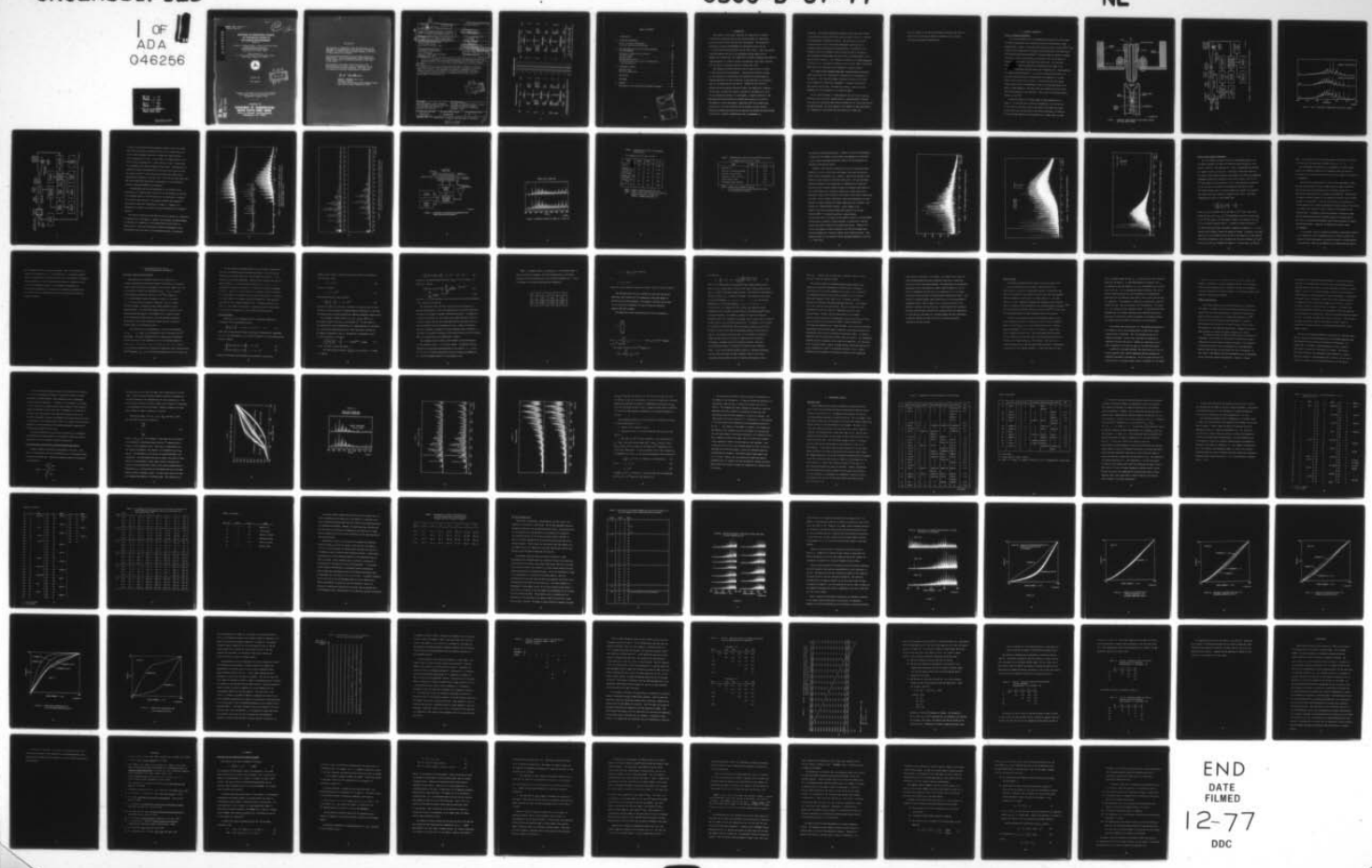
AD-A046 256 STANFORD RESEARCH INST MENLO PARK CALIF F/G 13/2
METHODS OF IDENTIFYING SOURCE OF PETROLEUM FOUND IN THE MARINE --ETC(U)
NOV 76 M E SCOLNICK, A C SCOTT, M ANBAR DOT-CG-81-74-1187
UNCLASSIFIED USCG-D-37-77 NL

1 OF
ADA
046256



END
DATE
FILMED
12-77
DDC

AD A046256

Report No. CG-D-37-77
Task No. 4243.2.8

12

METHODS OF IDENTIFYING SOURCE OF PETROLEUM FOUND IN THE MARINE ENVIRONMENT

REPORT II

Martin E. Scolnick, Arthur C. Scott, and Michael Anbar
Stanford Research Institute
~~Mass Spectrometry Research Center~~
Menlo Park, California 94025

Under contract to:
U.S. Coast Guard Research and Development Center
Avery Point, Groton, CT 06340

DDC
RECEIVED
NOV 2 1977
F

November 1976

FINAL REPORT

Document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161

AD No.
DDC FILE COPY

Prepared for

DEPARTMENT OF TRANSPORTATION
UNITED STATES COAST GUARD
Office of Research and Development
Washington, D.C. 20590

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of this report.

The contents of this report reflect the views of the Stanford Research Institute, which is responsible for the facts and accuracy of data presented. This report does not constitute a standard, specification or regulation.

D. L. Birkimer

DONALD L. BIRKIMER, Ph.D., P.E.
Technical Director
U.S. Coast Guard Research and Development Center
Avery Point, Groton, Connecticut 06340

(19) (15 USCG, CGR/DC)

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CG-D-37-77,	11/77	

4. Title and Subtitle (6)

METHODS OF IDENTIFYING SOURCE OF PETROLEUM FOUND IN THE MARINE ENVIRONMENT, REPORT II.

5. Report Date (11) November 1976

6. Performing Organization Code

7. Author(s) (10)

M. E. Scolnick, A. C. Scott, and M. Anbar

8. Performing Organization Report No. (12) 94p.

9. Performing Organization Name and Address

Stanford Research Institute
333 Ravenswood Avenue
Menlo Park, California 94025

10. Work Unit No. (TRAIS)
4243.2.8

11. Contract or Grant No. (15)
DOT-CG-81-74-1187, new

13. Type of Report and Period Covered (9)
Final Report, DOT-CG-22972-A
June 26, 1972 to May 17, 1976
26 Jun 72-17 May 76

12. Sponsoring Agency Name and Address

Department of Transportation
U. S. Coast Guard
Office of Research and Development
Washington, DC 20590

14. Sponsoring Agency Code

15. Supplementary Notes

The contract under which this report was submitted was under the technical supervision of the Coast Guard Research and Development Center, Groton, Connecticut 06340. R&D Center Report Number CGR&DC 11/77 has been assigned.

16. Abstract

The identification of oils by field ionization mass spectrometry is reported. Two multivariate data analysis models are described; a parametric statistical model that is based on the assumption of stochastic independence and an empirical model that can be used in a "learning machine" mode. The results of applying the empirical model to 154 quadrupole spectra and 42 sector magnet spectra are reported.

DDC RECEIVED NOV 2 1977 F

17. Key Words	18. Distribution Statement
Mass spectrometry, field ionization, multiscanning techniques (in mass spec), oil identification, statistical analysis, data analysis	Document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED	94	

Form DOT F 1700.7 (8-72) Reproduction of form and completed page is authorized

332500.

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.

Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

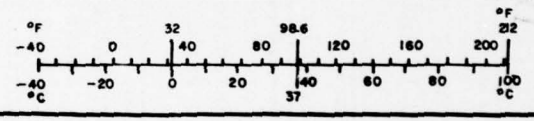


TABLE OF CONTENTS

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
SPECIAL
A

I INTRODUCTION

The contents of this report represent the culmination of a program to develop the technology and the data analysis methods for identifying oils by the use of field ionization mass spectrometry. This program was supported by contracts DOT-CG-22996-A and DOT-CG-81-74-1187 with the U. S. Coast Guard during the period June 26, 1972 to May 17, 1976. Our efforts involved primarily the use of two spectrometer systems which were assembled, for the most part, with commercially available components and capable of producing spectra of crude and refined oils spanning a mass range of 400 amu with the ability to reproduce spectra to better than ± 10%.

The problems of analyzing spectral data are many and are not limited to field ionization mass spectrometry. Among the more difficult problems for this program were the detection and deconvolution of fused peaks and problems associated with the variability in peak resolution of our spectrometers at ion masses greater than 300 amu. Perhaps the most difficult problem was the unavoidable existence of small, but significant, systematic errors which, for practical reasons, prevented us from making full use of all the information inherent in a spectrogram. A complete solution to the fused peak and variable resolution problems came recently as a result of the design of a third spectrometer, comprising a 60^{o} sector magnet mass analyzer and a new field ionization source (funded by another project). Resolution problems associated with the previous two systems have been avoided by the use of a spectral representation that is independent of

resolution. The problems presented by systematic errors have been avoided by abandoning our previous statistical analysis model in favor of a modified "learning machine" discriminant function. The application of this discriminant function to the "resolution-independent" renditions of 154 quadrupole spectra comprising the representations of 35 different oils produced correct and unambiguous identifications for 95% of the spectra. The discriminant function model can also be used for ranking spectra in the order of their similarity. This technique was applied to the older quadrupole data and to a set of 42 spectra produced by the new sector magnet spectrometer. These results are voluminous and are presented in separate binders.

The "Review of Data Analysis Techniques" describes several statistical models that proved to be unsuccessful or impractical for the analysis of mass spectra. The "Statistical Model" described in Section III is based on the assumption of stochastic independence, a condition that has not been satisfied by our data. This model did, however, serve as a useful background for the development of the "Empirical Model".

As this project matured, it became apparent that the cost of acquiring sufficient data to serve as training sets in a "learning machine" approach or for use in a statistical model would be prohibitive for routine applications. The empirical model, our final approach to the analysis of mass spectrometric oil "fingerprints", may provide the Coast Guard with a simple and

objective measure of the similarity between two spectra that could be used within the constraints imposed by systematic error, variable resolution and economic considerations.

II HISTORICAL PERSPECTIVE

Review of Hardware Development

The first spectrometer that was designed and constructed for this project included the multipoint field ionization source with wire mesh grid, shown schematically in Figure 1, and an ExB field velocity filter, the Colutron®.[1,2] The system is shown schematically in Figure 2 and was described at the Sixth International Mass Spectrometry Conference.[3] The mass range was scanned by varying the electric field of the velocity filter linearly with time by means of a ramp generator and high voltage operational amplifiers. The individual spectra that resulted from repetitively scanning the mass range as the organic specimen evaporated into the ionizer were integrated into a single spectrum in a 4096 channel multiscaler.[4] In an effort to optimize the stability of the data acquisition system, the multiscaler was triggered by a voltage marker in the velocity filter circuit. In this manner the first channel always received counts corresponding to ions of the same mass/charge ratio. Figure 3 shows examples of the spectra that were produced with this system. Note the nonlinearity of the mass scale. This is due to the proportionality between m/e and $1/E^2$.

A modified version of the above system is shown schematically in Figure 4. In this case the ion beam was focused onto a slit at the exit of the extractor lens which acted as the object whose image was focused by the einzel lens onto the plane of the electron multiplier ion detector. The ramp voltage supplied by the multiscaler as a "scope sweep" was used

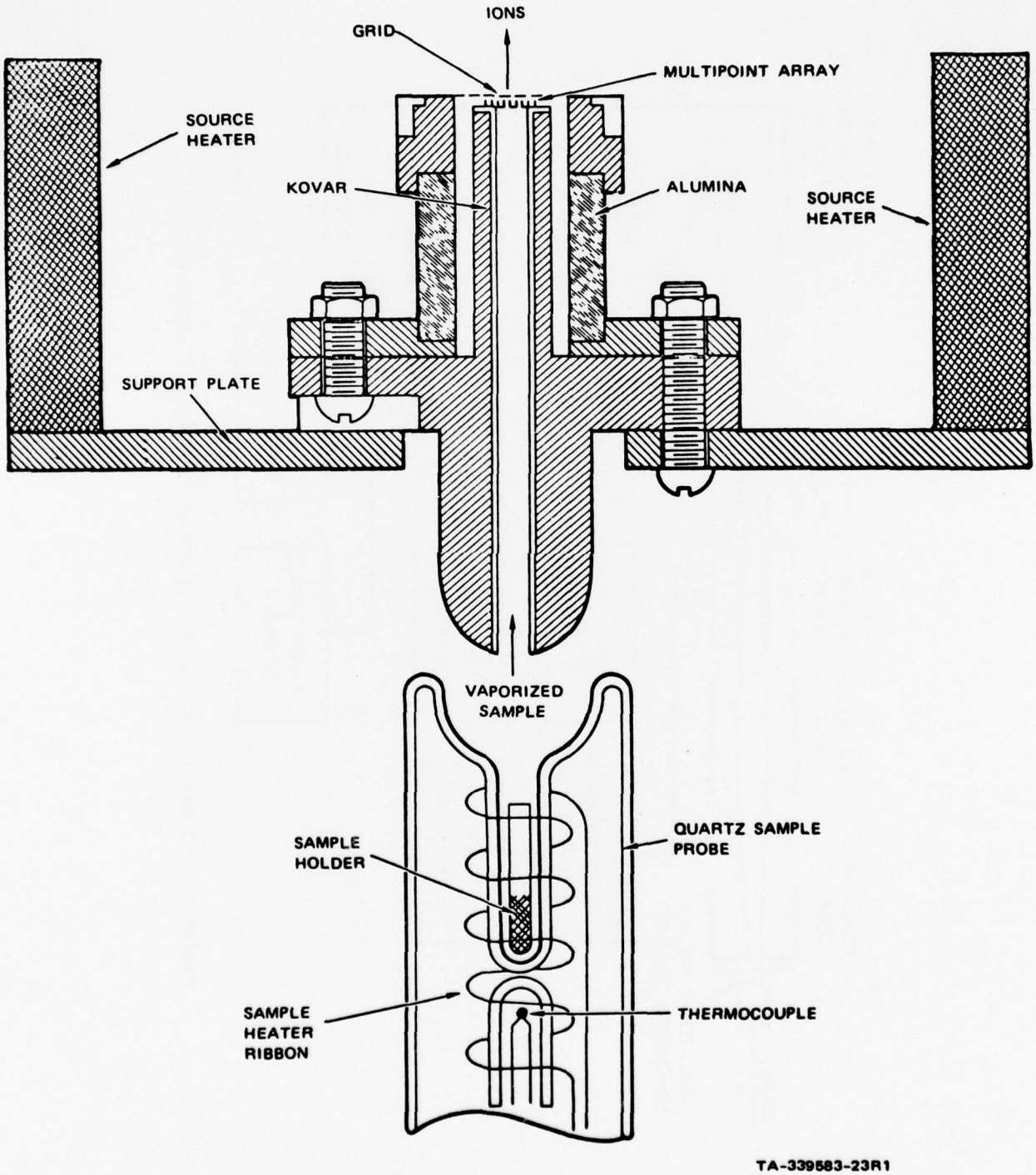


FIGURE 1 SCHEMATIC CROSS SECTION OF MULTIPOINT IONIZER AND SOLID SAMPLE PROBE

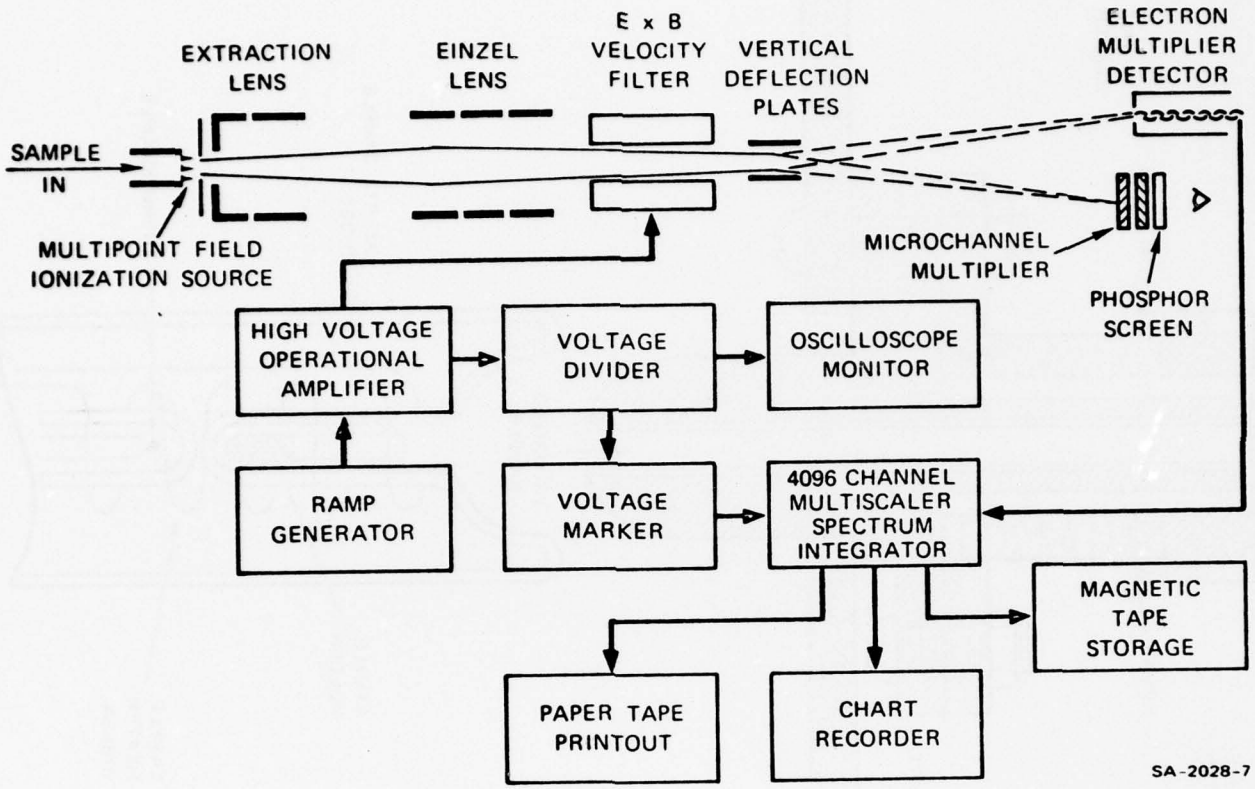


FIGURE 2 SCHEMATIC OF FIELD IONIZATION FINGERPRINT APPARATUS

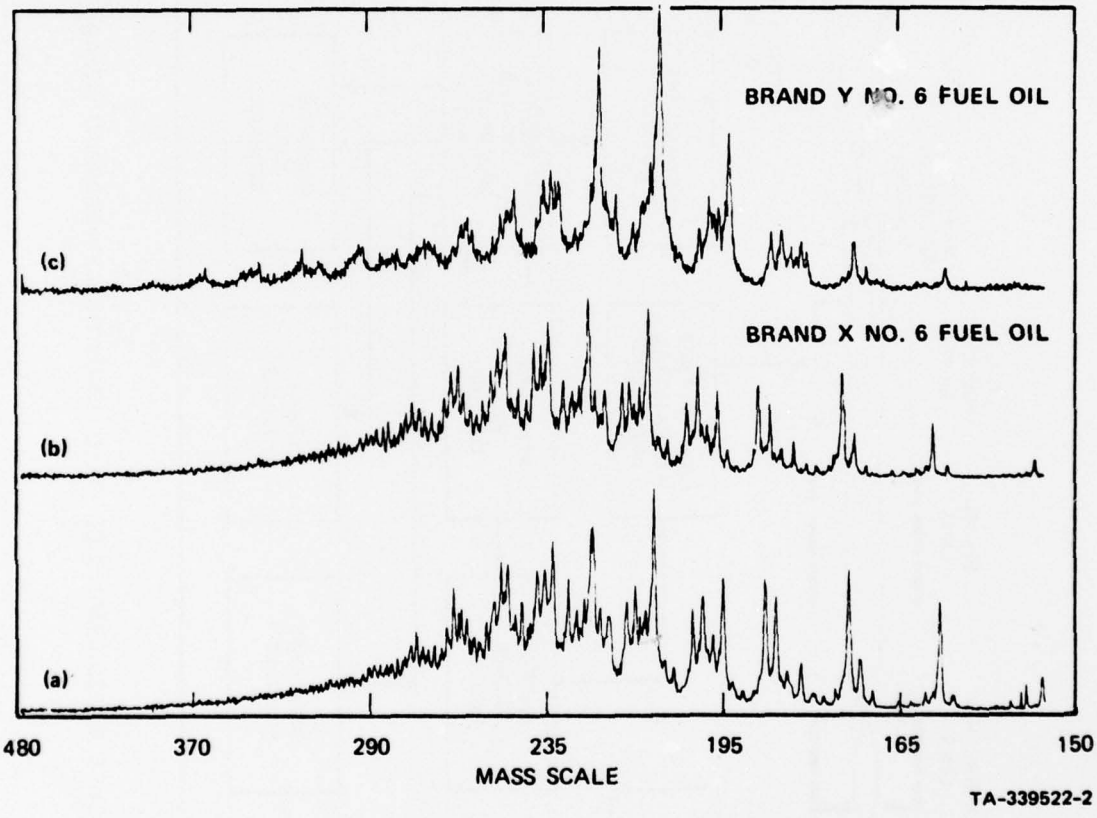


FIGURE 3 MASS "FINGERPRINT" COMPARISONS OF FUEL OIL SAMPLES

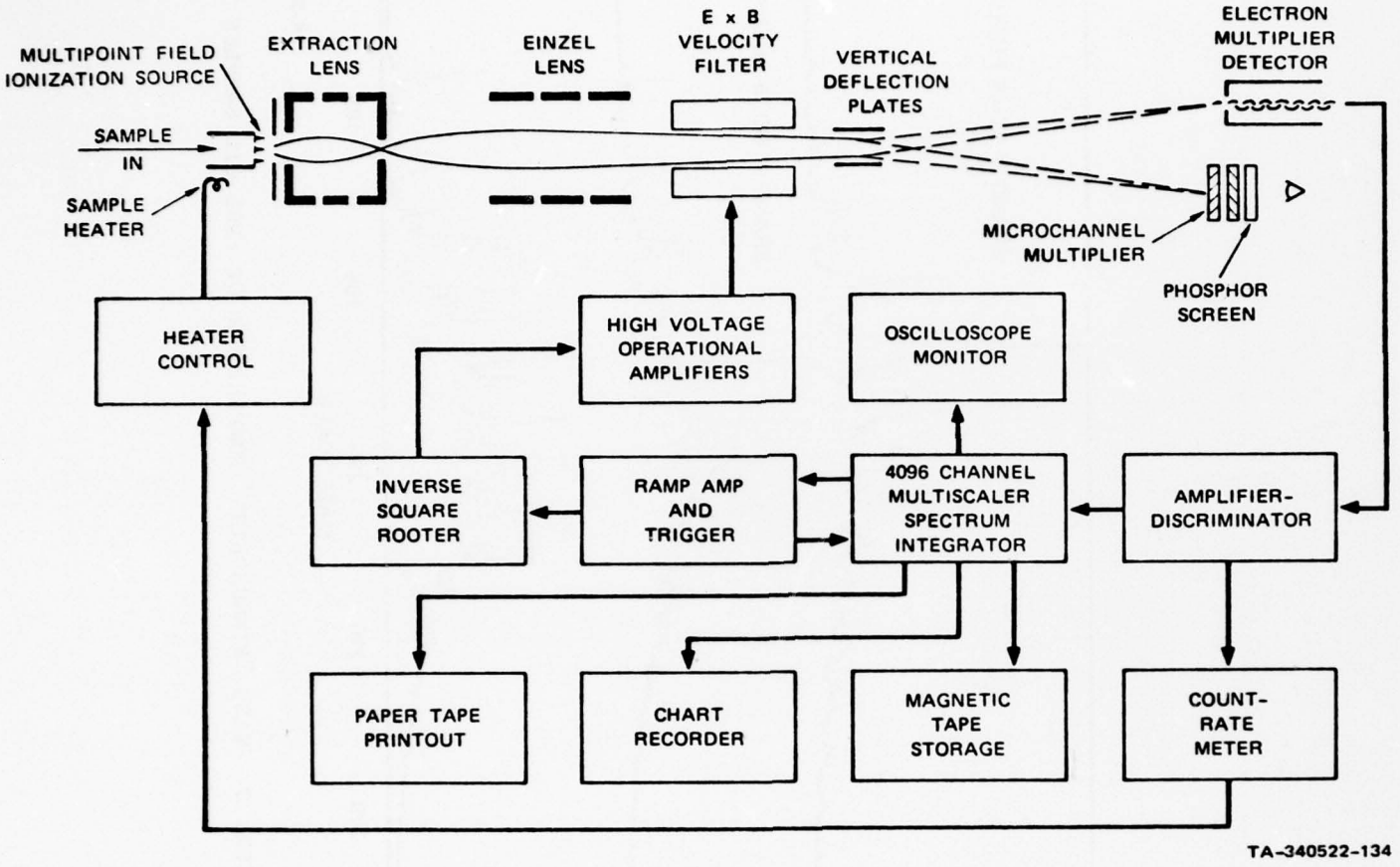


FIGURE 4 SCHEMATIC OF FIELD IONIZATION FINGERPRINT APPARATUS

in place of the external function generator voltage of the first system. The linear ramp voltage was used as the input to an inverse square root circuit whose subsequent output was a voltage that varied inversely with the square root of time. In this manner, the proportionality of m/e with $1/E^2$ was transformed into a linear function of time. Figures 5 and 6 are examples of the spectra produced by this system. The improvement in resolution is due to the modification in ion optics and to the fact that the velocity filter E - field sweep and the multiscaler memory address sweep were now controlled by the same clock. Note also the linear mass scale. The modified EXB field spectrometer was described in the International Journal of Mass Spectrometry and Ion Physics.[5]

Notwithstanding the improved performance of the velocity filter spectrometer, its operation was difficult due to the necessity of floating the ionization source at 10^4 volts and due to the difficulty in tuning it for wide mass range operation. We therefore assembled the quadrupole[6] spectrometer system shown schematically in Figure 7. Figure 8 is a representative example of the quality of spectra produced by the quadrupole analyzer.

The superior resolution of the velocity filter is evident in a comparison of Figures 5 and 6 with Figure 8. However, the quadrupole was more reliable as indicated by a comparison of the standard deviation data shown in Tables 1 and 2. Since the significance of physical measurements varies approximately with the inverse of the standard deviation, the quadrupole

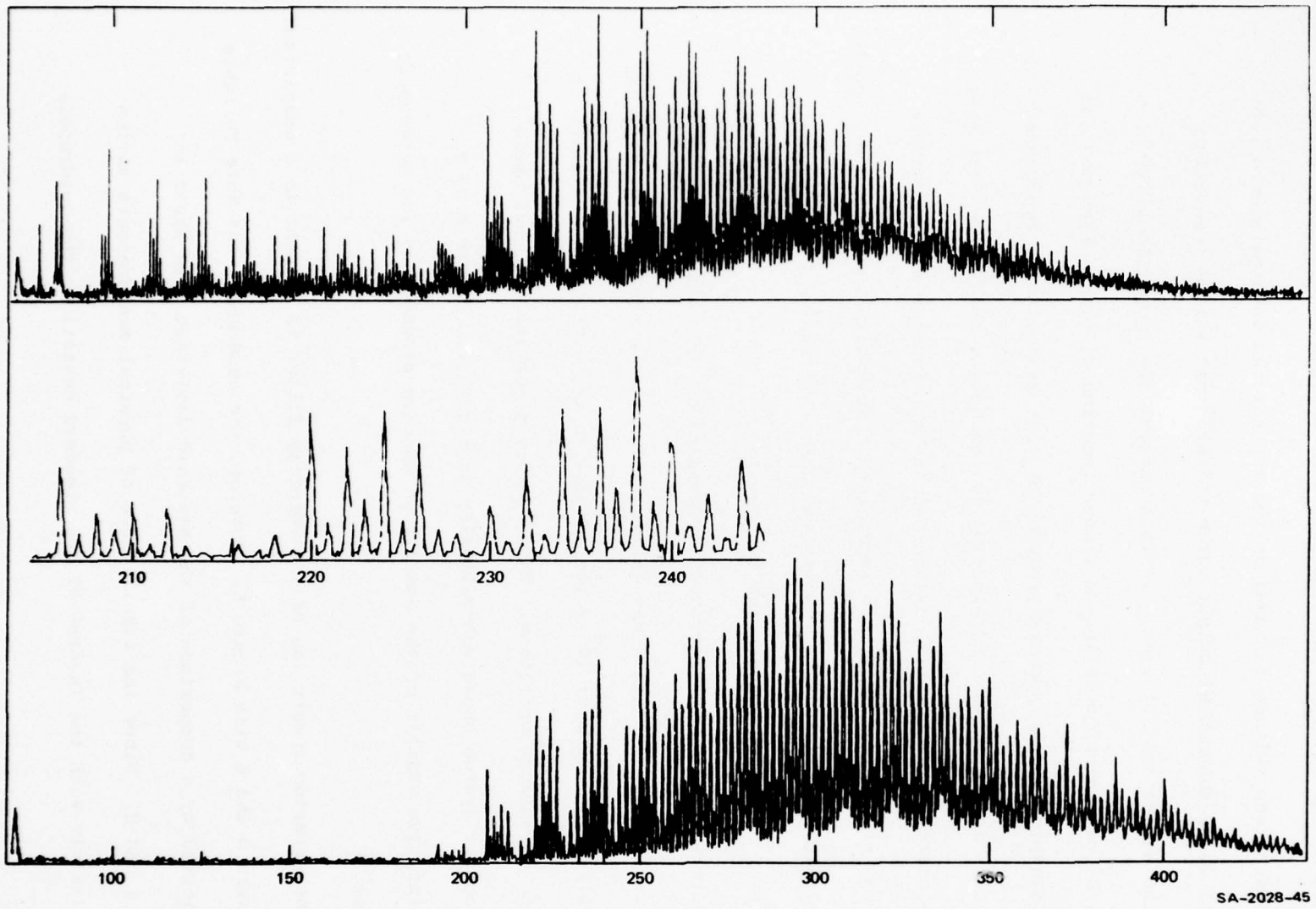


FIGURE 5 WIDE MASS SPECTRA OF STATESBURG, MISSOURI CRUDE OIL

Upper trace shows spectrum of molecular fragment ions produced by sustained electric discharge in source.

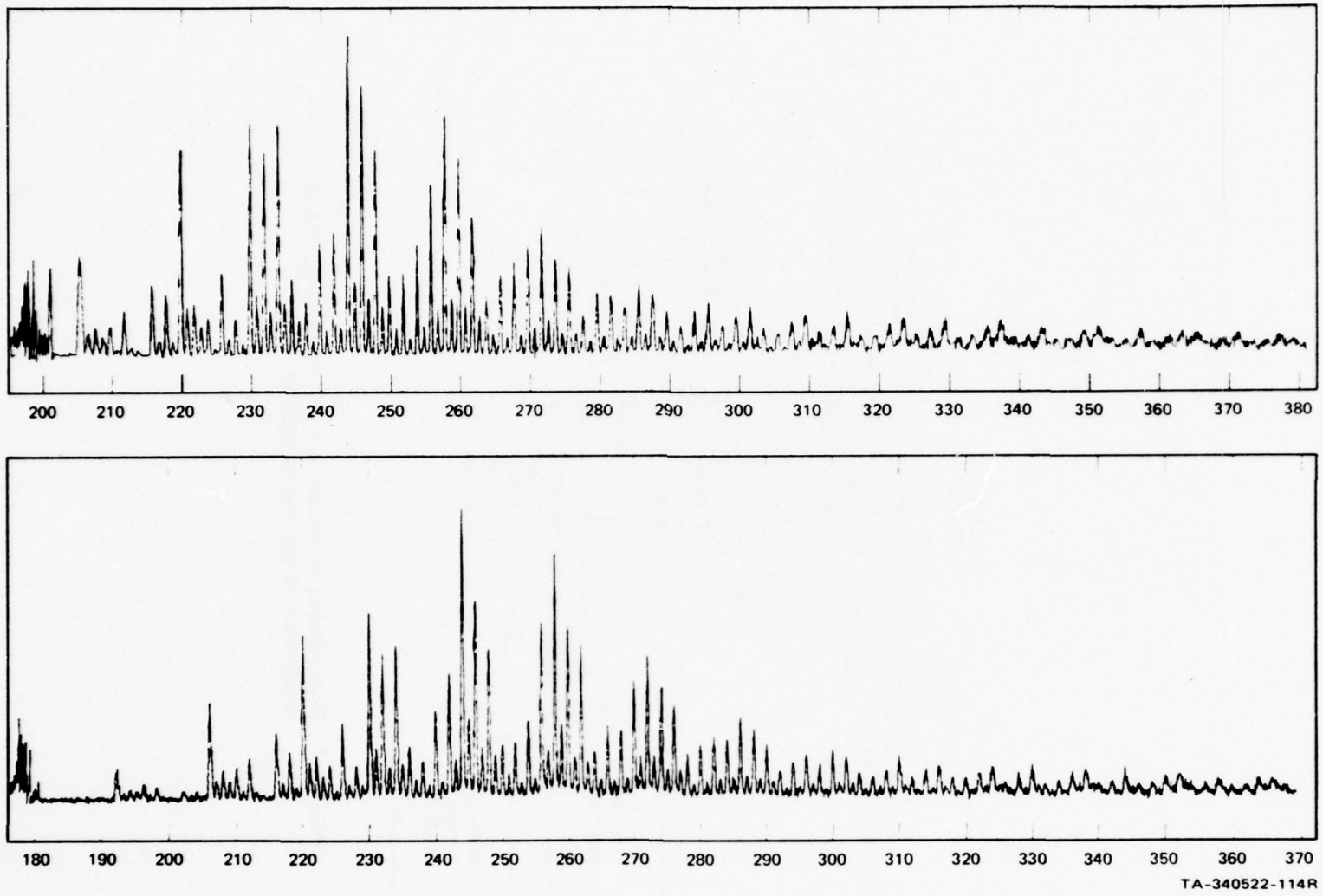


FIGURE 6 TWO FINGERPRINTS OF A SHELL NO. 6 FUEL OBTAINED 3 MONTHS APART

The difference in mass range between the two spectra is due to spectrometer tuning.

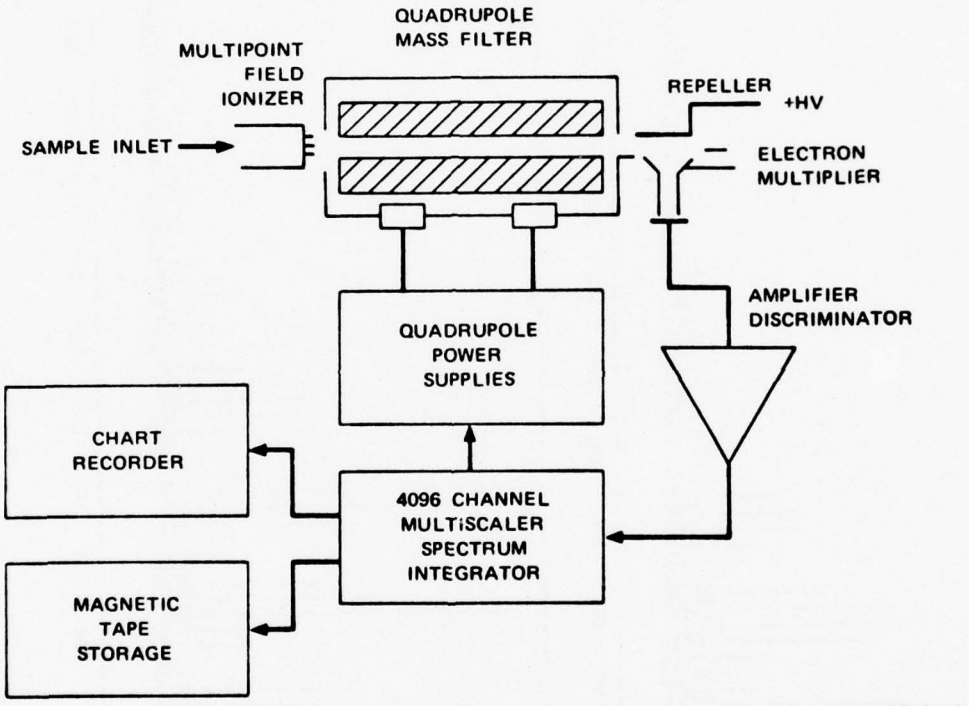


FIGURE 7 QUADRUPOLE INTEGRATING MULTISCANNING FIELD IONIZATION MASS SPECTROMETER

SHELL NO. 6 FUEL OIL

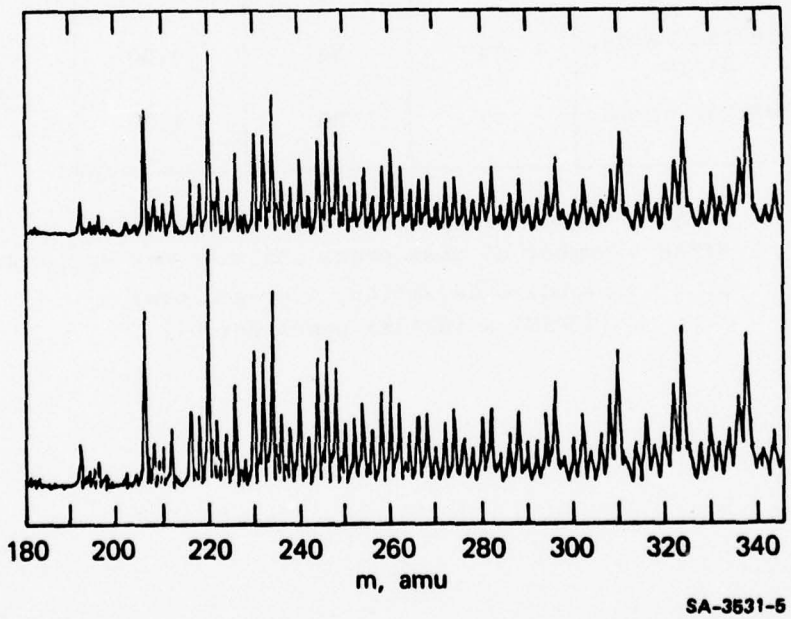


FIGURE 8 QUADRUPOLE SPECTRA OF SHELL NO. 6 FUEL OIL

TABLE 1 Reproducibility Data for Quadrupole Mass Analyzer

$\bar{\sigma}$, averaged over all 6 oils = 6.77%

NAME	NSPEC	NPEAK	$\bar{\sigma}$, %
Shell No. [illegible]	16	33	6.90
Gulf No. 6, Philadelphia	13	34	7.57
Gulf No. 6, Cincinnati	9	49	9.07
Quirequire Crude Venezuela	10	35	4.76
Zelten Crude, Libya	12	33	4.50
Bolivar Crude	10	30	7.85

NSPEC = number of spectra analyzed per oil
NPEAK = number of mass peaks analyzer per spectrum
$\bar{\sigma}$ = standard deviation, averaged over (NSPEC) x (NPEAK) peaks per oil

TABLE 2 Reproducibility Data for Colutron® Velocity Filter

$\bar{\sigma}$, averaged over all 6 oils = 18.2%

NAME	NSPEC	$\bar{\sigma}$, %
Shell No. 6	7	37.2
Gulf No. 6, Santa Fe Springs	6	12.3
Gulf No. 6, Port Arthur	5	11.1
Quirequire Crude, Venezuela	6	17.7
Zelten Crude, Libya	5	13.0
Statesburg Crude, Missouri	5	17.7

NSPEC = sample size = number of spectra

$\bar{\sigma}$ = relative standard deviation averaged over peaks at 9 different mass numbers

was chosen as the preferred analyzer. Figures 5, 6 and 8 also demonstrate a problem that was common to both systems, the degradation of resolution at mass numbers exceeding approximately 280 amu for the quadrupole and 300 amu for the velocity filter.

Normally, field ionization sources are "activated" by growing carbon whiskers on the tips of the points (see Figure 1) by using the source to ionize certain hydrocarbons (e.g., toluene). Deactivation consists of the removal of these whiskers by oxidizing constituents. The most persistent problem encountered in this program was an incompatibility between the multipoint field ionization source and some of the heavier constituents of oils. Each specimen was, therefore, prepared by a vacuum distillation technique to remove both the heavier constituents, which tended to deactivate the source, and the lighter constituents, which were considered to be more sensitive to sample handling and storage temperatures and, therefore, less reliable for "fingerprinting" purposes. Recent changes in the design of our field ionization sources have enabled us to activate them at 1000°C.[7] The resulting pyrolytic carbon whiskers are relatively inert with respect to the chemical action of oil constituents and have made it possible to obtain molecular ion profiles of crude and refined oils without preparation by vacuum distillation. Figures 9, 10 and 11 show examples of spectra obtained with a 60° sector magnet mass analyzer equipped with a pyrolytic whisker field ionization source. This system was used for the analyses of 28 oil specimens supplied to us by the U. S. Coast Guard.

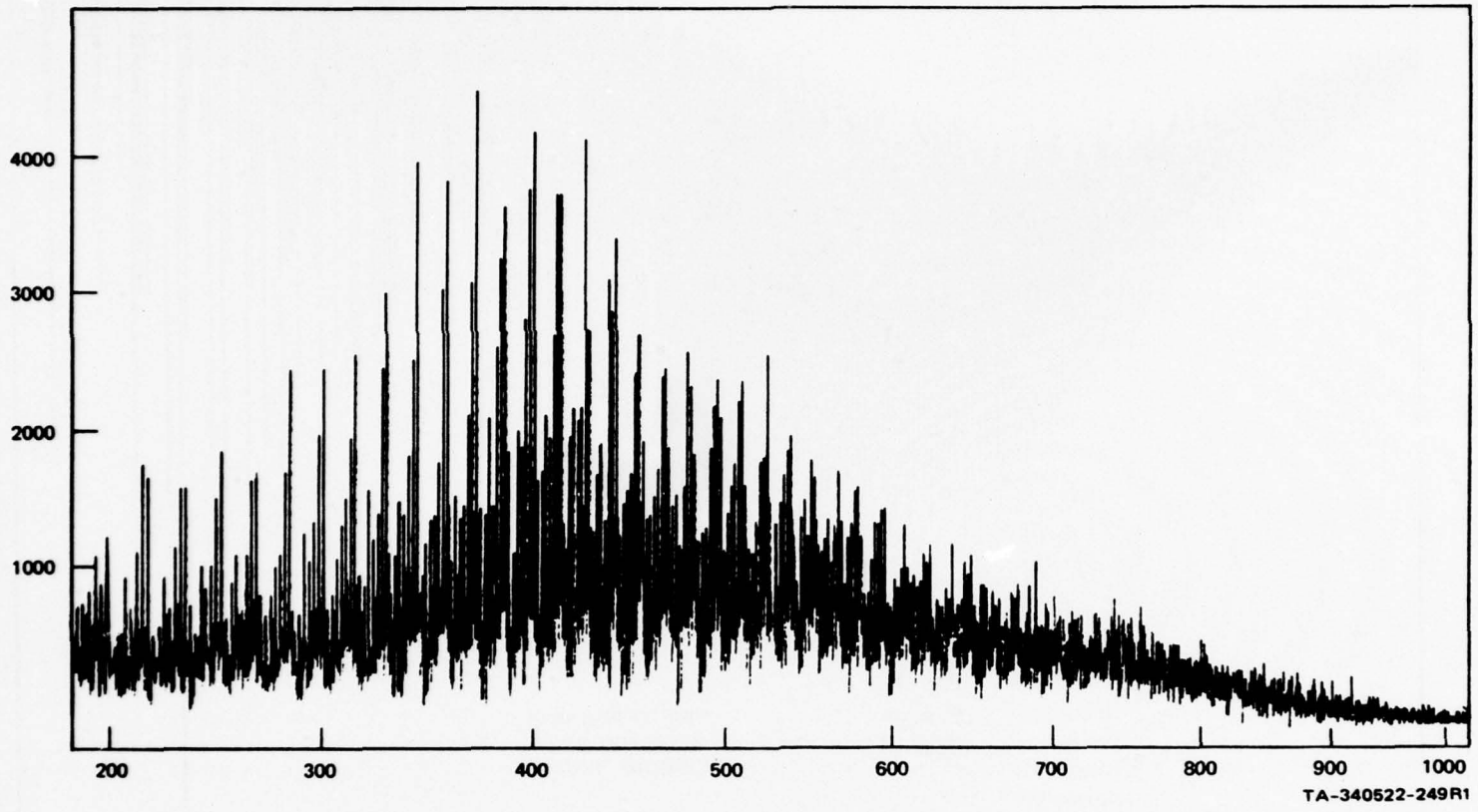


FIGURE 9 60° SECTOR MAGNET SPECTROGRAM OF IRANIAN CRUDE OIL

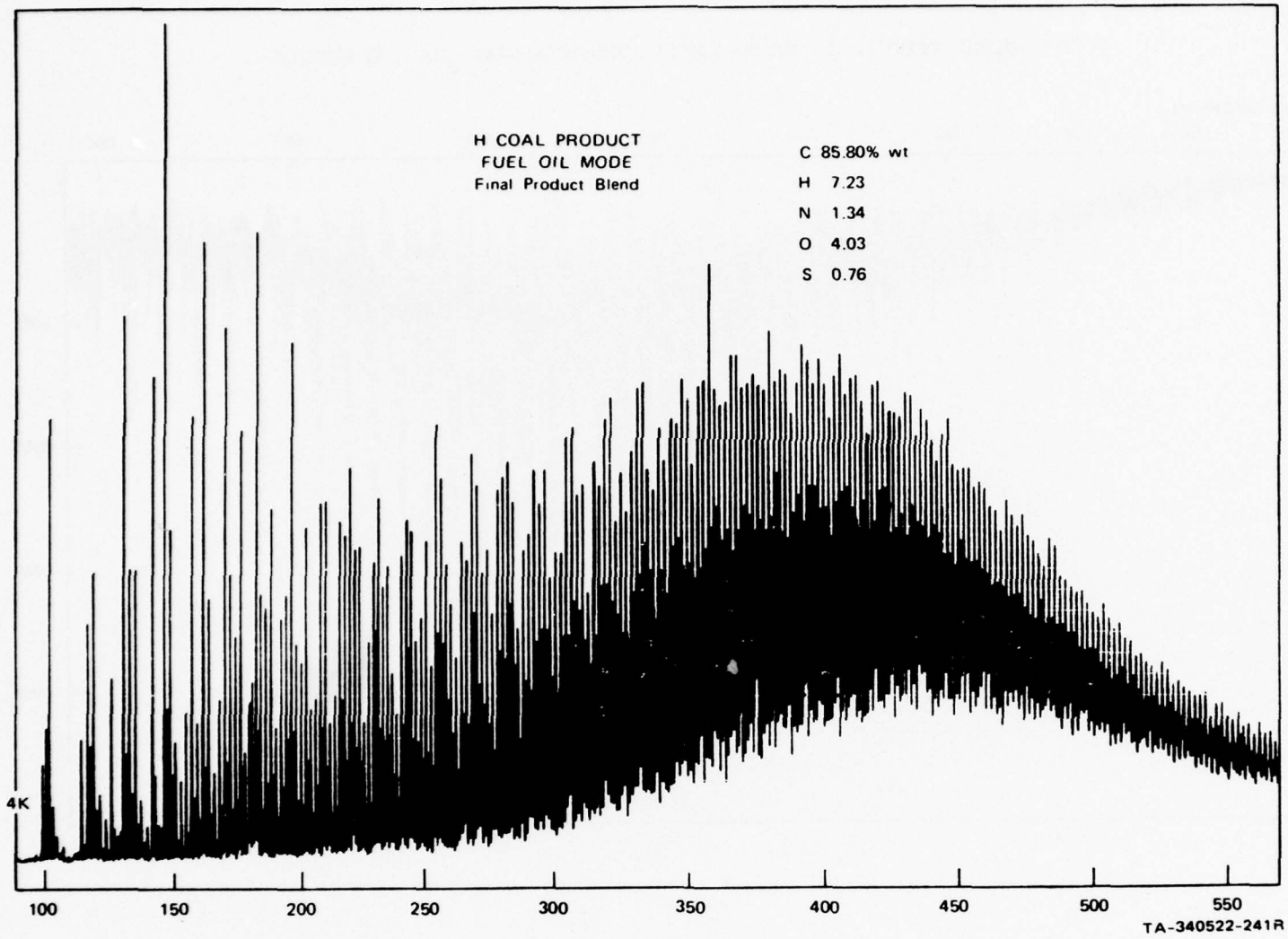
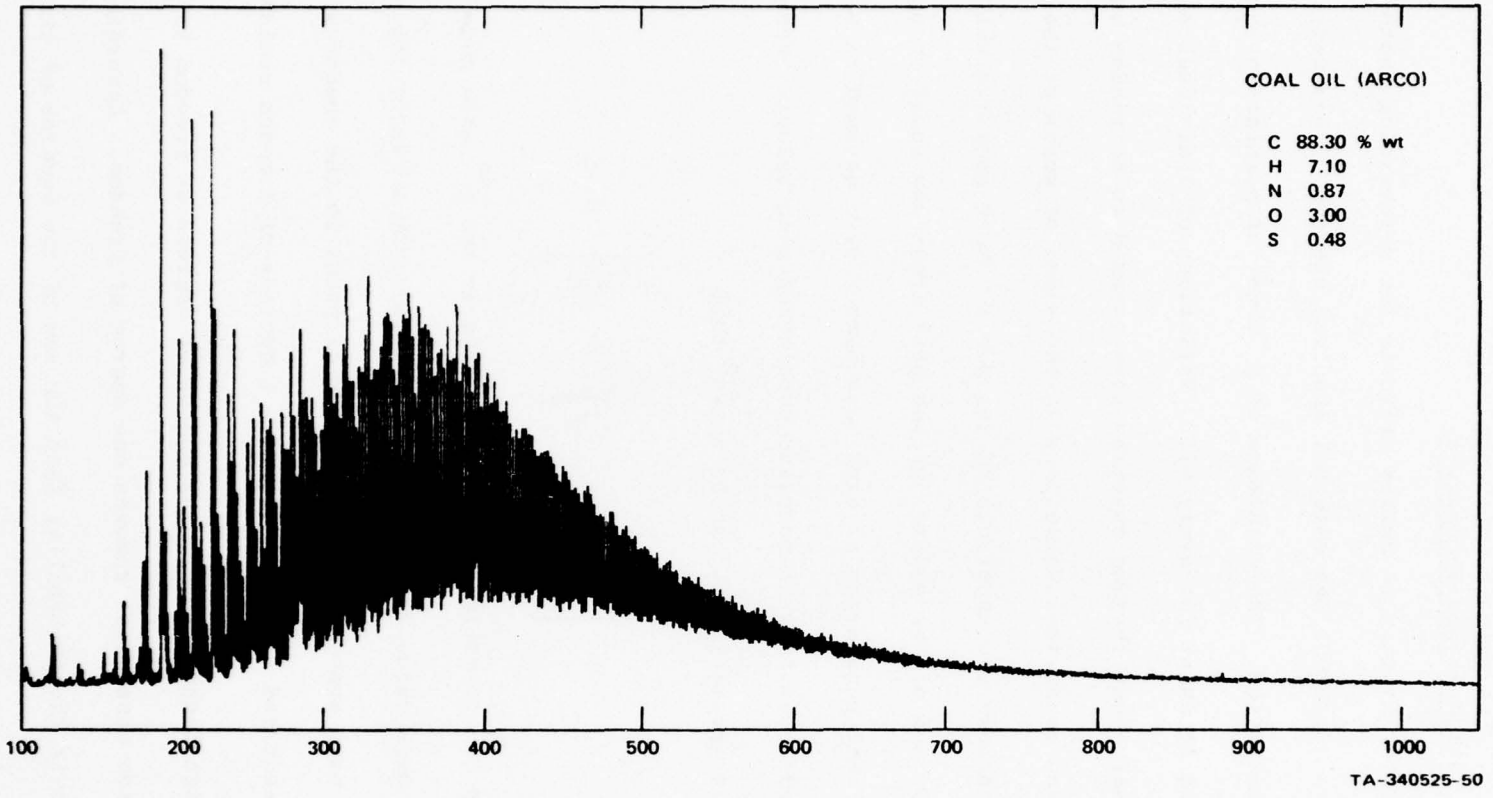


FIGURE 10 60^{o} SECTOR MAGNET SPECTRUM OF COAL PRODUCTS



6T

FIGURE 11 AN EXAMPLE OF A 60° SECTOR MAGNET WIDE MASS RANGE SPECTRAL ANALYSIS

Review of Data Analysis Techniques

Our first attempt to define criteria for determining whether or not two spectra represent the same oil involved the use of parametric multivariate statistics. The existence of a "true" or population mean spectrum was assumed to exist for every oil. Variations of individual spectra of the same oil about the true spectrum were assumed to be random and independent. Independence means that variations in the sizes of peaks at the j^{th} and k^{th} mass numbers are uncorrelated for $j \neq k$. Each spectrum was normalized so that the sum of the squares of the peak areas was equal to unity. The normalized peak areas at each mass number were assumed to be normally distributed about their respective population mean values. When these assumptions are valid, it can be shown[8] that

$$P\left[\sum_{j=1}^{J}\left(\frac{x_j - \mu_j}{\sigma_j}\right)^2 > \chi_\alpha^2\right] = \alpha$$

where x_j is the normalized area of the peak at the j^{th} mass number whose population mean value is μ_j, σ_j^2 is the population variance for peak areas at the j^{th} mass number, J is the number of peaks in the spectrum, and χ_α^2 is the Chi-squared statistic with $J - 1$ degrees of freedom evaluated at the 100% significance level. The number of degrees of freedom is $J - 1$, as the normalization constraint removes one degree of freedom. According to the above equation, α is the probability that the sum of the squares of all the spectral differences (normalized to unit variance) will exceed the threshold χ_α^2, given that the x_j's and μ_j's represent the same oil. In other words, χ_α^2 defines

$100(1 - \alpha)\%$ confidence limits for the measurement of molecular ion profiles of the oil whose true spectrum is defined by the μ_j, $j = 1, 2, \ldots, J$.

This approached failed because it was impractical to obtain enough spectra for reasonable estimates of the population means and variances and because it was discovered that the data would not support the assumption of independence.

Criteria for the identification of oils by mass spectrometric fingerprinting can be defined in terms of sample means and sample variances of spectra whose peak area distributions are correlated, by the use of a multivariate generalization of the Student-t distribution.[9] Unfortunately, the number of spectra required for this analysis is greater than the number of peaks in the spectrum. A test for the identity of the spectrum of an unknown specimen with a reference spectrum consisting of peaks at 30 different mass numbers would require at least 31 spectra of the unknown and the reference each. A technique in which this analysis is limited to peaks at mass numbers where spectral differences appear to be significant is not valid, as the generalized Student-t model cannot be applied to peaks at selected mass numbers. Therefore, the generalized Student-t model was abandoned.

At this point, we had to reassess the problem of spectral data analysis. If one regards the task of identifying oils as a forensic problem, the valid use of statistical models is attractive because it provides answers to the questions: "what are the chances for not identifying the culprit?"

and, presumably of equal or greater importance, "what are the chances for making a false identification?" The alternative is an empirical approach: define an identification criterion and test it on a "training set" of spectra of known identities. The following sections consist of a general overview of some of the statistical aspects of using spectral information for identification purposes and a description of the empirical technique that was finally used for classifying spectra and for measuring the similarity between two spectra.

III THE IDENTIFICATION OF OILS BY FIELD IONIZATION MASS SPECTROMETRY

Defining an Identification Criterion

Spectrograms can be ranked in order of their similarity to a reference spectrogram by counting the number of positions in the spectral range at which the relative peak heights or the relative "highs" and "lows" in the spectral envelope are approximately equal. If Z_j is a measure of the difference between two spectra at the j^{th} mass number and t is a fixed threshold value, the number of values of j for which $Z_j \leq t$ can be used as a measure of similarity. This is a binary decision technique; differences either exceed threshold or do not exceed threshold. For identifying complex mixtures by the use of field ionization mass spectrometry, the threshold value t should be chosen so that comparisons between spectra of the same specimen rank high on the similarity scale, while comparisons between spectra of different mixtures result in low similarity scores.

A mass spectrum can be represented by a vector X with components $x_1, x_2, \ldots, x_J$, where x_j is the height or area of the peak at the j^{th} mass number. The vector representation of a spectrum can be normalized so that the sum of all the components or all of the squared components equals unity. The difference between two spectra X and Y can be computed as the difference between their normalized J-dimensional vector representations. The j^{th} component, $x_j - y_j$, of this vector difference is one way to define Z_j.

For the purpose of deciding whether or not two spectra represent the same oil, an identification criterion must be chosen. The latter may be defined as an arbitrary lower limit for the degree of similarity. To be concise, H_o is the null hypothesis: spectra X and Y represent the same oil. Reject H_o when $Z_j > t$ for at least R values of j, j = 1, 2, . . ., J. The degree of similarity is given by J - R, where R is called the "Hamming distance." Clearly, if a small value for t is chosen, H_o will be rejected in a large number of cases where X and Y represent the same oil. Conversely, if a large value of t is chosen, H_o will be accepted in a large number of cases where X and Y represent different oils. A prudent decision on the value of t can be made on the basis of a statistical model or empirically, as a result of the analysis of a set of spectra of known identity.

Statistical Model

Assume H_o is true and suppose that for a prescribed probability α a threshold t can be determined such that

$$P\left[\left|Z_j\right| > t\right] = \alpha, \text{ for } j = 1 \text{ or } 2 \text{ or } \ldots \text{ or } J \qquad (1)$$

Assume that the spectral differences Z_j and Z_k are stochastically independent for $j \neq k$. The requirement $Z_j \leq t$ for all J values of j as the identification criterion implies

$$P\left[\text{accept } H_o \,\middle|\, H_o \text{ is true}\right] = (1 - \alpha)^J \qquad (2)$$

$$P\left[\text{reject } H_o \,\middle|\, H_o \text{ is true}\right] = 1 - (1 - \alpha)^J \qquad (3)$$

Equation (3) describes the probability for incorrectly rejecting H_o.

Suppose we are willing to tolerate this error in 100 A% of the spectrum pairs analyzed. Then

$$1 - (1 - \alpha)^J = A \tag{4}$$

Solving for α obtains

$$\alpha = 1 - (1 - A)^{1/J} \tag{5}$$

Combining equations (1) and (5) yields

$$P\left[\left|Z_j\right| > t\right] = 1 - (1 - A)^{1/J} \tag{6}$$

Equation (6) enables us to choose a threshold t so that the risk for error is equal to A when the number of possible spectral differences is J and when the identification criterion requires zero spectral differences greater than t.

As an example, consider the case $Z_j = (x_j - \mu_j)/\sigma_j$, where x_j is distributed normally with mean μ_j and variance σ_j^2. In other words, we are comparing the vector representation of a single spectrum of a particular oil to the vector representation of the "true" spectrum of the same oil. For spectra consisting of, say, J = 20 peaks and an acceptable risk of A = 0.05 the value of t for which

$$P\left[\left|\frac{x_j - \mu_j}{\alpha_j}\right| > t\right] = 1 - (0.95)^{0.05} \simeq 0.0025 \tag{7}$$

is 3.01, as found in statistics tables.

The identification criterion: $\left|Z_j\right| \leq t$ for at least J - 1 values of j, implies

$$P\left[\text{accept } H_o \;\middle|\; H_o \text{ true}\right] = (1-\alpha)^J + J\alpha(1-\alpha)^{J-1} \qquad (8)$$

In general, the identification criterion: $|Z_j| \leq t$ for at least $J - R$ values of j, implies

$$P\left[\text{accept } H_o \;\middle|\; H_o \text{ true}\right] = \sum_{r=0}^{R} \frac{J!}{(J-r)!\,r!}\,\alpha^r(1-\alpha)^{J-r} \qquad (9)$$

$$A = 1 - \sum_{r=0}^{R} \frac{J!}{(J-r)!\,r!}\,\alpha^r(1-\alpha)^{J-r} \qquad (10)$$

Note that $1 - A$ is the confidence level of this test for the null hypothesis: X and Y represent the same oil.

The statistical model elucidates two characteristic properties of spectral fingerprinting. First, the significance of a spectral difference decreases as the number of possible differences increases. In comparisons of spectra of the same oil, the chances for a large difference to occur at random are greater when the mass range spans 400 amu than where there are only 20 possible places for differences to occur. Second, the spectral difference threshold, required for a given confidence limit, decreases as the number of allowable differences in the identification criterion increases. These two characteristics are demonstrated in Table 3.

The threshold values in Table 3 were obtained in the following way: (a) The confidence level $1 - A = 0.98$ was chosen; (b) Equation (10) was solved for α by successive approximations for the cases $R = 0, 1, 2$ and $J = 10, 20, 40$; (c) A normal distribution with unit variance was assumed for Z_j; (d) The threshold values for t were found by using

TABLE 3: Threshold values t as functions of J, the maximum number of spectral differences possible, and R the maximum number of differences acceptable for the identification test at the 98% confidence level. Errors are assumed to be normally distributed and independent.

R	J = 10	J = 20	J = 40
1	3.09	3.29	3.48
2	2.28	2.54	2.78
3	1.87	2.18	2.45

$$\alpha = 1 - \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} \exp\ (-Z^2/2)\ dZ$$

or

$$\alpha = 2\ (1 - F(t)).$$

where F(t) is the cumulative normal distribution, found in statistics tables.

The true spectrum of an oil is usually not known and the cost of obtaining a good estimate of it by analyzing a large data sample is prohibitive for routine analyses. The Student-t statistic can be used to test the identity hypothesis in terms of means and variances computed from small samples.

The sample mean vector representation of a set of N spectra is

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_J \end{pmatrix}$$

where $\bar{x}_j = 1/N \sum_{i=1}^{N} x_{ji}$, $j = 1, 2, \ldots, J$ and x_{ji} is the j^{th} component of the i^{th} spectrum.

The sample variance for the j^{th} component is

$$S_j^2 = 1/N \sum_{i=1}^{N} (x_{ji} - \bar{x}_j)^2$$

The statistic

$$Z_j = \left[(\bar{x}_j - \bar{y}_j) - (\mu_j - \lambda_j) \right] \Big/ \sqrt{S_p^2 \, (1/N_x + 1/N_y)} \qquad (14)$$

where $\bar{x}$ is a sample mean of size N_x drawn from a normal population with mean μ_j, $\bar{y}_j$ is a sample mean of size N_y drawn from a normal population with mean λ_j, and S_p^2 is the pooled sample variance, is distributed as the Student-t statistic with $N_x + N_y - 2$ degrees of freedom. This statistic can be used to test the hypothesis: $\mu_j = \lambda_j$, $j = 1, 2., \ldots, J$ (i.e., the identity hypothesis) or to test the hypothesis: $\mu_j \neq \lambda$, $j = 1$ or 2 or . . . J. In both cases, it is assumed that the x_j and y_j have common variances, although the case of unequal variances known as the Behrens-Fisher[8] problem can also be analyzed. The Student-t statistic is used for finding the threshold value t as a function of the sample sizes N_x and N_y and as a function of the value of α that satisfies equation (10). In principle, this statistic can be used for estimating the risk of rejecting H_o when H_o is true (error of type 1) as well as the risk of accepting H_o when H_o is false (error of type 2). By estimating the latter error, it is possible to construct a "power function" which can be used for comparing tests of identity. For example, threshold values can be found for identity tests that involve different degrees of similarity but have the same confidence level. A test that correctly accepts H_o when all J spectral differences are less than t_o may have the same confidence level as a test that correctly accepts H_o when at least J-1 spectral differences are less

than $t_1 < t_o$. However, the two tests will, in general, differ in their ability to reject H_o when H_o is false.

The statistics model for choosing threshold values relies on the assumption that spectral differences are stochastically independent. A multivariate test for this assumption can be found in the literature.[9] The independence hypothesis was tested on samples of field ionization mass spectra ranging in size from 9 to 17. In general, spectral differences at different mass numbers were not independent. In retrospect this result could have been anticipated. Oils are introduced into the spectrometer's ionizer by means of a temperature controlled solid insertion probe. The more volatile constituents of these complex mixtures are mass analyzed at the beginning of the half hour data collection period, during which the entire mass range is scanned approximately 1000 times and integrated into a single spectrum. Since volatility decreases approximately with molecular weight, variations in operating parameters with time produce errors that are correlated with molecular weight. It should be emphasized that the correlations of these errors are crucial to the independence hypothesis but not necessarily to the total error magnitudes. The efficiency of the ionization source is known to change during a single mass analysis and may be responsible for some of the observed systematic errors. Similar problems can be anticipated in the analysis of spectral data obtained by

other analytical techniques. For example, the responsivities of gas and liquid chromatography detectors are generally functions of temperature and the flow rate of the carrier medium. The acquisition of a chromatogram consisting of 20 or more peaks will usually require enough time for the drifts in temperature and flow rate to be significant (i.e., to produce correlation coefficients that are significantly different from zero). Thus the use of statistics models, whose validities rely on the requirement of independent errors, does not appear to be a practical solution to the problem of analyzing real spectral data. Notwithstanding this conclusion, the model described above elucidates the relation between the significance of large spectral differences at a few mass numbers and small differences at many mass numbers and forms a basis for the empirical approach described in the next section.

Empirical Model

The empirical approach differs from the statistical model in the way in which threshold values are chosen. The definition of an identification criterion or "discriminant function" is essentially the same in both models with the exception that the empirical case threshold values are usually not restricted to a single value (i.e., the usual case is t = t(j)). Empirical techniques are described in the literature under the headings "Pattern Recognition" and "Learning Machine Techniques."[10-12] As this literature has become extensive in recent years,[13,14] we will limit the following discussion to those techniques used in this project.

To define a discriminant function for determining whether or not X and Y represent spectra of the same oil, we need to define threshold values t_j such that $|Z_j| \leq t_j$, for at least J-R values of J, implies identity and $|Z_j| > t_j$, for at least R + 1 values of j, implies that X and Y represent different oils. This discriminant function differs from the identification criterion used in the statistical model only in the use of J threshold values t_j. Note, however, that the use of a single threshold value in the statistical model was merely a mathematical convenience and not a logical necessity. In the ideal empirical case

the t_j are simply chosen so that $Z_j \leq t_j$ for all X and Y that represent spectra of the same oil. As additional spectra are acquired, the t_j are adjusted so that the condition $Z_j \leq t_j$ is maintained for all vector pairs in the set. If, by applying this empirical method to the sets of spectra representing many oils, every spectrum can be unambiguously identified with its respective class (type of oil), we say that the data are "separable." The assumption is made that if sufficiently large sets of data are used - they are referred to as "training sets" - the values of t_j will converge to their respective limits. At this point the vector representation of an unknown spectrum can be classified correctly by applying the discriminate functions iteratively to differences between the unknown spectrum and spectra from each of the training sets.

Two problems that can arise are: (1) The unknown spectrum may not be a member of any of the training classes, in which case correct identification is impossible. (2) The training sets may not be completely separable, in which case a spectrum will sometimes be identified with more than one oil. Making the wrong choice in this case is equivalent to an error of type 2 in the statistical model.

In addition to the above problems, the identification of oils for forensic purposes often involves comparisons between weathered and unweathered specimens of the same oil. For an oil whose spectrum falls outside the set of training classes, such as a weathered oil, the degree

of similarity enables us to specify the class that most resembles the unknown. Furthermore, in cases of ambiguous identity, it facilitates a choice between candidate classes. Therefore, the use of the degree of similarity, defined in the statistics model, is a desirable adjunct to the empirical classification technique.

Reduced Dimensionality

Well resolved spectra contain copious quantities of information. For example, the number of distinctly different mass spectra that are possible with unit mass resolution over a range of 100 amu and with peaks that can be resolved into 10 significant magnitudes is 10^{100}. State of the art field ionization mass spectrometry is capable of producing molecular ion profiles of complex mixtures, such as crude and refined oils, that span more than 400 amu. Examples of wide mass range molecular ion spectra are shown in Figures 9, 10, and 11.

The computer analysis of 400-dimensional vectors is expensive and unnecessary. The envelope of a spectrum can be digitized by using a technique that imitates analog to digital conversion. The envelope of a 400-peak spectrum can thus be transformed into a histogram whose spectral range (abcissa) is partitioned into, say, 10 intervals. In other words, a 400-component spectral representation can be transformed into an arbitrarily smaller dimensionality. Moreover, a unique

advantage of field ionization mass spectrometry is that the envelopes of molecular ion distributions of crude oils attain their maximum values at high mass numbers, while those of refined oils peak at relatively low mass numbers. In fact, it is usually possible to identify oils as crudes or refined products by visual inspection of their spectrograms, as illustrated by Figures 5 and 6. Therefore, in the case of field ionization spectra, the cumulative spectral distribution appears to be useful as a means of describing the gross characteristics of oil spectra. For example, the fifty percentile point is equal to the mass number at which the spectrum can be divided into two equal areas. The fifty percentile points for refined oils will, in general, be smaller than those of crude oils. Thus the use of the digitized cumulative spectral distribution provides a means of categorizing oil spectra as well as reducing the dimensionality of the computer analysis.

The use of the digitized cumulative distribution of a spectrum rather than its constituent peaks precludes many of the problems associated with spectrometer resolution and the deconvolution of fused peaks. To a large extent, it provides a common denominator for sets of spectra whose quality (resolution in particular) is variable. The obvious price that is paid for the advantages of this technique is a loss of spectral information. Notwithstanding this limitation, we were able to classify 154 spectra into 35 different categories with 95% success.

The use of digitized cumulative spectral distributions was conceived as a means of programming a computer to separate the spectra of crudes from those of refined products, thus avoiding the cost of unnecessary detailed spectral comparisons. A computer can be programmed, for example, to avoid identity tests between spectra whose respective fifty percentile points are separated by more than 20 amu. Furthermore, the resolution of our data, over long periods of time, has been inconsistent due to the fact that considerable "hardware" development has occurred during the course of this project, and techniques based on the use of cumulative distributions appeared to have the potentiality for enabling meaningful comparisons to be made between "good" and "bad" data.

The following method for data analysis was devised as an inexpensive and mathematically simple approach to the problem of identifying complex mixtures by the use of cumulative spectral distributions.

Oil Identification by the Use of "No-Resolution Mass Spectrometry"

Consider a smooth and continuous spectrogram of the form $y = y(m)$, where y is the height of the chart recording and m is a continuous variable in units of fractional amu. The percent cumulative spectral distribution can be defined as

$$F(m_1) = \frac{100 \int_{m_0}^{m_1} y\,dm}{\int_{m_0}^{m_f} y\,dm} \qquad (15)$$

where m_0 and m_f are the lower and upper limits respectively of the mass range. Figure 12 shows the percent cumulative spectral distributions of 34 spectra comprising the representations of three different oils. The three distinct families of curves in Figure 12 are evidence of "clustering" or the separation of data into classes. Examples of spectra from these three classes are shown in Figures 13, 14 and 15.

Define M_p by $F(M_p) = p\%$ (e.g., if $m_1 = M_{20}$, then $F(m_1) = 20\%$). The following vector can be constructed

$$X = \begin{pmatrix} x_{10} \\ x_{20} \\ \cdot \\ \cdot \\ \cdot \\ x_{80} \end{pmatrix} \qquad (16)$$

where $x_j = M_{j+10} - M_j$ = the increment in mass range that is traversed as the cumulative distribution passes from the j^{th} percentile point to the $(j+10)$th percentile point. The choice of dimensionality for this vector was arbitrary, the omission of the components x_0 and x_{90} was not. The components of this vector are crude derivatives of the form $\Delta m/\Delta y$. For spectrograms that have no peaks at the upper or lower limits of the spectrometer's sweep range a shift of the mass scale relative to the spectrum has no effect on the vector representation of equation 16 (i.e., the mass scale can be shifted until one of the mass range extremites intercepts a peak). In other words, there is no need for assigning mass numbers to individual peaks. The inclusion of x_o

FIGURE 12

PERCENT CUMULATIVE AREA VS. MASS NUMBER

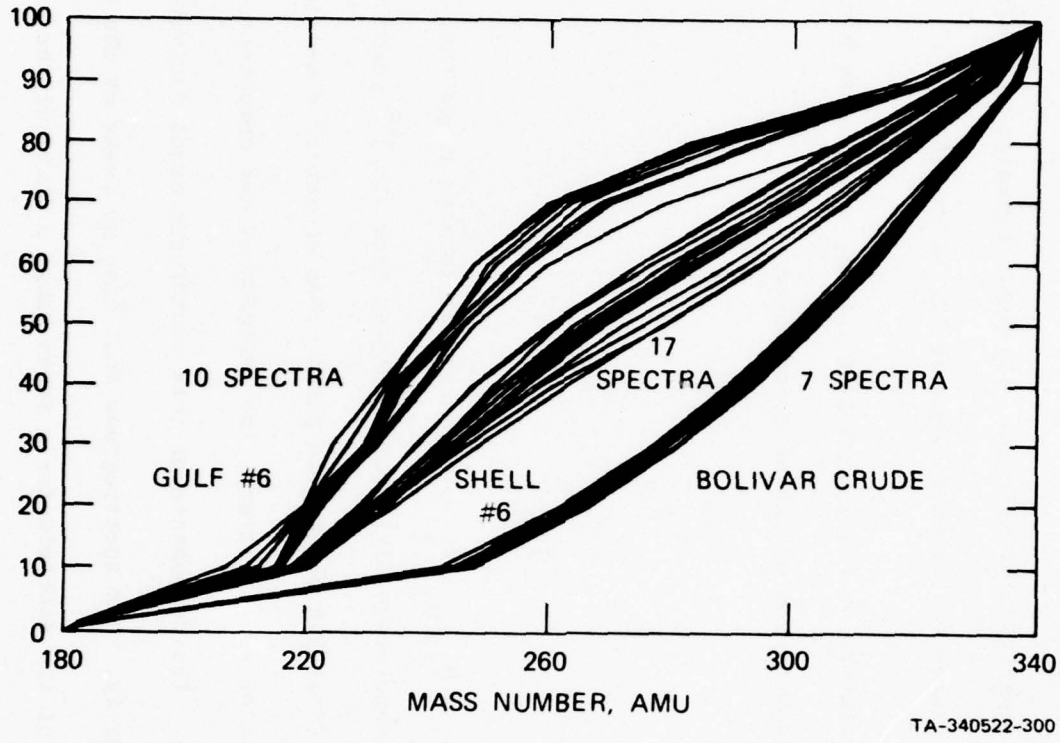


FIGURE 13

GULF NO. 6 FUEL OIL
CINCINNATI REFINERY

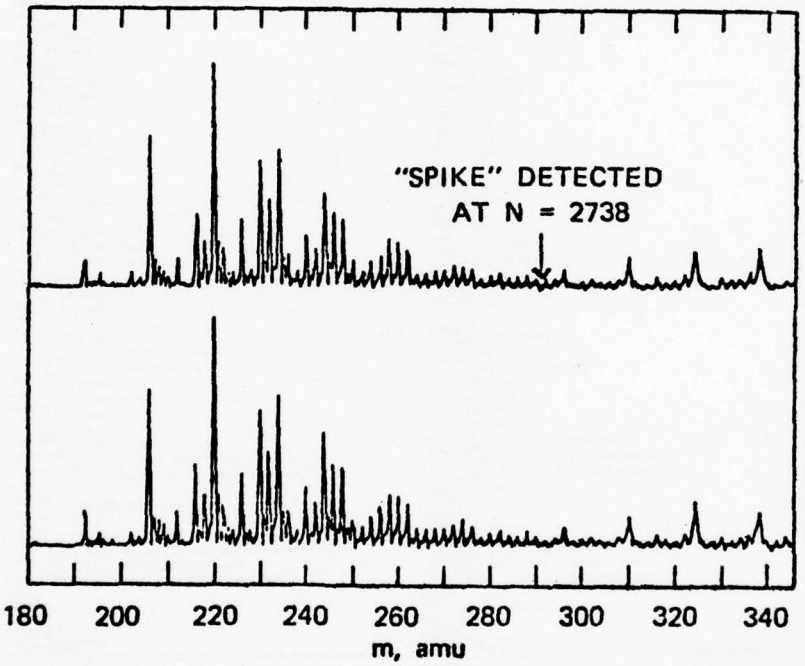


FIGURE 14

SHELL No. 6 FUEL OIL

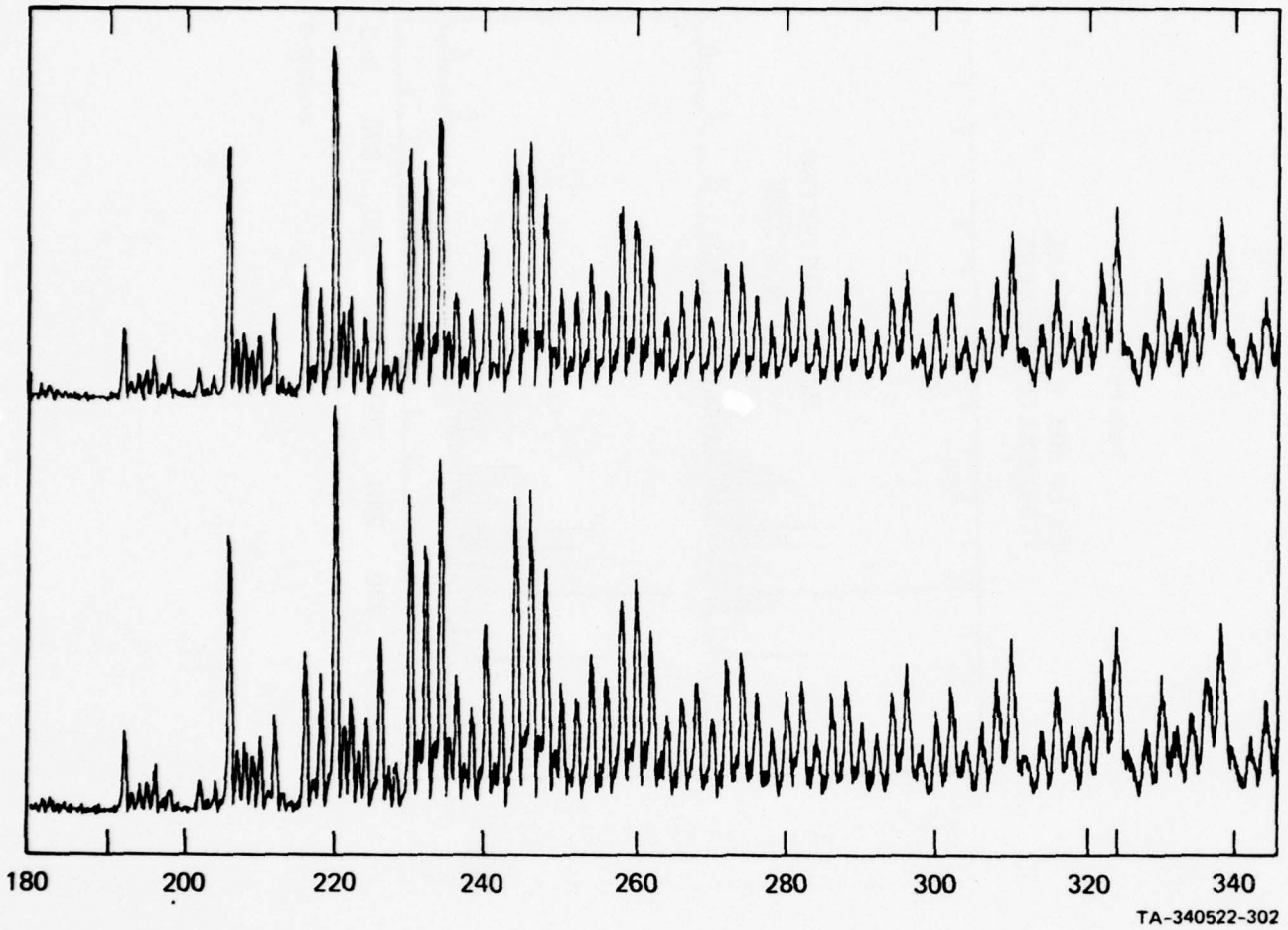
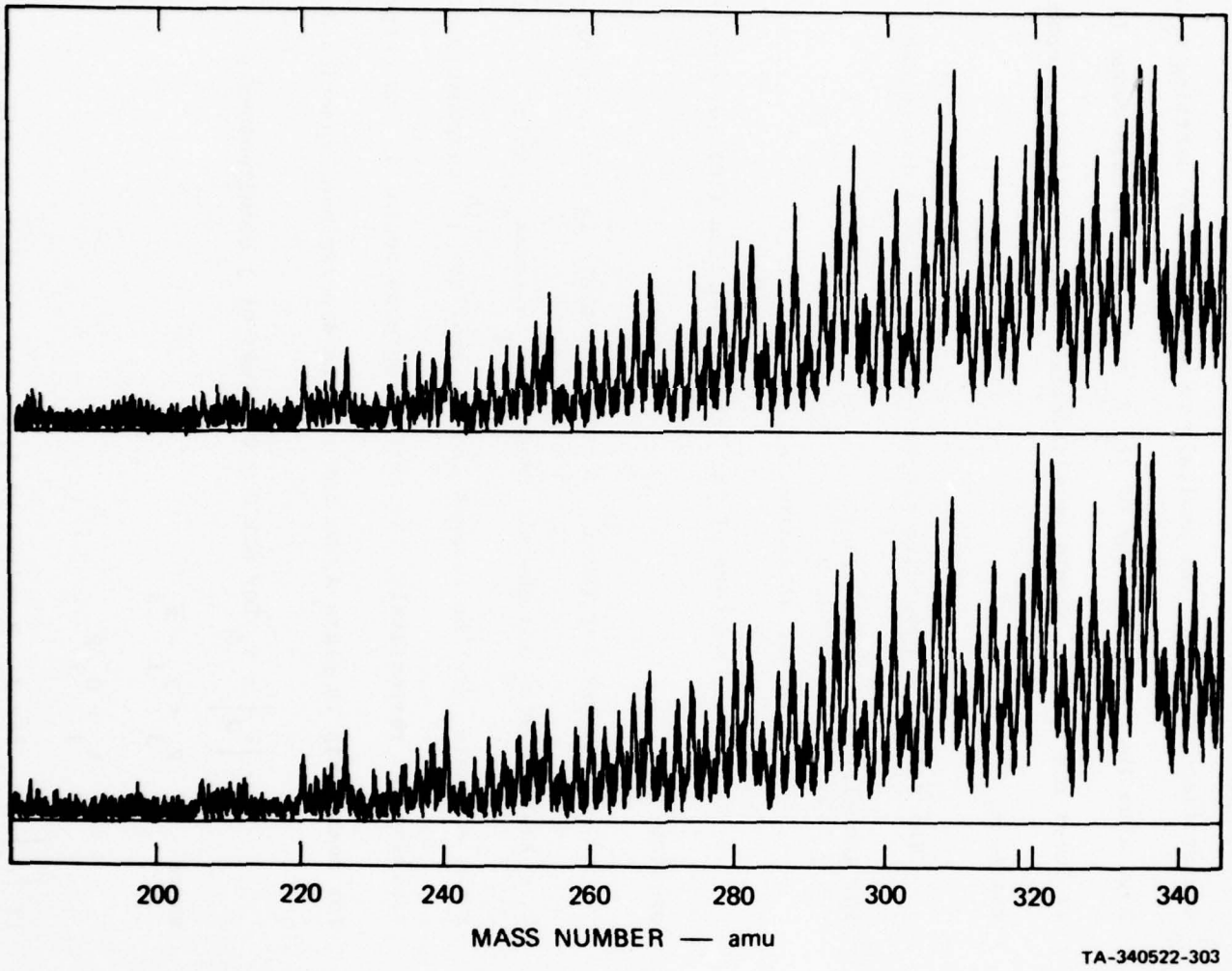


FIGURE 15

BOLIVAR CRUDE



and x_{90} in equation (16) requires that the data analysis span precisely the same mass range for each spectrum. In view of the additional calibration computations required, the analysis is simplified by omitting x_0 and x_{90}. Note also that since the sum of the x_j's equals the mass range (a constant), at least one of them should be eliminated as it is linearly dependent on the rest.

The following procedure can be used to define a discrminant function for identifying spectra of oil A.

1. Obtain a set of spectra of oil A.

2. Construct vectors of the form of equation (16) for each of the spectra.

3. The range for the j^{th} vector component in the training set is $D_j = xmax_j - xmin_j$ and the mid-range is $\tilde{x}_j = (xmax_j + xmin_j)/2$, where $xmax_j$ and $xmin_j$ are the largest and the smallest j^{th} components of the training set respectively. To test an unknown vector Y with components y_j for membership to class A, we can use the discriminant function defined by

$$\left|Z_j\right| \leq t_j \text{ for all } J = 8 \text{ values of } j \text{ simultaneously,} \qquad (17)$$

where

$$Z_j = y_j - \tilde{x}_j \qquad (18)$$

$$t_j = D_j/2 \qquad (19)$$

If $\left|Z_j\right| \leq t_j$ for $J - R$ values of j, we can say that the oil represented by $\tilde{X}$ is an $(R + 1)^{th}$ choice for the identity of Y.

This discriminant function insures the correct classification of all members of the training set. It does not exclude the possibility of identifying a spectrum from a training set with more than one oil, however. The frequency with which training set spectra are classified ambiguously provides a measure for evaluating the merits of this discriminant function before applying it to spectra of unknowns. The identification criterion, equation (17), is equivalent to a partitioning of the vector space into 8-dimensional parallelopipeds with centers at $\tilde{X}_1, \tilde{X}_2, \ldots \tilde{X}_N$, where N is the number of classes in the training set. The success of this technique relies on the analysis of training sets that are large enough to provide good estimates of the dispersion of the data, measured in terms of the range, and of the true vector representation, measured in terms of the mid-ranges of each component. From a practical point of view, this technique appears at first to offer no advantage over statistical models, in which the population means and variances must be estimated. Both models require large sample sizes (i.e., $N \geq 10$). However, the first objective of mass spectrometric fingerprinting is to identify oils and the empirical approach provides a means to meet this objective without the constraints of a specific model for the data.

IV EXPERIMENTAL RESULTS

Quadrupole Data

Table 4 summarizes the results of applying the procedure outlined above to 154 field ionization mass spectra comprising 35 different classes ranging in size from 2 to 17. A quadrupole mass separator was used for analyzing specimens prepared by vacuum distillation. The class number, class code name and class size are tabulated in the first three columns respectively. The use of Table 4 can be described by a few examples: The zero in the "False ID" column of the first class means that none of the 17 spectra of Shl-6 oil was identified with a class other than its own. The oil whose molecular ion distribution is more similar to that of Shl-6 than any of the other oils is G-6, Phil. The "4th Choice" column of the K = 1 row shows that G-6, Phil was the fourth choice for the identity of seven of the Shl-6 spectra. In the case of the nine Zel-Cr spectra (K = 2) the fourth choice was Tobl-Cr for three of them and one class each for three others. The average dispersion of each class of spectral representations is tabulated in the "$\bar{R}$%" column where the entries are equal to $D_j/\bar{x}_j$ averaged over the eight values of j. Most of the classes are smaller in size than 7, in which cases the estimates of range are unreliable. However, considering that the 8-dimensional vectors represent only the gross characteristics of an oil spectrum it is remarkable that ambiguities resulted in only 5% of the spectra and that all of these were falsely identified with the same oil (viz. K = 2).

Table 4 Classification of 154 Oil Spectra Into 35 Classes

K	Code Name	Size	False ID	2nd Choice	3rd Choice	4th Choice	$\bar{R}$%
1	Shl-6	17	0	0	0	7(G-6,Phil	39.4
2	Zel-Cr	9	0	0	0	3(Tobl-Cr)	32.5
	"					Sts-Cr	
	"					SBD1-Cr	
	"					Apa-Cr	
3	G-6,Cin	10	0	0	0	0	60.1
4	QQ-Cr	10	0	7(AmoD1-4)	3(AmoD1-4)	4(ShoD4-6)	11.8
5	G-6,Phil	13	0	0	0	2(shl-6)	31.1
6	Bol-Cr	7	5(Zel-Cr)	2(Zel-Cr)	0	3(SBD1-Cr)	9.4
	"			SBD1-Cr			
7	G-6,S.Fe	3	0	AmoD1-4	AmoD1-4	0	6.3
8	Sts-Cr	3	0	Zel-Cr	Zel-Cr	Zel-Cr	6.8
9	Say-6	3	0	0	2(Zel-Cr)	Zel-Cr	3.6
10	Apa-Cr	3	Zel-Cr	Zel-Cr	0	Zel-Cr	15.7
11	G-6,P.A.	3	0	0	3(Zel-Cr)	Bol-Cr	3.3
12	Ira-Cr	3	0	AmoD1-4	2(AmoD1-4)	Zel-Cr	6.5
13	Tob-Cr	3	0	3(Zel-Cr)	Bol-Cr	SBD4-Cr	3.3
14	TobL-Cr	4	0	2(Zel-Cr)	Zel-Cr	Zel-Cr	8.7
15	Has-Cr	3	0	0	3(Zel-Cr)	0	5.2
16	Say-2	3	0	0	0	AmoD1-4	16.4
17	Say-Lub	2	0	0	0	0	1.4
18	SB-Cr	3	Zel-Cr	2(Zel-Cr)	2(Bol-Cr)	Apa-Cr	4.3
	"			Bol-Cr			
19	SBD1-Cr	3	Zel-Cr	2(Zel-Cr)	Bol-Cr	Bol-Cr	5.3
20	SBD4-Cr	3	0	2(Zel-Cr)	Zel-Cr	Bol-Cr	5.9
	"					TobL-Cr	
21	Shl7	3	0	0	Shl7D1	Shl7D1	13.0
22	Amo-4	3	0	0	0	0	9.5

Continued . . .

Table 4 (Continued)

K	Code Name	Size	False ID	2nd Choice	3rd Choice	4th Choice	$\bar{R}$%
23	Sh17D1	3	0	Sh17	Sh17	0	6.4
	"				Sh17D4		
24	Sh17D4	3	0	0	0	Sh17D1	19.8
25	AmoD1-4	4	0	0	0	Sh1-6	52.7
26	AmoD4-4	3	0	0	0	AmoD1-4	6.9
27	Sh33-6	3	0	2(Sh1-6)	Sh1-6	G-6,Phil	9.8
	"					Sh33D1-6	
28	Sh33D1-6	3	0	0	3(Sh1-6)	0	6.8
29	Sh33D4-6	3	0	0	Sh1-6	2(Sh1-6)	4.1
30	Sho-6	3	0	0	0	0	4.5
31	ShoD1-6	3	0	3(Zel-Cr)	Apa-Cr	Apa-Cr	8.4
32	ShoD4-6	3	0	0	ExoD4-2	AmoD1-4	14.3
33	Exo-2	3	0	3(AmoD1-4)	AmoD1-4	0	4.9
34	ExoD1-2	3	0	2(AmoD1-4)	AmoD1-4	0	6.2
35	ExoD4-2	3	0	0	2(ShoD4-6)	2(AmoD1-4)	7.8
	"					QQ-Cr	

K = class number

$\bar{R}$ = average spectral range in percent

Oil codes: Cr = crude, -N = number N refined oil, D1, D4 = weathered for 1 and 4 days

In an effort to devise an oil identification scheme that can provide a measure of similarity without the requirement of ten or more spectra per training set specimen, the empirical definition for the threshold values was modified. Consider a set of vectors of the form described in equation (16), representing the cumulative spectral distribution of a set of oil specimens. From this set we pick a reference vector X_r. A threshold value for the j^{th} component of X_r can be defined as $t_{rj} = fx_{rj}$, $j = 1, 2, \ldots, J$, where $0 < f < 1$. In this case we will identify a vector X_i with the reference oil whose representation is X_r when $|x_{ij} - x_{rj}| \leq fx_{rj}$ for at least J - R values of j. This discriminant function provides two degrees of freedom, the degree of similarity, J - R, and the threshold factor, f. In this case, where we are trying to avoid the large numbers of spectra required in both the statistics and the learning machine models, we expect to determine both degrees of freedom empirically after analyzing sufficiently large sets of data in which the number of spectra per specimen is generally less than three or four. The requirement for complete similarity (i.e., the requirement R = 0) has been relaxed to avoid the large sample sizes needed for estimating the ranges of data. Since there is no way to estimate appropriate threshold values a priori because the data are not independent in the statistical sense, we will determine these values empirically in terms of how well they separate "known similars" from "known dissimilars."

To apply this analysis all the possible pairs in a set of vectors are analyzed with one member of the pair acting as reference. The analysis is performed iteratively as the parameter f is varied systematically. The results are examined for the best combinations of J - R and f.

The individual quadrupole spectra comparising the 35 classes listed in table 4 are identified with their respective class numbers and oil code names in table 5. Table 6 shows the results of using the set of 154 quadrupole spectra in a similarity measurement test. One vector from each of the six largest classes was used as a reference and compared to the other 153 8-dimensional representations. For each threshold factor f the numbers of correct and incorrect identifications are tabulated as functions of the degree of similarity J - R. For example, using spectrum no. 21 as reference, the remaining 9 members of class 3 were correctly identified while 8% of the 143 spectra from other classes were incorrectly identified when a similarity index of J - R = 2 was used with a threshold factor f = 0.16.

TABLE 5 Class Identification and Spectrum Numbers of 154 Quadrupole Spectra

N	K	NAME	N	K	NAME	N	K	NAME
1	1	SHL-6	30	6	BOL-CRU	58	1	SHL-6
2	1	"	31	6	"	59	1	"
3	1	"	32	6	"	60	1	"
4	1	"	33	7	G-6,S Fe	61	12	IRA-CRU
5	1	"	34	7	"	62	12	"
6	1	"	35	7	"	63	12	"
7	1	"	36	8	Sts-CRU	64	3	TOB-CRU
8	1	"	37	8	"	65	3	"
9	1	"	38	8	"	66	3	"
10	1	"	39	9	Say-6	67	35	TOBL-CRU
11	1	"	40	9	"	68	35	"
12	1	"	41	9	"	69	35	"
13	2	ZEL-CRU	42	10	Apa-CRU	70	35	"
14	2	"	43	10	"	71	14	HAS-CRU
15	2	"	44	10	"	72	14	"
16	2	"	45	11	G-6,PA	73	14	"
17	2	"	46	11	"	76	4	QQ-CRU
18	2	"	47	11	"	77	4	"
20	1	SHL-6	48	3	G-6,CIN	78	4	"
21	3	G-6,CIN	49	3	"	84	3	G-6,CIN
22	3	"	50	3	"	85	5	G-6, PHIL
23	3	"	51	5	G-6 PHIL	86	3	G-6,CIN
24	4	QQ-CRU	52	5	"	87	4	QQ-CRU
25	4	"	53	5	"	88	4	"
26	4	"	54	2	ZEL-CRU	89	4	"
27	5	G-6 PHIL	55	2	"	90	4	"
28	5	"	56	2	"	91	3	G-6,CIN
29	5	"	57	1	SHL-6	92	3	"

N = spectrum number
K = class number

TABLE 5 (Continued)

N	K	NAME	N	K	NAME	N	K	NAME
93	6	BOL-CRU	132	22	SH17D1	170	31	SHOD4-6
94	6	"	133	22	"	171	32	EXO-2
95	6	"	134	22	"	173	32	"
96	6	"	135	23	SH17D4	174	32	"
98	15	Say-2	137	23	"	175	33	EXOD1-2
99	15	"	138	23	"	176	33	"
100	15	"	139	24	AMOD1-4	178	33	"
101	16	Say-LUB	140	24	"	179	34	EXOD4-2
102	16	"	141	24	"	180	34	"
105	5	G-6,Phil	142	24	"	181	34	"
106	5	"	144	25	AMOD4-4			
107	5	"	145	25	"			
108	5	"	146	25	"			
109	5	"	147	26	SH33-6			
110	5	"	149	26	"			
113	17	SB-CRU	150	26	"			
114	17	"	151	27	SH33D1-6			
115	17	"	152	27	"			
116	18	SBD1-CRU	154	27	"			
118	18	"	156	28	SH33D4-6			
119	18	"	157	28	"			
120	19	SBD4-CRU	158	28	"			
121	19	"	160	29	SHO-6			
123	19	"	161	29	"			
124	20	SH17	162	29	"			
126	20	"	164	30	SHOD1-6			
127	20	"	165	30	"			
128	21	AMO-4	166	30	"			
129	21	"	168	31	SHOD4-6			
131	21	"	169	31	"			

N = spectrum number

K = class number

TABLE 6 Percentages of Correct (Left Entry) and Incorrect (Right Entry) Classifications of Quadrupole Spectra of 6 Different Oils. Total Sample Size = 154.

REF	f \ J-R	8	7	6	5	4	3	2
1	0.10	8, 0	12, 0	25, 0	25, 2	25, 5	44, 6	69, 8
	0.12	12, 0	19, 0	25, 1	25, 5	25, 5	69, 6	69, 9
	0.14	12, 0	25, 0	25, 4	25, 5	44, 6	81, 6	100, 10
	0.16	19, 0	25, 1	25, 4	31, 6	56, 7	87, 7	100, 13
13	0.10	12, 1	25, 2	37, 8	37, 14	37, 24	37, 33	50, 51
	0.12	12, 1	25, 3	37, 10	37, 18	37, 30	50, 40	75, 57
	0.14	12, 3	25, 6	37, 12	37, 23	37, 36	62, 49	87, 61
	0.16	25, 7	37, 13	37, 20	37, 29	50, 39	62, 55	100, 66
21	0.10	0, 0	11, 0	11, 0	11, 0	22, 0	56, 0	78, 2
	0.12	0, 0	11, 0	11, 0	11, 0	22, 0	67, 0	78, 3
	0.14	0, 0	11, 0	11, 0	22, 0	33, 0	78, 1	78, 6
	0.16	0, 0	11, 0	11, 0	22, 0	44, 0	89, 1	100, 8
24	0.10	22, 0	78, 0	100, 1	100, 3	100, 17	100, 37	100, 55
	0.12	22, 0	100, 1	100, 4	100, 8	100, 26	100, 49	100, 62
	0.14	89, 0	100, 1	100, 6	100, 15	100, 38	100, 52	100, 66
	0.16	89, 0	100, 2	100, 8	100, 19	100, 43	100, 53	100, 69
27	0.10	0, 0	0, 0	33, 0	50, 0	75, 1	100, 4	100, 11
	0.12	0, 0	17, 0	33, 0	58, 0	83, 4	100, 8	100, 20
	0.14	17, 0	25, 0	50, 1	75, 4	83, 7	100, 11	100, 23
	0.16	17, 0	33 0	58, 1	83, 6	92, 7	100, 15	100, 25
30	0.10	33, 3	83, 5	100, 10	100, 17	100, 28	100, 43	100, 54
	0.12	33, 3	100, 5	100, 14	100, 22	100, 35	100, 48	100, 59
	0.14	33, 4	100, 7	100, 18	100, 27	100, 36	100, 50	100, 59
	0.16	50, 6	100, 13	100, 23	100, 29	100, 39	100, 54	100, 64

TABLE 6 (Continued)

Ref. No.	Class	Size	Name
1	1	17	Shell No. 6
13	2	9	Zelten Crude
21	3	10	Gulf No. 6 Cencen
24	4	10	Quirequire Crude
27	5	13	Gulf No. 6 Phila.
30	6	7	Bolivar Crude

In general, table 6 shows that for five of the six classes there are pairs of parameters that enable all of the members of a specified class to be identified correctly while only 10% or less for the remaining spectra are identified incorrectly. However, to avoid using large training sets for each class it is necessary to demonstrate the existence of a single pair of parameters that can be used successfully for the identification of most of the classes.

The entries in table 7 are the result of averaging the respective entries in table 6 over the six classes. Only one pair of parameters, $J - R = 2$, $f = 0.16$, produces an identification criterion that results in all members of the six classes being classified correctly. Unfortunately, an average of 41% of the remaining spectra are also identified with the same six classes. From a forensic point of view this is equivalent to arresting all of the culprits and 41% of the bystanders. If an average of 10% incorrect identification is arbitrarily taken as acceptable, table 6 shows that approximately half of the valid identifications will go undetected (viz. the entry at $J - R = 5$, $f = 0.12$). It appears, therefore, that on the basis of the 154 quadrupole spectra we must conclude that either large numbers of spectra per oil are necessary or that it is necessary to characterize spectra with more detail than is possible with an 8-dimensional vector representation of the cumulative spectral distribution.

TABLE 7 Percentages of Correct (Left Entry) and Incorrect (Right Entry) Classifications Averaged Over the 6 Classes of Table 6

J-R	8	7	6	5	4	3	2
f							
0.10	12, 1	35, 1	51, 3	54, 6	60, 12	73, 20	83, 30
0.12	13, 1	41, 2	51, 5	55, 9	61, 17	81, 25	92, 35
0.14	27, 1	48, 2	54, 7	60, 12	66, 20	87, 28	94, 37
0.16	33, 2	51, 5	55, 9	62, 15	74, 23	90, 31	100, 41

60^{o} Sector Magnet Data

Twenty-nine oil specimens, representing six oil spill cases, were supplied to us by the U.S. Coast Guard. Two of these specimens were mass analyzed in duplicate, one was analyzed thirteen times. The digitized data for the specimen with code name RDC 17 was destroyed by an accident in the laboratory and one of the thirteen replicate spectra (tag #31) for the oil with code name RDC 26 was inadvertently omitted from some of the computer analyses. Table 8 shows the Coast Guard's RDC code numbers, the tag numbers used by our computer for spectrum identification and the case numbers for the 42 spectra comprising this data set.

The thirteen replicate spectra are shown in Figure 16. These measurements were performed before the completion of some of the electronic circuitry for the 60 degree sector magnet spectrometer and with a prototype field ionization source that enabled us to obtain spectra without the usual specimen preparation by vacuum distillation. Due to the developmental stage of the spectrometer, we obtained the following results: there are nonlinearities in the mass scale and there are systematic errors that can be correlated with mass number or with volatility. The latter problem is a consequence of using manual control on the solid insertion probe heater in which the oil was placed so that the sample feed programming was not constant over the thirteen analyses. The systematic error is manifested by the variability in the envelopes of the spectra; some are fairly flat, others tend to peak at 400 amu. The number of peaks detected by computer algorithm

TABLE 8 Identification Code (RDC) Numbers and Spectrum Tag Numbers for 42 Sector Magnet Spectra Comprising 28 Oil Specimens

CASE #	RDC #	TAG #
43	1	34
	2	15
	3	39,44
	4	19
	5	4,12
66	6	33
	7	36
	8	38
	9	26
	10	23
	11	13
68	12	11
	13	32
	14	20
	15	22
	16	42
77	17	-
	18	46
	19	6
	20	24
	21	7
	22	43
	23	47
80	24	30
	25	16
82	26	9,14,17,21,25,28,31,35,41,45,48,49,50
	27	37
	28	27
	29	10

CASE 82: SPECTRA SHOWING 13 REPLICATE MASS ANALYSES OF CODE NUMBER 26 OIL

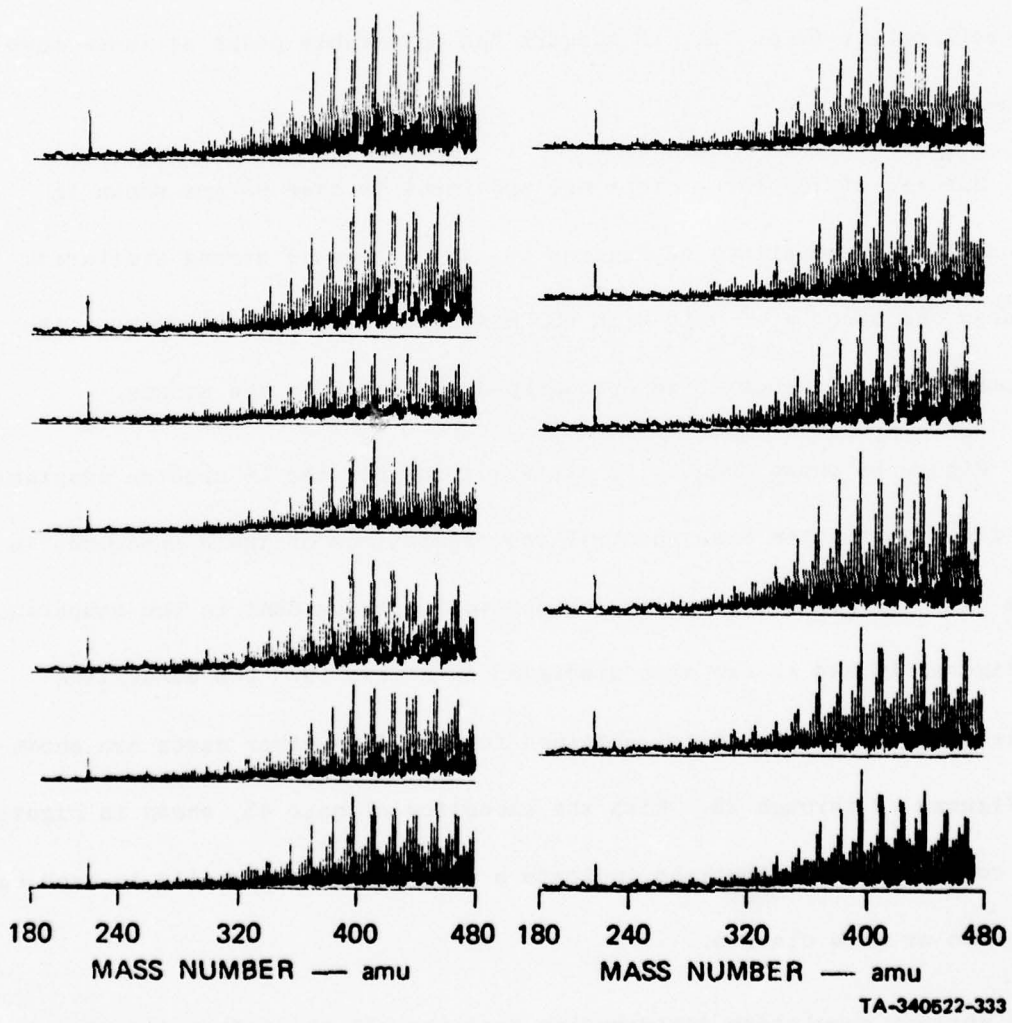


FIGURE 16

varied from 117 to 172 peaks per spectrum with an average of 145. The number of ions counted per peak for the subset of peaks used in this report was of the order of 10^2. Therefore, an average relative standard deviation of the order of 10% would be expected from ion counting statistics alone. The actual standard deviation, computed after each spectrum was normalized to unit total area, was 16%, averaged over the 38 mass numbers selected for reliability (i.e., all 13 spectra had detectable peaks at these mass numbers).

Spectra of the three other oil specimens in case 82 are shown in Figure 17. A comparison of Figures 16 and 17 shows a strong similarity between the spectra of oils with RDC numbers 26, 28 and 29, whereas the spectrum of oil number 27 is obviously different from the others.

Figure 18 shows cumulative distributions for the 16 spectra comprising the field ionization mass spectral representations of the 4 specimens in case 82. The similarities and the dissimilarity evident in the comparison of Figures 16 and 17 are also displayed in Figure 18. The cumulative distributions of the spectra obtained for the five other cases are shown in Figures 19 through 23. With the exception of case 43, shown in Figure 19, the cumulative distributions indicate a separation of the oils in each case into two or more classes.

While a cumulative distribution contains all the information inherent in the original spectrum from which it was derived, its appearance emphasizes the gross characteristics of the molecular ion mass distribution.

CASE 82: MOLECULAR ION MASS DISTRIBUTIONS OF THREE DIFFERENT OIL SPECIMENS

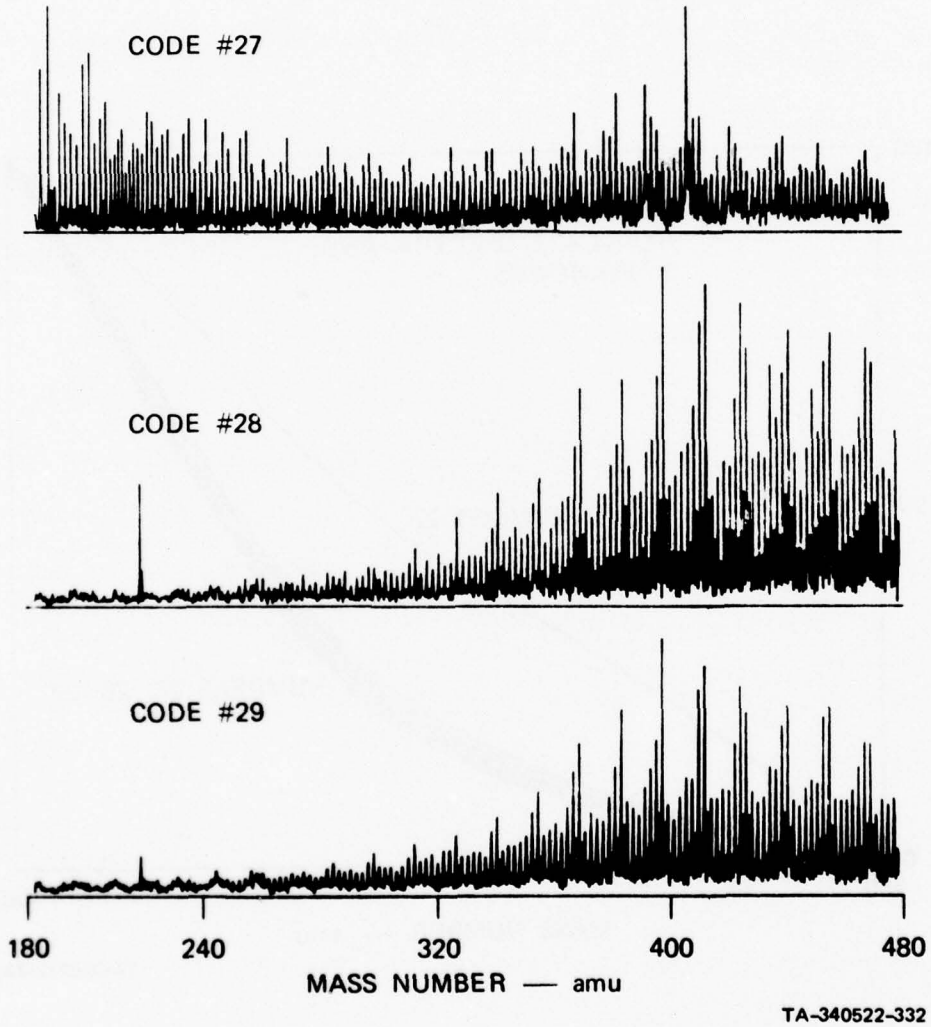


FIGURE 17

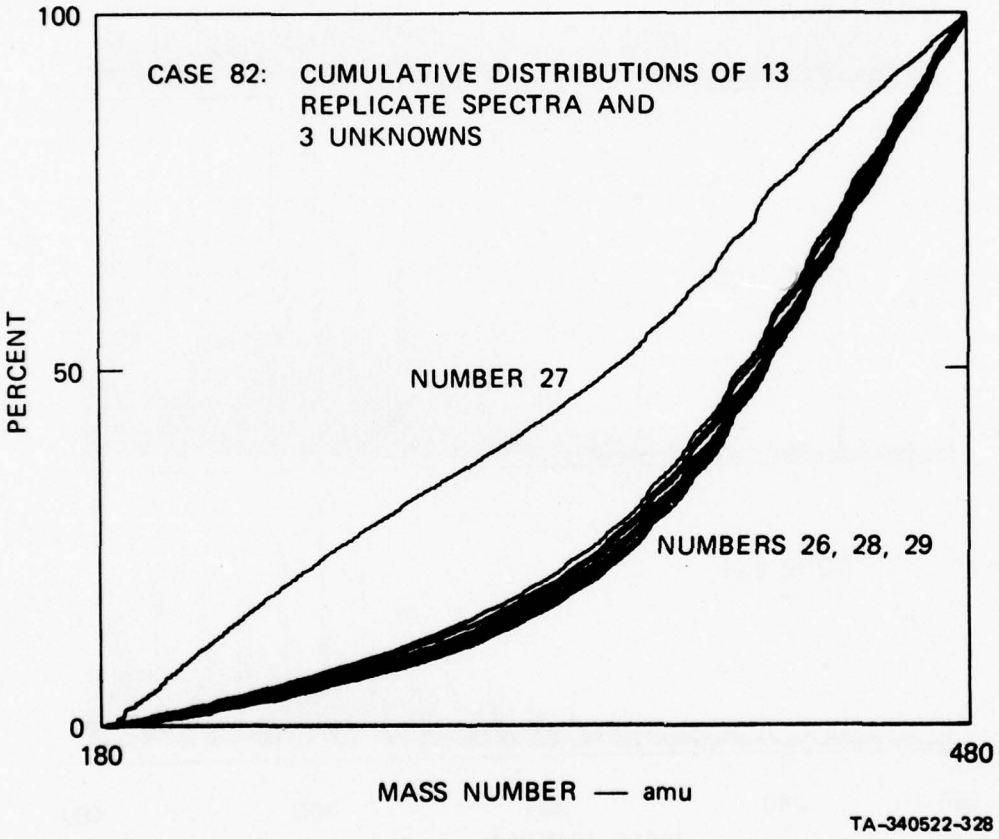

100
CASE 82: CUMULATIVE DISTRIBUTIONS OF 13
REPLICATE SPECTRA AND
3 UNKNOWNS
PERCENT
50
NUMBER 27
NUMBERS 26, 28, 29
0
180
480
MASS NUMBER — amu
TA-340522-328

FIGURE 18

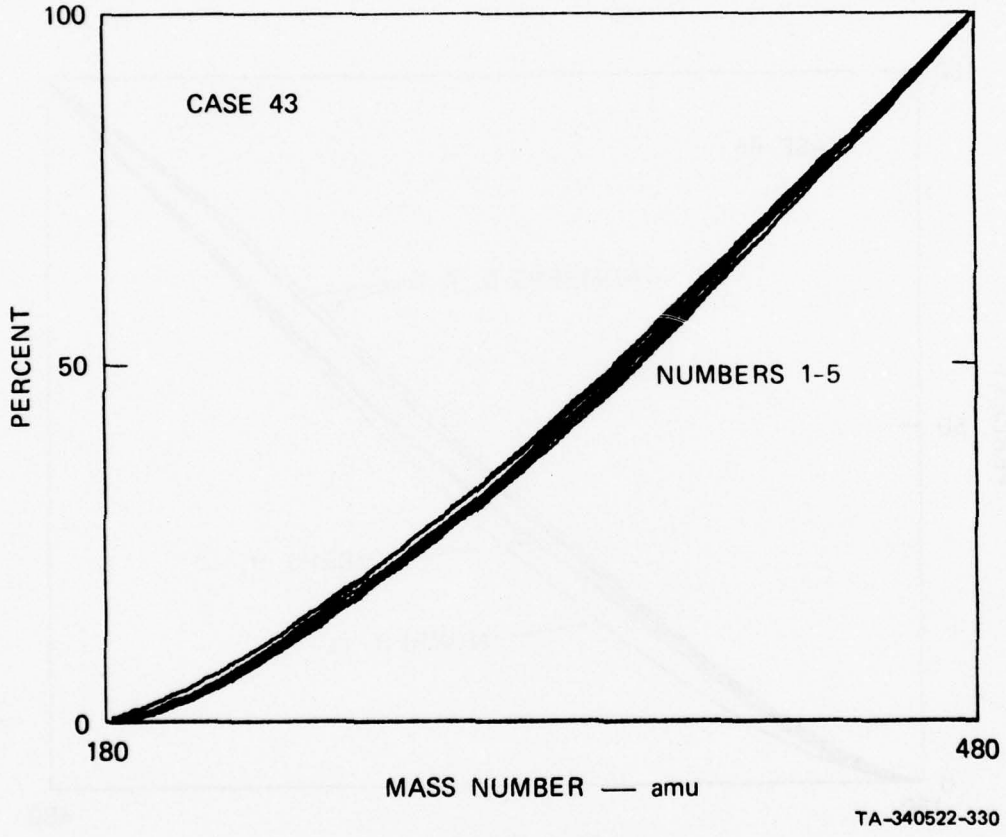


FIGURE 19 CUMMULATIVE DISTRIBUTIONS OF 7 SPECTRA COMPRISING 5 OIL SPECIMENS COMPRISING CASE 43

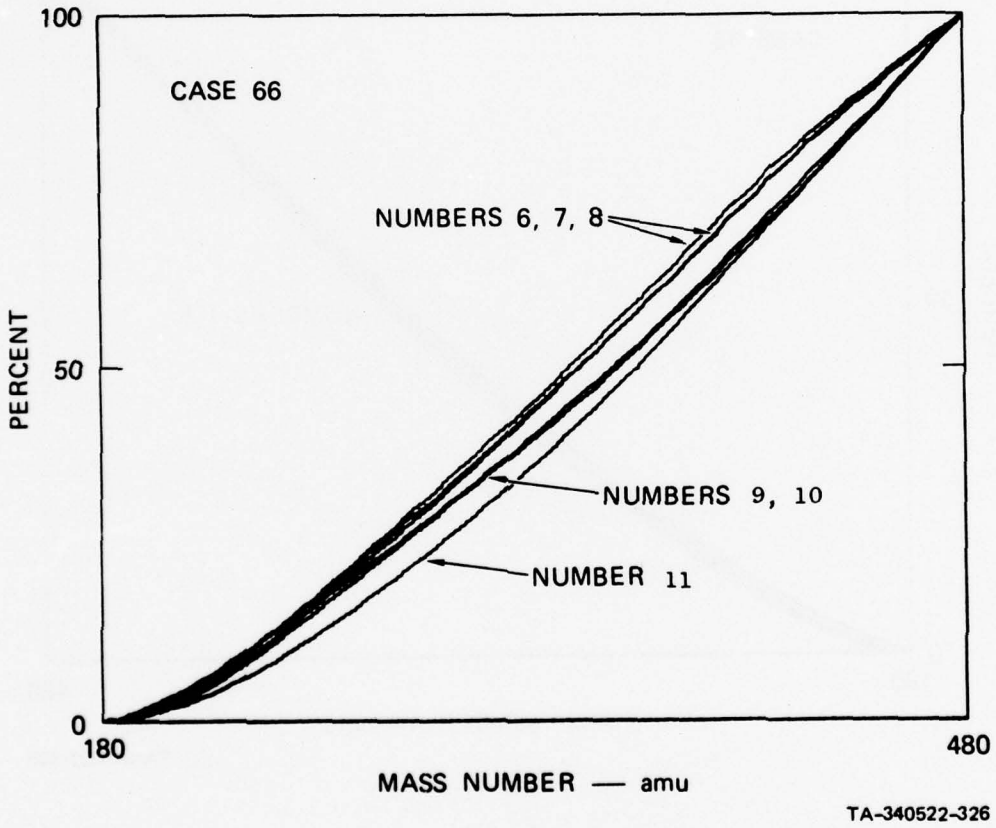


FIGURE 20 CUMMULATIVE DISTRIBUTIONS FOR 6 OIL SPECIMENS COMPRISING CASE 66

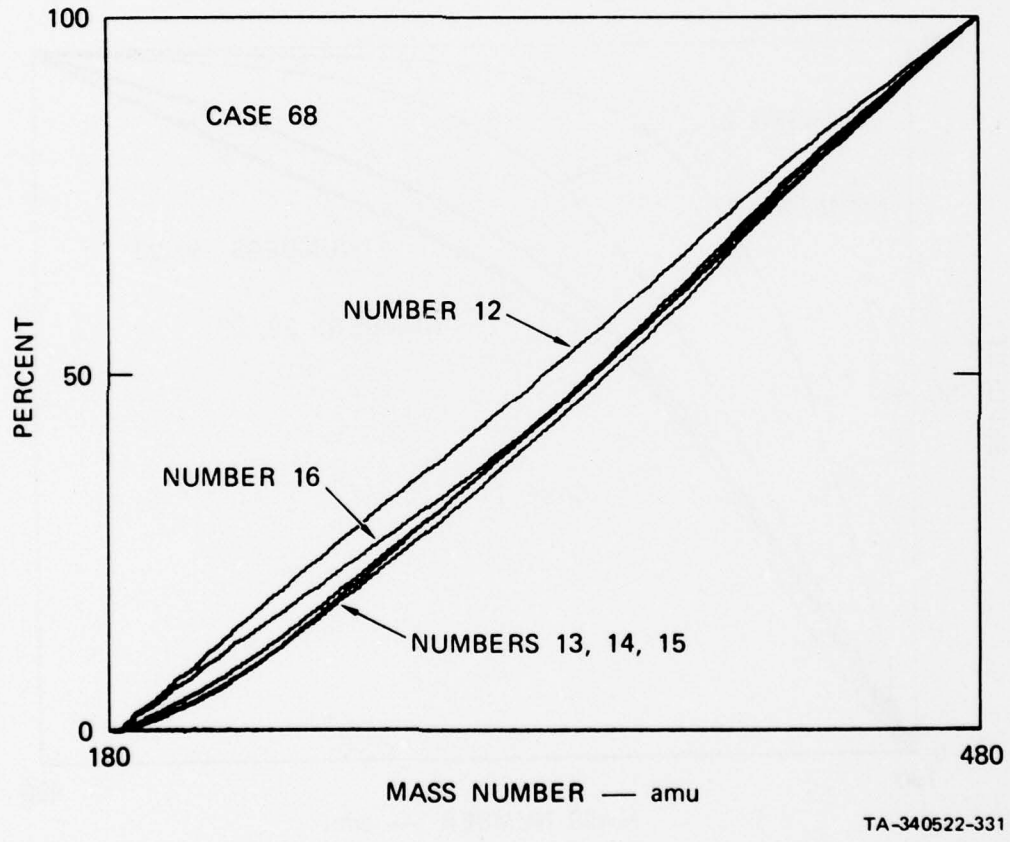


FIGURE 21 CUMMULATIVE DISTRIBUTIONS FOR 5 OIL SPECIMENS COMPRISING CASE 68

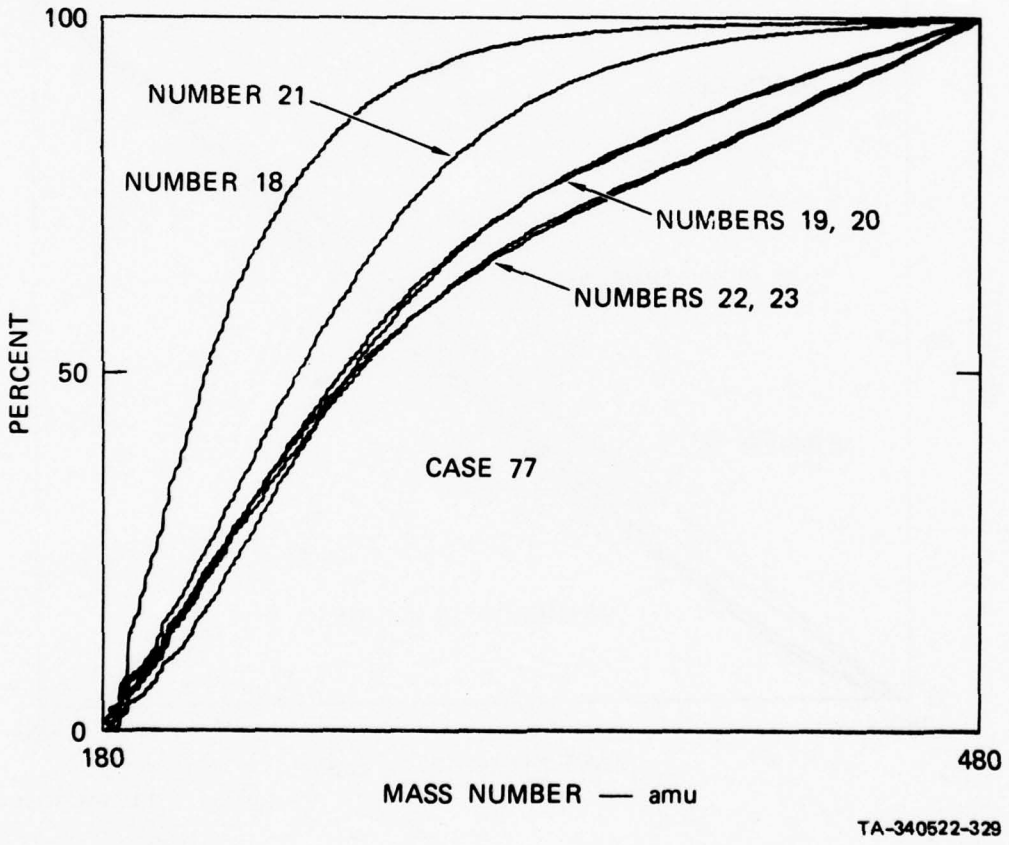


FIGURE 22 CUMMULATIVE DISTRIBUTIONS FOR
6 OIL SPECIMENS COMPRISING CASE 77

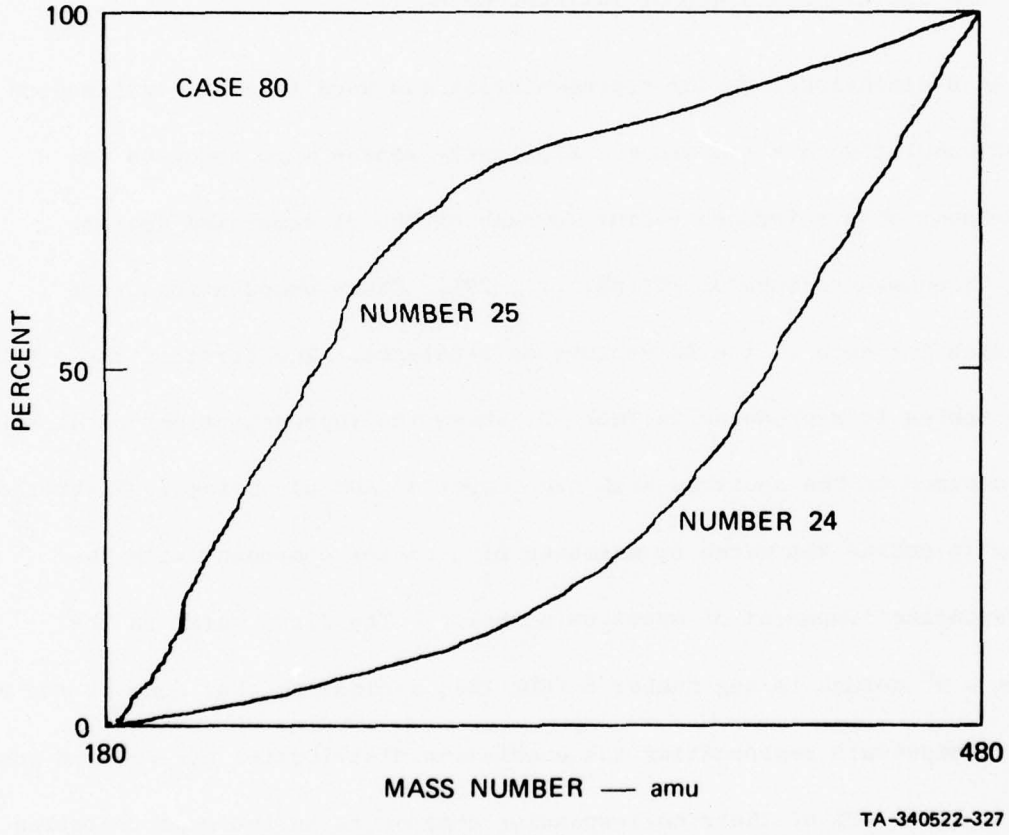


FIGURE 23 CUMMULATIVE DISTRIBUTIONS FOR 2 OILS COMPRISING CASE 80

The two distributions in Figure 23, for example, are obviously different - that is, the separation between the two curves is large in comparison to the range of the replicate analyses of RDC 26 oil in Figure 18. Furthermore, the RDC 25 curve in Figure 23 has its fifty percentile point at 252 amu whereas number 24 oil crosses the fifty percentile point at 407 amu, implying that the former is a relatively "light weight" refined oil while the latter is a crude or residual product.

An 8-dimensional vector representation was used to characterize each of the 42 cumulative distributions. Similarity scores were computed for comparisons of a reference vector to each of the 41 remaining vectors, using threshold factors of 2%, 4%, ..., 20%. These computations were performed for each of the 42 vectors as reference. The first of these 420 score tables is reproduced in Table 9, where the representations of 41 spectra are compared to the spectrum with tag number 4 (RDC 5), using a 2% threshold factor to define the match or mismatch of a vector component with the corresponding component of spectrum number 4. The first entry in the "score = 0" column is tag number 6 (RDC 19), indicating that none of the 8 vector components representing the cumulative distribution of spectrum number 6 fell within $\pm$2% of their corresponding components in the representation of spectrum number 4. Since RDC 5 and RDC 19 are the code names for oils from different spill cases (see Table 8), it is reasonable to assume that these are different oils so that the lack of similarity is not surprising. Reference to Table 8 shows that RDC 5 was mass analyzed in duplicate (viz

Table 9 Similarity Data for 43 Sector Magnet Oil Spectra Comprising 28 Specimens

REF TAG= 4
THRESHOLD= 2%

SCORE= 8	7	6	5	4	3	2	1	0
0	0	0	0	0	0	0	0	6
0	0	0	0	0	0	0	0	7
0	0	0	0	0	0	0	0	9
0	0	0	0	0	0	0	10	0
0	0	0	0	0	0	0	0	11
0	0	0	0	0	0	0	12	0
0	0	0	0	0	0	0	13	0
0	0	0	0	0	0	0	14	0
0	0	0	0	0	15	0	0	0
0	0	0	0	0	0	0	0	16
0	0	0	0	0	0	0	0	17
0	0	0	0	0	19	0	0	0
0	0	0	0	0	0	20	0	0
0	0	0	0	0	0	0	0	21
0	0	0	0	0	0	0	22	0
0	0	0	0	0	0	0	0	23
0	0	0	0	0	0	0	0	24
0	0	0	0	0	0	0	0	25
0	0	0	0	0	26	0	0	0
0	0	0	0	0	0	0	0	27
0	0	0	0	0	0	0	0	28
0	0	0	0	0	0	0	0	30
0	0	0	0	0	0	0	31	0
0	0	0	0	0	0	32	0	0
0	0	0	0	0	0	33	0	0
0	0	0	0	0	0	0	0	34
0	0	0	0	0	0	0	0	35
0	0	0	0	0	0	0	36	0
0	0	0	0	0	0	0	0	37
0	0	0	0	0	0	0	38	0
0	0	0	0	0	0	39	0	0
0	0	0	0	0	0	0	0	41
0	0	0	0	0	0	42	0	0
0	0	0	0	0	0	0	0	43
0	0	0	0	0	0	44	0	0
0	0	0	0	0	0	0	0	45
0	0	0	0	0	0	0	46	0
0	0	0	0	0	0	0	0	47
0	0	0	0	0	0	0	0	48
0	0	0	0	0	0	0	0	49
0	0	0	0	0	0	0	0	50

tag numbers 4 and 12), Table 9 shows that tag numbers 12 got a similarity score of 1 while tag numbers 15 (RDC 2) and tag 19 (RDC 4) got scores of 3 each. Tag numbers 4, 12, 15 and 19 all correspond to case number 43. The cumulative distributions shown in Figure 19 indicate that the spectra in this case were indistinguishable within the present constraint of a 16% standard deviation.

The 420 similarity score tables were reduced to "case" tables. For example, Table 9 reduces to Table 10 where the comparisons to spectrum number 4 are limited to the other spectra in case 43. In general, the similarity score for comparing spectrum "X" to reference "Y" is different from the score obtained when spectrum "Y" is compared to reference "X". The two scores are not independent, however. The analysis of 12 replicate spectra of RDC 26 oil (using a threshold factor of 8%) resulted in 12 x 11 = 132 similarity scores, half of which correspond to "X compared to reference Y" while the other half correspond to "Y compared to reference X". The correlation coefficient obtained by analyzing the 66 pairs of scores was 0.88. It was decided, therefore, to use the average of each pair. For example, tag 12 got a similarity score of 1, when compared to tag 4 as reference and tag 4 got a similarity score of 3 when compared to tag 12 as reference. Therefore, a score of $(3 + 1)/2 = 2$ was used for the similarity in a comparison of the spectra with tag numbers 4 and 12, using a 2% threshold factor.

Table 10 Case 43: Similarity Scores for 6 Oil Spectra Compared to Spectrum Number 4 (RDC 5) Entries = Tag Nos.

REF TAG = 4

THRESHOLD = 2%

SCORE =	8	7	5	4	3	2	1	0
					.	.	12	.
					15	.	.	.
					19	.	.	.
					.	.	.	34
					.	39	.	.
					.	44	.	.

Table 11 shows similarity scores for the 7 spectra in case 43 using threshold factors of 8% and 4%. The 8% threshold data show that when the duplicate analyses of RDC 3 oil were compared, a similarity score of 8 (i.e., maximum similarity) was obtained. The duplicate analyses of RDC 5 also resulted in a high score of 7. On the other hand, when one of these spectral representations, RDC 5(2), was compared to the representation of any other oil in case 43, a score of 8 was obtained. When the threshold factor was reduced to 4% so that the requirements for a spectral match were more stringent, Table 11 shows that the duplicate spectra of RDC 5 were less similar than RDC 5(2) was to any of the other spectra in case 43. On the basis of these results, we cannot distinguish between any of the specimens in case 43. This result is consistent with the undistinguishability of the cumulative distributions shown in Figure 19. The use of other threshold factors resulted in the same conclusion.

In an attempt to determine the significance of differences in similarity scores, 12 replicate analyses of RDC 26 were analyzed. Table 12 shows the similarity scores for comparisons between twelve replicates of RDC26 and one analysis each for RDC numbers 27, 28 and 29. The $\bar{X}$ row shows the average of the scores resulting from comparisons with the replicates of RDC26. The entries under the diagonal line were used merely to facilitate the computation of the average scores, otherwise they are redundant. According to these results, if we assume that the individual scores for comparisons of replicate

Table 11 Case 43: Similarity Scores for Comparisons between 7 Spectra Representing 5 Oil Specimens

Threshold = 8%

RDC #	2	3(1)	3(2)	4	5(1)	5(2)
1	8.0	8.0	7.5	8.0	7.0	8.0
2		8.0	7.0	7.5	6.0	8.0
3(1)			8.0	8.0	6.5	8.0
3(2)				6.5	5.5	8.0
4					6.0	8.0
5(1)						7.0

Threshold = 4%

RDC #	2	3(1)	3(2)	4	5(1)	5(2)
1	4.0	8.0	5.0	6.0	5.0	7.0
2		7.5	5.0	6.0	5.0	5.0
3(1)			7.0	5.0	3.0	8.0
3(2)				3.0	3.0	7.0
4					4.0	5.0
5(1)						4.0

Table 12 Case 82: Similarity data for 12 replicate analyses of RDC 26 oil and for single analyses of oils with code numbers 27, 28 and 29. Dimensionality = 8, threshold = 8%.

RDC #	26(1)	26(2)	26(3)	26(4)	26(5)	26(6)	26(7)	26(8)	26(9)	26(10)	26(11)	26(12)	27	28	29
26(1)		6.5	3.0	4.5	3.0	2.0	3.5	3.0	3.5	3.5	2.5	2.5	0	3.0	5.5
26(2)	6.5		5.5	8.0	8.0	6.0	6.0	7.0	6.5	5.0	6.0	8.0	0	7.0	7.5
26(3)	3.0	5.5		7.0	7.0	7.0	6.0	7.5	8.0	3.0	7.0	8.0	0	7.0	1.5
26(4)	4.5	8.0	7.0		8.0	8.0	8.0	7.0	6.5	4.0	7.0	8.0	0	7.0	4.5
26(5)	3.0	8.0	7.0	8.0		8.0	8.0	7.0	6.5	4.5	7.0	8.0	0	7.0	3.0
26(6)	2.0	6.0	7.0	8.0	8.0		7.5	7.5	7.0	4.0	8.0	8.0	0	8.0	2.5
26(7)	3.5	6.0	6.0	8.0	8.0	7.5		6.0	8.0	2.0	7.0	8.0	0	8.0	2.5
26(8)	3.0	7.0	7.5	7.0	7.0	7.5	6.0		7.0	3.5	5.0	7.0	0	5.0	3.0
26(9)	3.5	6.5	8.0	6.5	6.5	7.0	8.0	7.0		3.0	8.0	8.0	0	7.0	3.0
26(10)	3.5	5.0	3.0	4.0	4.5	4.0	2.0	3.5	3.0		3.0	4.0	0	2.5	7.0
26(11)	2.5	6.0	7.0	7.0	7.0	8.0	7.0	5.0	8.0	3.0		7.0	0	8.0	2.0
26(12)	2.5	8.0	8.0	8.0	8.0	8.0	8.0	7.0	8.0	4.0	7.0		0	7.0	3.0
$\bar{X}$	3.41	6.59	6.27	6.50	6.82	6.64	6.36	6.14	6.54	3.59	6.14	6.95	0	6.37	3.75
27														0	0
28															2
29															

RDC 26: Grand average = 6.00 (n = 66)

$S^2 = 3.81$

$S^2_{\bar{X}} = 1.42$ (n = 11)

spectra are distributed normally with mean = 6.00 and variance 3.81, approximately 70% of the similarity scores will fall in the range $4 \leq X \leq 8$ when two spectra represent the same oil. If we use the average of eleven scores, 90% of the average scores will fall in the range $4 \leq \bar{X} \leq 8$. As a result of these estimates, the following conclusions concerning case 82 were drawn:

(1) RDC 27 is different from the other oils in case 82.

(2) RDC 28 can be identified with RDC 26, as the score $\bar{X} = 6.37$, computed from comparisons with all 12 spectra of RDC 26, agrees well with the grand average score of 6.00 for RDC 26 spectra compared to each other.

(3) RDC 29 does not match well with RDC 26. As a null hypothesis, assume that the two specimens are from the same source. Using the student-t statistic

$$t = (\bar{X}_1 - \bar{X}_2) / \quad S_p^2 \, (1/n_1 - 1/n_2),$$

with $\bar{X}_1 = 6.00$

$n_1 = 66$

$\bar{X}_2 = 3.75$

$n_2 = 12$

$S_p^2 = 3.86$

obtains $t = 3.65$ with 76 degrees of freedom. The probability, $P[t > 3.65]$, is 5×10^{-4}, therefore the null hypothesis was rejected. By inference, this result also implies that RDC 29 and RDC 28 are different oils. Furthermore, the direct comparison between these

two oils resulted in a low similarity score of 2 (see Table 12), which is outside the range of the 66 RDC 26 intragroup scores.

The results of comparing the six specimens of case 66 are shown in Table 13. The curves in Figure 20 imply the existence of three different oils represented by the specimens with RDC numbers (6,7,8), (9,10) and 11. The scores in Table 13 reflect the apparent clustering in Figure 20 with high scores for comparisons between the spectra of oils within each cluster and low scores for comparisons between oils in different clusters.

Table 13 Case 66: Similarity Scores for Mass Analyses of 6 Oil Specimens
Dimensionality = 8, threshold = 8%

	7	8	9	10	11
6	8.0	7.0	4.5	4.5	2.0
7		7.0	4.0	4.0	4.0
8			5.0	5.0	3.0
9				8.0	4.0
10					4.5

The scores for the five oils in case 68 are shown in Table 14 where, in spite of the fact that the RDC 12 curve in Figure 21 appears to be well-isolated from the other curves, the comparison between RDC 12 and RDC 14

resulted in a score of 5. This result implies that the RDC 12 and RDC 14 curves are parallel to within ± 8% over one-half of the total mass range (i.e., the 8-dimensional vectors were constructed out of eight of the ten percentile intervals for these curves).

Table 14 Case 68: Similarity Scores for Spectra Representing 5 Oil Specimens Dimensionality = 8, threshold = 8%

	13	14	15	16
12	3.0	5.0	3.5	2.5
13		6.0	8.0	5.0
14			7.0	5.0
15				6.0

The scores for case 77 are shown in Table 15.

Table 15 Case 77: Similarity Scores for Spectra Representing 6 Oil Specimens Dimensionality = 8, threshold = 8%

	19	20	21	22	23
18	0	0	0	0	0
19		5.0	0	0	0
20			0	0	2.0
21				0	0
22					7.0

The similarity score for the two spectra in case 80 was 1, indicating that in spite of the difference between the two curves in Figure 23, these cummulative distributions are parallel to within $\pm$ 8% over one of the ten percentile mass intervals. Judging from the appearance of Figure 23, this interval is in the center of the mass range.

V CONCLUSIONS

Identification criteria can be evaluated in terms of the frequency with which either of two kinds of errors are made: (a) the common identity of a pair of spectra is incorrectly rejected and (b) the common identity of a pair of spectra is incorrectly accepted. The significance of the total difference between two spectra is dependent on (a) the number of possible differences (i.e., the number of peaks in the spectra or the dimensionality of their representations), (b) the definition of a spectral difference (i.e., the threshold values), (c) the number of spectral differences, (d) how well the true spectra are known and (e) the dispersion of the data as measured by the variances or the ranges, etc. The relationship between these five parameters and the two kinds of error that determine the success or failure of an identification criterion can be described precisely by assuming that errors in measurements are random and independent and applying a statistical model. Notwithstanding the fact that these assumptions were invalidated experimentally, the conclusions of the statistical model provided guidelines for constructing an empirical discriminant function. The application of this discriminant function to a set of 8-dimensional vectors, that described only the gross characteristics of oil spectra, resulted in the classification of 154 spectra into 35 classes with an error rate of 5%. The success of the reduced dimensionality representation is probably due to the fact that the envelope of a field ionization mass spectrogram represents the molecular mass distribution in a complex mixture.

The "degree of similarity" is a precise and conceptually simple way to rank spectra according to their similarity to a reference spectrum. This technique can be applied to entire spectra or to their reduced dimensionality representations.

REFERENCES

1. Colutron velocity filter, Model 300-6, Colutron Corp., Boulder, Colo. 80302.

2. L. Wahlin, Nucl. Instrum. Methods 27, 55 (1964).

3. W. H. Aberth, C. A. Spindt, M. E. Scolnick, R. R. Sperry, and M. Anbar, Proc. 6th International Mass Spectrometry Conf., Edinburgh, Scotland, in Advances in Mass Spectrometry, A. R. West, ed., Vol. 6 (Elsevier's Applied Science Publishers, Ltd., Essex, London, 1974) p. 437.

4. Model ND-2400 4096-channel multiscaler analyzer, Nuclear Data, Inc. Golf and Meacham Roads, Schaumburg, Illinois 60172.

5. M. E. Scolnick, W. H. Aberth and M. Anbar, Int. J. Mass Spectrom. Ion Phys. 17, 139 (1975).

6. Extranuclear Laboratories, Inc., P. O. Box 11512, Pittsburgh, Penn. 15238.

7. R. H. Cross, H. L. Brown and M. Anbar, Rev. Sci. Instrum., in press.

8. R. L. Wine, Statistics for Scientists and Engineers, (Prentice-Hall, New York, 1964).

9. T. W. Anderson, An Introduction to Multivariate Statistical Analysis John Wiley and Sons, Inc., New York, 1958.

10. R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, (John Wiley & Sons, New York, 1973).

11. N. J. Nilsson, Learning Machines, (McGraw-Hill, New York, 1965).

12. J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles, (Addison-Wesley Publishing Company; Reading, Mass., 1974).

13. B. R. Kowalski, Anal. Chem. 47, 1152A (1975).

14. P. S. Shoenfeld and J. R. DeVoe, Anal. Chem. 48, 403R (1976).

VI APPENDIX

Description of Oil Identification Computer Programs

Mass spectra in the form of sequences of integers

$$S = \{y(n);\ n = 1,2, \ldots, 4096\} \qquad \text{A1}$$

are recorded on 9-track magnetic tape in the laboratory. For a mass range sweep period of t seconds, the nth integer, y(n), is equal to the number of ions accumulated in a "window" or "channel" of width $\tau/4096$ seconds. The tape is read by a computer (Burroughs B6700) and the data are stored on magnetic disk in a file called RAWDATA. The 9-track tape is retained as an archive.

Individual mass peaks usually span 12 to 28 channels. Electromagnetic interference from sources external to the spectrometer can result in extraordinarily large numbers at random positions in the spectrum. This noise appears as a set of "spikes" in a spectrogram whose signal to noise ratio is uniform elsewhere. The RAWDATA file is read by a program called "SMOOTH" which removes isolatable noise, and smooths the data by a least squares fit technique.[16]

To detect and correct isolatable noise at n=n, the following algorithm is used

$$\text{IF} \quad y(n_1) > B + T\left[y(n_1-1) + y(n_1+1)\right]/2, \qquad \text{A2}$$

$$\text{THEN} \quad y(n_1) = \left[y(n_1-1) + y(n_1+1)\right]/2, \qquad \text{A3}$$

where B is a base, below which no corrections are to be made, and T is a threshold value. For example, for T = 3, y(2000) is defined as a spike if its value is greater than baseline by more than three times the average of its immediate neighbors y(1999) and y(2001). Correction consists of replacing a spike with the average value of its neighbors, as indicated in equation A3.

For data smoothing, we assume that the individual peaks can be described adequately by a polymonial in n of degree 5. The value of $y(n_1)$ in the smoothed spectrum is computed by applying a least squares fit to data points (n,y) over the range $n_1-\Delta \leq n \leq n_1 + \Delta$, where $1 + 2\Delta$ is the number (i.e., the nearest odd integer) of channels per amu. The smoothed spectrum is stored in a file called SMOOTHDATA.

Spectral peaks are detected, peak areas are computed and mass numbers are assigned to the spectral position variable n by the program "PKFIND".

A mass peak is detected in a smoothed spectrum at a $n=n_1$, according to the following criteria:

$$\text{IF} \quad y(n_1) > 2 \times \text{noise} \tag{A4}$$

$$\text{and} \quad \text{IF} \quad y(n_1-1) \leq y(n_1) \geq y(n_1+1) \tag{A5}$$

$$\text{and} \quad \text{IF} \quad y(n_1-w) < Ry(n_1) > y(n_1+w), \; o < R < 1 \tag{A6}$$

Then n, is the position of a peak maximum. "Noise" is defined as either the minimum of the spectrum or some arbitrarily small positive integer (threshold value). Equation A4 eliminates baseline fluctuations as candidates for mass peaks. Equation A5 is a local maximum test to determine whether or not $y(n_1)$ is larger than its two immediate neighbors. Equation A6 is a peak width test. For example, if 4w is the number of channels per amu, 2w is an approximate upper limit for the full width at half maximum of peaks in a well resolved spectrum. Thus, for $R \simeq 0.5$, equation A6 discriminates against fused peaks and unsmoothed "bumps" in the spectrum. After scanning the entire spectrum for peaks and storing the value of n corresponding to the largest peak, the program enters a mass calibration routine.

The numbers of channels between the jth peak and the (j-1)th peak and the jth peak and the (j-2)th peak are computed for j=3,4, ..., NPEAK, where NPEAK is the total number of peaks detected. By visually inspecting a few spectra in the data set, one can compute a range for the number of

channels between adjacent peaks (e.g., 4096 channels spanning 200 amu = 20 to 21 channels per mass unit). The numbers of channels between the jth peak and its neighbors are compared to a range that has been selected according to two criteria.

1. The range must be small enough to discriminate between (a) two peaks that are separated by approximately one mass unit and (b) peaks that are separated by two or more mass units or less than one mass unit (i.e., "bumps" that had enough modulation to evade the screening of equation A6.

2. The range must be large enough to accomodate the variations in the number of channels per amu that are due to systematic non-linearities, signal noise and the limit in resolution imposed by the finite width of a channel.

When the "distance" between two peaks falls in this range, the reciprocal distance (units = amu per channel) and the value of n corresponding to the jth peak are stored. In other words, this algorithm stores data that discribes the slope of a mass number versus channel number calibration curve, as a function of channel number. These data are used to compute a regression curve for the slope of the calibration curve as a quadratic function of n.

At this point in the program, the computer displays the channel number at which the largest or $(jmax)^{th}$ peak occurred and requests a mass number assignment. By inspecting a spectrogram of the data being analyzed and referring to spectrometer calibration data, the user can supply the computer with the proper mass number. Next, the number of channels between the largest peak and the (jmax-1) peak is computed and multiplied by the slope (evaluated at the channel number of the largest peak) to determine the number of mass units that separate the two peaks. The latter number is subtracted from the mass number of the largest peak to obtain a mass assignment for the $(jmax-1)^{th}$ peak. The slope is now re-evaluated at the channel number of the $(jmax-1)^{th}$ peak and the number of mass units to the next peak on the list is computed. The latter number is subtracted from the mass number of the $(jmax-1)^{th}$ peak to obtain the mass number for the $(jmax-2)^{th}$ peak. This procedure is iterated until the mass number of the first peak on the list is computed. The analysis now returns to the $(jmax)^{th}$ peak and procedes iteratively toward the end of the list until the last peak is assigned a mass number.

Peak areas are computed by summing the numbers that lie in the channel range whose boudaries are the minima that lie to the right and to the left of the peak maximum. To avoid integrating the areas of

individual unsmoothed ripples, the algorithm for finding the boundary minima is limited to small ranges that are centered at 1/2 mass unit from the peak maximum.

There are two options for subtracting noise; "noise", as defined above can be multiplied by the number of channels that lie between the two peak minima and subtracted from the peak area or the computer can perform the equivalent of drawing a line tangent to the two minima and subtracting that portion of the peak area that lies below the line.

"PKFIND" writes a file, called PEAKS, that consists of NPEAK + 1 records for each spectrum. Each record consists of two entries. The format is as follows: (TAG, NPEAK), (A_1, m_1), (A_2, m_2), $\cdots$ (A_{NPEAK}, m_{NPEAK}), where TAG is the identification number for the spectrum and NPEAK is the number of peaks in the list, A_j is the area of the jth peak whose mass number is m_j.

Notwithstanding the curve smoothing effect of the least squares fit algorithm and the modulation requirement in the peak detection algorithm, the file PEAKS contains some ambiguous entires (viz more than one peak area with the same mass assignment). A program called "MASSTRAP" detects ambiguous entries in a spectrum and deletes all peak areas with the same mass number except the largest one. The assumption implied by using this program is that the smaller peaks correspond to bumps on the "real" mass

peak. Examination of PEAKS shows that in cases where ambiguity exists there is usually a dominant entry. "MASSTRAP" writes a cleaned-up version of PEAKS called XPEAKS.

The degradation in resolution with increasing mass number that occurred with the Colutron® and quadrupole mass analyzers presented several data analysis problems. Without the peak width test of equation A6 there is no way to distinguish between fused and unfused peaks. For example, the fusion of peaks may exist in the range $m > 280$ in one spectrum of a given oil while in another spectrum of the same oil it may extend over the range $m > 282$. In a comparison of the two spectra the computer would compare a large peak at $m = 280$ in the first spectrum to a small peak in the second one and a peak with zero area at $m = 281$ in the first spectrum to a peak with finite area in the second one. Furthermore, a fused peak has a maximum that lies between the two maxima of its constituents so that the systematic degradation of resolution with increasing mass number results in an apparent nonlinearity of the mass scale.

Oil identification by mass spectrometry is a pattern recognition problem. The assignment of mass numbers to peaks is merely a means of labeling them to facilitate the comparison of spectra. Therefore, an apparent nonlinearity in the mass scale is only an inconvenience. In-

consistency in the resolution of replicate spectra, however, does present a more serious problem. In view of the large numbers of peaks available per spectrum, it was decided to use a peak width test and to limit the characterization of an oil to the mass numbers at which peaks were well resolved in all of the replicate spectra simultaneously.

The computer program "COMMASS" reads the replicate spectra of a given oil from the file XPEAKS, selects the mass numbers at which all of the spectra have peaks and writes a file for that class in which each spectrum contains peaks at the same mass numbers.

A program called "STAT" reads class files (i.e., sets of spectra of a single oil) and performs the following computations and analyses:

(1) Each spectrum in the file is normalized to a total peak area of unity.

(2) A normalized class average spectrum is computed.

(3) The relative error is computed for each mass number of each spectrum

$$e_{ij} = (x_{ij} - \bar{x}_j)/\bar{x}_j, \quad \begin{matrix} i = 1,2 \cdots, \text{NSPEC}, \\ j = 1,2 \cdots, \text{NPEAK}, \end{matrix} \qquad \text{A7}$$

where e_{ij} is the relative error of the ith spectrum evaluated at the jth mass number, x_{ij} is the normalized peak area of the ith spectrum at the jth mass number and $\bar{x}_j$ is the peak area at the jth mass number, averaged over all the spectra in the class.

(4) For each spectrum the computer plots the relative error versus the peak number j.

(5) Graphs generally show errors that monotonically increase or decrease with mass number at rates that are approximately constant. Therefore, regression lines are computed for the relative errors of each spectrum as linear functions of mass number. As an approximation, we assume

$$e_{ij} = A_i + B_i m_j + r_{ij} \qquad \text{A8}$$

where A_i and B_i are constants whose values are computed in the regression analysis and r_{ij} is a random error. Equation A8 represents an attempt to resolve the relative error into systematic and random components.

Combining equations A7 and A8 and solving x_{ij} obtains

$$x_{ij} = \bar{x}_j + \bar{x}_j(A_i + B_i M_j + r_{ij}) \qquad \text{A9}$$

The transformation, $x'_{ij} = x_{ij} - \bar{x}_j(A_i + B_i M_j)$, A10

yields

$$x'_{ij} = \bar{x}_j(1 + r_{ij}) \qquad \text{A11}$$

Therefore, to the extent that equation A8 is a valid approximation, the transformed spectrum whose jth normalized peak area is described by equation A11 differs from the class average spectrum by random errors only.

It should be noted that the normalization requirement for the x'_{ij} implies an additional equation in the regression analysis.

(6) The normalized spectra are transformed according to equation A10.

(7) A covariance matrix is computed for the variations of peak areas about their respective mean values.

(8) The hypothesis, that the variations of peak areas about their respective means are distributed normally, is tested.

(9) A correlation coefficient matrix is computed.

(10) The hypothesis, that the variations in peak area are stochastically independent, is tested.

(11) The computer prints the results of the two tests, the normalized peak areas of the class average spectrum, the standard deviations for peak areas at each mass number in the spectrum and the standard deviation averaged over all peaks in the spectrum.

In general, after the correction for systematic error which accounts for approximately 10% of the average variance, the data appear to be normally distributed but fail to pass the stochastic independence test.

☆ U. S. GOVERNMENT PRINTING OFFICE: 1977--701-610--158