similar separation exists in the 22-dimensional measurement space. This conclusion is reinforced by the results of k-nearest neighbor (35) which was also used to analyze these 22 chromatographic peaks (see Table 4). The k-nearest neighbor classification method categorized the fuel samples in the training set according to their proximity to other samples (of preassigned categories) in the 22-dimensional measurement space. It is evident on the basis of k-nearest neighbor and the principal component map that information is contained in these 22 peaks characteristic of fuel type.

The 22 GC peaks identified by the genetic algorithm served as a starting point for the SIMCA study. The training set data

---

**Figure 6.** Principal component map of the 84 GC peaks obtained from the 156 neat jet fuel gas chromatograms. The two largest principal components explain 65% of the total cumulative variance. 1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, and 6 = JP-8. Only JP-4 is linearly separable from the other fuels in the map.

**Figure 7.** Principal component map of the 22 GC peaks obtained from the 156 neat jet fuel gas chromatograms. The two largest principal components explain 47% of the total cumulative variance. 1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5, and 6 = JP-8. All the fuels are linearly separable in the map.

**Table 4. K-NN Classification Results for the Training Set**

<table>
<thead>
<tr>
<th>class</th>
<th>no.</th>
<th>1-NN</th>
<th>3-NN</th>
<th>5-NN</th>
<th>7-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>JP-4</td>
<td>41</td>
<td>41</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Jet-A</td>
<td>41</td>
<td>41</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>JP-7</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>JPTS</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>JP-5</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>JP-8</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>total</td>
<td>156</td>
<td>154</td>
<td>151</td>
<td>150</td>
<td>148</td>
</tr>
</tbody>
</table>

---


were divided into six classes according to fuel type: JP-4, Jet-A, JP-7, JPTS, JP-5, and JP-8 fuels. Principal component analysis was performed on each fuel class in the training set to separate the information and noise sources. Truncation of the noninformative sources of variation led to a principal component model for each fuel class. The model size or complexity for each fuel class was determined using a procedure called cross validation.36 For each fuel class, one principal component was used to model the data. The gas chromatograms in the training set were then fitted to the principal component models, and the residual (the sum of the square differences between an original gas chromatogram and one reproduced by a principal component model) was computed for every gas chromatogram and model. Each gas chromatogram in the training set was classified on the basis of its goodness of fit via an F-test, which was performed using the residual variance of the sample together with the residual variance of the class.32 The F-test determined whether a sample was inside or outside its assigned class and outside the other classes in the data. A gas chromatogram was assigned to the fuel class for which it had the lowest variance ratio. If the variance ratio exceeded the critical F-value for the class, then it was not assigned to it. Results from the SIMCA six-way classification study involving the training set samples are summarized in Table 5. The recognition rate for each principal component model was reasonably high.

To test the predictive ability of these GC peaks and the principal component models associated with them, a prediction set of 33 gas chromatograms was employed (see Table 3). Every neat jet fuel sample in the prediction set was correctly classified. This result suggests that our SPME sampling technique is robust since variations in sampling time or fiber type did not adversely influence the classification of the neat jet fuels.

The recovered fuels posed a more challenging identification problem. Of the 13 recovered jet fuels, 11 were correctly classified. Nine of the 11 were uniquely classified,32 which would suggest a strong classification of the data. (The two diesel fuel samples were not assigned to any fuel class in the training set, which we construed as a correct classification, since diesel fuels were not included in the training set.) However, two of the four JP-5 fuels were classified as Jet-A, which is not surprising since Jet-A and JP-5 are similar in their chemical composition. In a previous study,31 we were able to correctly classify all of the weathered fuels (which included the same four JP-5 fuel samples) using chromatographic data obtained by directly injecting fuels diluted with methylene chloride into the gas chromatograph. Furthermore, the recognition rates obtained for the samples in the training set were higher31,37 (Using SIMCA, a recognition rate of 100% was achieved for the training set samples in the previous study versus 92.3% in the current study.) Evidently, there is a loss in classification power when the headspace of the jet fuels is sampled using the SPME technique since only the most volatile components are sampled by the fiber.

**ACKNOWLEDGMENT**

B.K.L. acknowledges the financial support of the Air Force Office of Scientific Research, (Bollings Air Force Base, Washington, DC) through their Summer Faculty Research Program. J.R. also acknowledges the financial support of the Air Force Office of Scientific Research through their Summer Graduate Research Program. Any references to specific trade names or products do not constitute an endorsement on behalf of the United States Air Force.

Received for review May 11, 1999. Accepted October 26, 1999.

---

**Table 5. SIMCA Classification Results**

<table>
<thead>
<tr>
<th>class</th>
<th>no. in class</th>
<th>right</th>
<th>wrong</th>
<th>% right</th>
</tr>
</thead>
<tbody>
<tr>
<td>JP-4</td>
<td>41</td>
<td>41</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Jet-A</td>
<td>41</td>
<td>38</td>
<td>3</td>
<td>93</td>
</tr>
<tr>
<td>JP-7</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>JPTS</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>83.33</td>
</tr>
<tr>
<td>JP-5</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>JP-8</td>
<td>38</td>
<td>31</td>
<td>7</td>
<td>81.6</td>
</tr>
<tr>
<td>total</td>
<td>156</td>
<td>144</td>
<td>12</td>
<td>92.3</td>
</tr>
</tbody>
</table>

---