

T-3345

PREDICTION OF LIQUEFACTION
REACTIVITY AND STRUCTURAL ANALYSIS
OF COALS USING PYROLYSIS
MASS SPECTROMETRY WITH
COMPUTERIZED PATTERN RECOGNITION

by Steven L. Durfee

ProQuest Number: 10796310

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10796310

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

T-3345

A thesis submitted to the Faculty and Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Geochemistry).

Golden, Colorado

Date 11/24/86

Signed: Steven L. Durfee
Steven L. Durfee

Approved: Kent J. Voorhees
Dr. Kent J. Voorhees
Advisor

Golden, Colorado

Date: 25 November 1986

George H. Kennedy
Dr. George Kennedy
Head of Department
Chemistry-Geochemistry
Department

ABSTRACT

Coal is opaque, insoluble, non-volatile, inhomogeneous and very complex chemically. Consequently, there are few chemical methods for detailed examination of its chemical structure. One method which works well is pyrolysis followed by mass spectrometry (Py/MS). The resulting spectra are too complicated to be interpreted without the assistance of computerized pattern recognition.

A series of Py/MS experiments was performed on whole coals and flotation-separated sporinite macerals in order to determine their structural properties and to correlate those properties with known characteristics of the coals. The pattern recognition techniques of nonlinear mapping, Fisher weighting, hierarchical clustering and principal components analysis were used to uncover patterns in the data which related to rank and liquefaction reactivity in tubing bombs and the Gulf continuous flow reactor.

These observations indicated that the principal components could be used for linear modeling of reactivity. The multivariate statistical techniques of principal components analysis followed by stepwise

multiple linear regression with cross validation were used to develop linear models which predicted total equilibrium liquefaction reactivity and the yield of specific products from the liquefaction process. The validity of the model was established by treating some of the coal samples as unknowns.

In a later set of experiments, reactivity in the Exxon donor solvent process was also modeled using the same technique; however, the technique was not successful for modeling kinetic reactivity because of the nonlinear nature of this measurement. Karhunen-Loeve projections showed that kinetic reactivity could be qualitatively predicted, however.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	iii
LIST OF FIGURES	viii
LIST OF TABLES	xi
ACKNOWLEDGEMENTS	xii
I. INTRODUCTION	1
A. Description of the Problem	1
B. Statement of Objectives	6
C. Liquefaction Background	8
D. Pyrolysis Background	12
1. Overview	12
2. Two temperature regimes	13
3. Maturation	17
4. Macerals	19
5. Role of Oxygen in Maturation and Liquefaction	22
6. Justification for Peak Assignments in Py/MS	27
7. Py/MS for Prediction of Liquefaction Reactivity	31
E. Pattern Recognition Background	33
1. Overview	33
a. Unsupervised Learning	37
b. Supervised Learning	37
c. A Valid Approach for Coals	39

2.	Data pretreatment	42
	a. Normalization	43
	b. Scaling	49
3.	Pattern Recognition Techniques	54
	a. Feature Selection	54
	b. Unsupervised Learning	61
	c. Supervised Learning	69
II.	RESOURCES	75
	A. Instrumentation	75
	B. Computer Resources	77
	C. Samples	79
III.	UNSUPERVISED METHODS	83
	A. Tubing Bomb Reactivity	83
	B. GCF Reactivity	91
	C. Stirred Bomb Reactivity	100
VI.	SUPERVISED METHODS	105
	A. Experimental Method	106
	1. Samples	106
	2. Collection of Data	108
	B. Prediction and Evaluation	110
	1. Tubing Bomb Reactivity	111
	2. GCF Reactivity	133
	C. Discussion of Results	136
VII.	EXXON COALS	139

A.	Experimental Method	139
1.	Samples	139
2.	Analysis	140
3.	Pattern Recognition	142
B.	Equilibrium and Kinetic Reactivities . .	142
VI.	CONCLUSIONS	168
IX.	REFERENCES	176
	APPENDIX I. DOCUMENTATION FOR PROGRAM SPIN	180
	APPENDIX II. PROCEDURE FOR MOVING DATA BETWEEN PROGRAMS	197

LIST OF FIGURES

1.	Comparison of Conversion Measurements	10
2.	van Krevelen diagram	23
3.	O content vs. rank	24
4.	Early Py/MS - reactivity non-linear map . . .	32
5.	Py-MS of two selected coals	85
6.	Nonlinear map of Py-MS data from tubing bomb coals	87
7.	Hierarchical clustering dendrogram of Py-MS data from tubing bomb coals	89
8.	Nonlinear map of Pennsylvania State University coals	94
9.	Hierarchical clustering dendrogram of Py-MS data from Pennsylvania State University coals . .	96
10.	Bivariate plot of m/z 94 vs. 108	98
11.	Bivariate plot of m/z 83 vs. 85	99
12.	Nonlinear map of Py-MS data of CSM coals . .	102
13.	Accuracy of predicting "unknown" reactivities as coefficients are added through stepwise linear regression	113
14.	Prediction of reactivities with six coefficients	116
15.	Predicted reactivity of group Y2 coals from a model based on group Y1 coals	118
16.	Predicted reactivity of group Y1 coals from a model based on group Y2 coals	119
17.	Correlation of mass spectral peaks with tubing bomb conversion	121

18.	Factor spectrum for tubing bomb conversion .	125
19.	Comparison of predictions for tubing bomb reactor reactivity with and without standard deviation multiplication	127
20.	Factor spectrum for vitrinite reflectance .	128
21.	Comparison of predictions of rank with and without standard deviation multiplication .	129
22.	Correlation between rank and conversion in tubing bombs	131
23.	Factor spectrum for reactivity in the Gulf continuous flow process.	135
24.	Predicted THF rate constant	144
25.	Predicted toluene rate constant	145
26.	Predicted pentane rate constant	146
27.	KL projection for pentane k	149
28.	KL projection for pentane k with rotation .	151
29.	KL projection for toluene k	152
30.	KL projection for toluene k with rotation .	153
31.	KL projection for THF k with rotation . . .	155
32.	Predicted THF conversion (three factor model)	156
33.	Predicted toluene conversion (eight factor model)	157
34.	Predicted pentane conversion (five factor model)	158
35.	Predicted pseudoequilibrium reactivity (five factor model)	159
36.	Factor spectrum for pentane conversion . . .	164

T-3345

37.	Factor spectrum for toluene conversion . . .	165
38.	Factor spectrum for THF conversion	167

LIST OF TABLES

I.	Compounds identified in coal pyrolysis products	28
II.	Compounds identified in the pyrolysis of GC/MS analysis of coal ILLS-23	30
III.	Coal samples available	80
IV.	Summary of coal samples studied in tubing bomb reactor (Group 1 coals)	83
V.	Distance table for various coals studied in tubing bomb reactor (Group 1 coals)	86
VI.	Fisher Weights for Tubing Bomb Coals (Group 3 coals)	91
VII.	Characterization data for Gulf continuous flow reactor coals (Group 4 coals)	92
VIII.	Distance table for the Gulf continuous flow reactor coals (Group 4 coals)	93
IX.	Fisher Weights for Gulf Continuous Flow Reactor (Group 4 coals)	100
X.	Coals Used in Stirred Batch Reactor (Group 2 coals)	101
XI.	Distance Table for the Coals Studied in the Stirred Tank Reactor (Group 2 coals)	103
XII.	Reactivity Predictions	115
XIII.	Regression statistics for most reasonable equation calculated from factor analysis and stepwise linear regression	147
XIV.	Prediction of reactivities for Exxon coals	162

ACKNOWLEDGEMENTS

The research reported in this thesis was supported by several teaching assistantships, a one year fellowship from the Colorado Energy Research Institute (CERI), a one year fellowship from the Colorado School of Mines Mining and Minerals Research Institute, and two one year fellowships from the Colorado Scholars program. In addition, I was supported by a research assistantship from the Department of Energy. Support for the instrument was provided in part by the Gates Foundation. To all of these sources of funding I owe a sincere debt of gratitude. The research would not have been possible without them.

During my years at the Colorado School of Mines, I was privileged to be a part of a stimulating research group. In particular, I wish to acknowledge Rushung Tsao, Steve DeLuca, Jim Hickey and Mike Malley for many fruitful discussions and for maintaining a positive, cooperative approach to the work we accomplished in the mass spectrometry lab. I also wish to acknowledge the capable leadership of Professor Kent J. Voorhees, my thesis advisor. Kent was an advisor in every sense of the word - I am indebted to him for his guidance in

selecting coursework, for directing the course of my research, for supplying and maintaining the laboratory, for prodding me at appropriate times to publish and present papers at conferences, for finding financial support when I needed it, and for his friendship and personal advice.

Finally, I wish to acknowledge the support of my family. My wife Amy was a continual source of advice, encouragement, understanding and enthusiasm. The time and energy I was able to devote to my higher education is reflected equally by sacrifices on her part. My children, Casey, Benjamin and Lilly, my parents and brother Chuck were always understanding of my many commitments, and assisted whenever possible throughout my course of study.

I. INTRODUCTION

A. Description of the Problem

A significant body of literature indicates a considerable effort to understand the chemistry of coal. A substantial portion of that literature deals with thermal reactions which coal or compounds modeling coal undergo. It would seem, from the amount of effort which has been expended on understanding high temperature reactions of coal, that liquefaction reactivity would be well understood and easily predicted. This is not the case. The chief reason for this breach in the understanding of coal is the lack of methods for measuring composition and reactions of coal with detail, reproducibility and interpretability.

For several reasons, much of the research which has been performed is unsatisfactory for directly predicting liquefaction accurately: 1) Measurements of only the products of liquefaction fail to relate directly to the properties of the coal itself because the liquefaction process itself is a complex, poorly controlled and incompletely understood process, and because the distribution of products does not relate directly to the catalytic components which affect

liquefaction yields. 2) Failure to recognize geological and chemical constraints limits the extrapolation of models based on one set of coals to a different set of coals. 3) Representative sampling is difficult to achieve, since coal appears to vary dramatically in composition both horizontally and vertically. 4) If the range of reactivities is not broad enough, the model will have limited applicability to coals which lie outside the range of coals used to develop the model. 5) There is no well-accepted definition of reactivity. 6) There are errors in projecting laboratory scale model processes (such as the tubing bomb or the stirred batch reactor) to industrial-scale processes, since they do not model the process exactly.

Of these problems, the most serious is the lack of an agreed-upon definition of reactivity, since the others can be overcome by proper experimental design, sampling and a suitable instrumental method of analysis. The lack of a unified definition of reactivity is a result of two principal factors. 1) There is a wide range of liquefaction processes, which produce different product slates; therefore, a

predictor of liquefaction reactivity should provide an estimate which is specific for a given process. If it does, then the measurement of the reactivity in that process becomes the working definition of reactivity for the model. An alternative would be a sufficiently detailed reactivity measurement (and corresponding model) to apply to a variety of processes. 2) Within an individual process there are a number of parameters which influence the yield of product. The desired product slate influences the selection of these operational parameters, since the most desirable product slate may not result in the greatest total yield of liquid products. Therefore a liquefaction reactivity model which can be used to predict the yield of specific products under a given set of operating conditions would be more useful than a definition which provides a single number representing all of the products.

There are several alternatives to the acceptance of a single predictive equation:

1. Find a feature or set of features which provide a measurement which is more universal than the simple measurement of the yield of products, such

as the kinetic reactivity measurement proposed by Furlong et al. (1982).

2. Adopt a different predictive equation for each process, or even for each set of operating conditions within a given process modeled.
3. Generate a model which produces an intermediate set of parameters. This intermediate set of parameters defines reactivity, but separate predictive equations relate the liquefaction parameters to the actual process modeled, giving a single prediction of reactivity specific to each process.

The approaches used in this thesis appear quite promising because they are not limited to a restricted definition of reactivity or a specific process. In these methods, the structure of the coal is examined in greater detail than other direct chemical techniques can provide, and those aspects of the structure which are related to a given process or liquefaction definition are incorporated into the model. In the process of developing the best predictive model, much information which is irrelevant to predicting liquefaction reactivity is discarded.

Reactivity in large scale processes is typically predicted from models based on reactivity in small scale or pilot processes, such as the tubing bomb (Yarzac et al., 1980). Error because of scaling differences between the laboratory and the industrial processes will be combined with analytical uncertainty, reducing the accuracy of the prediction. The analytical/mathematical approaches described in this thesis permit a direct prediction of reactivity on an industrial scale by means of a quick, cost effective chemical test. Furthermore, a separate model may be developed for different processes using these approaches, since reactivity is not a single generalizable constant. Finally, the methods described in this thesis provide structural details which reflect the chemical differences responsible for the reactivity differences observed in different processes with an equivalent coal.

There are many potential uses for the reactivity information obtained. It could be used to provide quick tests for purposes of purchasing and blending of coal and adjusting the liquefaction process. Improved homogeneous catalysts might be suggested by better

understanding the nature of components correlated with high reactivity and the relationship between structure and reactivity. This technique can be used to develop a process for a given area. For instance, an optimum process for low-sulfur Cretaceous Western Region coals would probably be quite different from a process for Carboniferous, high-sulfur Interior Region coals. The process will be demonstrated to be relatively insensitive to the effects of weathering. Therefore, elaborate precautions during collection, storage and analysis do not appear to be necessary. This rapid, cost-effective predictive method can be used to estimate the variability in yield from potential feedstocks.

B. Statement of Objectives

As the title of this thesis implies, there are two general objectives for these experiments which are connected by the data analysis. The chief objective is to develop separate predictive models for liquefaction reactivity in a variety of industrial- and laboratory-scale processes according to several definitions of reactivity. The supplementary objective is to obtain

as much liquefaction specific structural information as possible by combining an understanding of the predictive models and information about coal pyrolysis products which has been reported in the literature.

The experiments and data analysis methods were dictated to some extent by the samples available and experience with the pattern recognition methods used. The following specific objectives will be attained in order.

1. Determine whether the yield of liquefaction products in tubing bombs is related to information in the pyrolysis mass spectra by using unsupervised pattern recognition.
2. Determine whether information related to the yield of liquefaction products in the Gulf Continuous Flow process is revealed by the same methods.
3. Determine whether the yield of liquefaction products in the stirred batch reactor is represented in pyrolysis mass spectra using unsupervised pattern recognition.
4. Develop an accurate predictive model, establish its validity and predict the reactivity for a set of coals with unknown reactivity using a large data

set consisting of 26 coals with known tubing bomb reactivity.

5. Determine the Py/MS peaks which are correlated with reactivity in tubing bombs.
6. Develop a predictive model for reactivity in the Gulf Continuous Flow (GCF) process using a data set of 11 coals with eight replicates each. Make predictions for unknowns.
7. Determine the Py/MS peaks which are correlated with reactivity in the GCF process.
8. Develop an equation to predict the reactivity of coals in the Exxon Donor Solvent (EDS) process.
9. Develop separate models for a variety of liquefaction reactivity definitions in order to predict the yield of specific products and define structural differences related to yield of each product for the EDS process.

C. Liquefaction Background

In the last decade, researchers at Pennsylvania State University have expended a considerable amount of effort to investigate the performance of coals from the Pennsylvania State coal bank (PSOC coals), in the GCF

reactor (Given et al., 1979) and in tubing bombs (Shadle and Given, 1982; M. Abdel- Baset et al., 1978; Yarzab et al., 1980). At Exxon, Neavel has investigated the performance of a different suite of coals in the EDS liquefaction process (Neavel, 1981; Neavel et al., 1981).

The measurement of liquefaction products directly from the industrial-scale process is expensive and time-consuming. Therefore a tubing bomb reactor (an easily-constructed device for laboratory-scale liquefaction) was developed at Pennsylvania State University. The bomb has made it possible to process a large number of coals, to look at regional and compositional differences, and to investigate factors affecting liquefaction over a broad range of reactivities because of its small size and ease of use. The correspondence between tubing bomb reactivities and GCF reactivities is reasonably good, as data from Yarzab et al., (1980) indicate (Figure 1), but the correspondence between tubing bomb reactivities and other full-scale liquefaction processes has not been established. Also, the procedure is time-consuming compared to analytical pyrolysis, and the tubing bomb

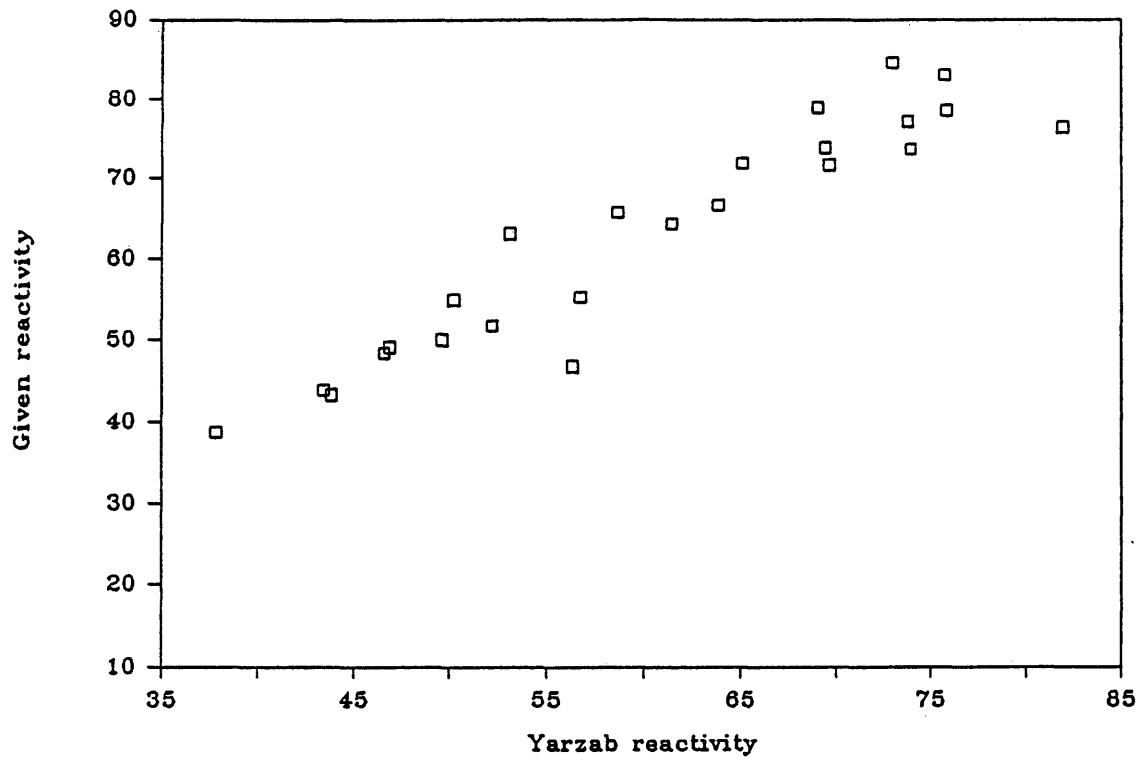


Figure 1. Comparison of Conversion Measurements

reactivity indicates nothing about the individual products of liquefaction.

Plots of conversion in a tubing bomb reactor study (M. Abdel-Baset et al., 1978) against dry and mineral matter free (dmmf) carbon content, percent S and total reactive macerals (TRM) indicated that there are pronounced differences between Eastern, Interior and Western province coals. TRM is defined as the sum of the amounts of vitrinite, pseudovitrinite and liptinite. The equation $\text{percent conversion} = 3.4 (\text{total S}) + 0.31 (\text{TRM}) - 0.66 (\text{percent C}) + 91.9$ was derived, which explained 93.8 percent of the variance.

Given et al. (1979) investigated the liquefaction performance of 104 coals in the GCF process. They also recorded a wide variety of physical and chemical data for these coals, including the yield of specific products after liquefaction and the effect of maceral distribution on some of the coals' behavior. They concluded that rank alone was insufficient to predict reactivity in the GCF process. After a principal component analysis using 15 chemical and physical properties of the coals, three discrete groups of coals were apparent from the data analysis. These groups

could be identified more or less with the Eastern, Interior and Western geographic provinces. There were, however, a few exceptional coals which were assigned to groups dominated by members of a different province.

Yarzab et al. (1980) measured the conversion of a different set of 104 PSOC coals in the tubing bomb reactor. As in the previous study, three groups, corresponding somewhat with geographic province, were indicated by factor analysis. For groups 1 (predominantly Eastern) and 2 (predominantly Interior) separate regression equations were developed which included terms for vitrinite reflectance, H/C, vitrinite content, volatile matter and total sulfur. The models for groups 1 and 2 explained 80.0 percent and 79.2 percent of the variance, respectively.

D. Pyrolysis Background

1. Overview

Analytical pyrolysis is a powerful procedure for understanding the underlying structure of complex mixtures of organic compounds when precautions are taken to ensure a uniform and reproducible pyrolysis. Pyrolysis has been used for a wide variety of organic

substances, and an extensive review of applications of pyrolysis to organic geochemistry has been published (Larter and Douglas, 1982). Meuzelaar et al. (1982a) have published a compilation of pyrolysis mass spectra and an overview of current pyrolysis mass spectrometry (Py/MS) techniques. Unfortunately, as the complexity of the sample increases, the amount of information in the spectra increases very rapidly, and the spectra become difficult to interpret visually. Because the results of the analysis are too complex for direct interpretation, computerized pattern recognition is the only method currently available capable of extracting the information from the pyrolysis mass spectra of substances like coal.

2. Two Temperature Regimes

According to Girling (1963), the first introduction of coal or coal products directly into a mass spectrometer was when Hirota et al. (1954) heated whole coal under vacuum and collected the gas evolved at various temperatures up to 500 degrees C. The gas was analyzed in a mass spectrometer, which indicated that paraffins, olefins and benzene were present. Apparently, the earliest work reported in the English

literature involving the pyrolysis of whole coal into a mass spectrometer was by Holden and Robb (1958). Pyrolysis was accomplished by inductive heating in a tubular furnace with a temperature range of 60-450 degrees C. Prominent homologous series of alkyl-substituted benzenes, naphthalenes, phenols and naphthols were observed. In later work, (Holden and Robb, 1960), indane, acenaphthalene and anthracene series were also observed. The spectra were collected over a period of days, with long holding times at each increment in temperature. This approach gave information on the amount of evolution of each of the components with temperature. At lower temperatures, the phenomenon observed was probably desorption rather than pyrolysis. At temperatures above 300 degrees C, the mechanism was predominantly pyrolysis, indicated by a dramatic increase in the evolution of alkyl aromatic fragments. The fraction of the sample finally degassed or pyrolysed was in the range 2-28 percent. Therefore, as a measure of the gross structural features of the coal, the pyrolysis was less than optimum.

Using gas chromatography (GC) with a flame ionization detector (FID), Barker (1974b) observed two

maxima in the total evolution of hydrocarbons in the mass range C_1 to C_{10} during slow desorption/pyrolysis of petroleum source rocks. The maxima occurred at about 130 and 480 degrees C. With increasing maturation, the amount of hydrocarbons generated at 130 degrees C increased, apparently at the expense of the 480 degree C pyrolysate. In contrast, Barker (1974a) observed only one maximum for vitrinite, which occurred between 420 and 660 degrees C. With increasing rank, the temperature of maximum yield of pyrolysate increased. There was no simple relationship between rank and amount of hydrocarbons produced. The pyrolysis products were not separated and identified and the components analyzed were only a small part of the total coal. Girling (1963) used pyrolysis gas chromatography (Py/GC) to provide much more detailed information on the evolution of specific products as a function of temperature. He also observed low temperature (<300 degree C) and high temperature (300-600 degree C) product evolution indicating production by desorption and decomposition, respectively. This was true for most components. The amount of hydrocarbons and

aromatics produced by desorption was highest for middle-ranked coals.

These studies emphasized thermal effects on the moderate temperature evolution of readily recognizable organic constituents. Because they are common in any coal, these components are not as useful for discriminating and characterizing coals as the more unusual pyrolysis fragments. A more detailed technique is desirable for purposes of fingerprinting, classification, prediction, and estimation of reproducibility. One such procedure will be described in this thesis.

In a novel approach, Romovacek and Kubat (1968) attempted to calculate activation energies for evolution of benzene and toluene by dropping samples of coal into molten tin at various temperatures between 600 and 950 degrees C. The pyrolysate was passed either directly into an FID or through a GC and then into an FID. Plots of the ratio of aromatic to aliphatic constituents, identified by retention times, indicated rank.

3. Maturation

In order to examine the influence of oxygen functionalities on simulated maturation, Rouxhet et al. (1979) used infrared spectrophotometry (IR) to look at the effect of pyrolysis on sporopollenin, air-oxidized sporopollenin and lignite. The authors suggested that structural reorganization of the coal required prior removal of oxygenated functions. Villey et al. (1979) obtained similar results from subjecting sporopollenin (Recent-aged trilete spores) and lignite to low temperature pyrolysis at 436 degrees C. Using electron microscopy, differential thermal analysis (DTA), thermogravimetric analysis (TGA), IR and electron spin resonance (ESR) spectroscopy they investigated the decomposition of coal. This group intended to use the low temperature pyrolysis as a simple model for petroleum formation from disseminated kerogen. The results gave fairly precise temperatures for outgassing of H₂O and CO₂, production of tars, and "plastification". The authors speculated that kerogen (in this case lignite) is originally present as stacks of 2-3 molecules containing 5-12 aromatic rings each, distributed at random and connected by non-aromatic

groups. During maturation the stacks don't grow, but suddenly become oriented in the oil production zone. The amount of orientation possible decreases when oxygen content is higher, perhaps because oxygen restricts mobility of the stacks and prevents molecular orientation from developing by cross-linking the stacks.

Van Graas et al. (1980) used Curie-point pyrolysis mass spectrometry (Py/MS) to examine 23 coals of various ranks from the United States and the Netherlands. In this paper, the pattern recognition techniques of distance tables and non-linear mapping alone were not sufficient to predict rank, although visual selection of five peaks indicated that the pyrolysis mass spectra contained information related to rank. A factor analysis of these Py/MS data showed a good correlation between the first principal component and rank (van Graas et al., 1979). The peaks most correlated with rank were m/z 80, 107, 108, 132, 144, 146 and 158. From Py/GC/MS these were identified as the alkyl-substituted homologs of phenol, indene, benzofuran, and indane. Pyrolysis mass spectra of

fusinite, alginite, sporinite and vitrinite were reported for a single coal.

Work at the University of Utah has involved the use of Py/MS with computerized pattern recognition to investigate differences between coals and coal liquefaction products among a highly diverse set of Rocky Mountain coals. Differences between coal type (humic vs. sapropelic) and rank were observed (Meuzelaar et al., 1982b). The specific mass spectral peaks which correlated with rank were similar to ones found in this laboratory (Durfee et al., 1982), although the pattern recognition techniques used were different. In later work (Meuzelaar et al., 1984b), differences between Py/MS spectra of coals from the same seam were found to be small, even when large differences in petrographic composition were apparent. Although oxygen-containing compounds were found to correlate negatively with rank, aromatic vs. aliphatic content seemed more indicative of depositional environment.

4. Macerals

Work at the University of Newcastle upon Tyne has focused on determining differences in pyrolysate from

different types of kerogen (insoluble sedimentary organic matter) in order to increase understanding of thermal maturation, petroleum generation, and for petroleum/source rock correlation. Coal macerals have been used in much of this work.

Larter et al. (1977) compared total material evolved from vitrinite, sporinite and alginite macerals at 110 degrees C and 300 degrees C using an FID. They found that, although vitrinites generated 2 to 3 times less volatile material than the other macerals, a greater fraction of the volatiles in the vitrinite is low molecular weight material. Therefore, plots of ratios of the normalized amount of material evolved at different temperatures distinguished between exinites and vitrinites.

At higher temperatures, in the range 500 - 740 degrees C, vitrinites were found to yield predominantly low molecular weight alkyl and phenolic aromatic hydrocarbons, although some straight-chain material was present (Larter and Douglas, 1978). Alginites contained mostly straight-chain paraffins and olefins, and sporinite was intermediate between vitrinites and alginites, producing also some saturated branched and

cyclic material. They reported that, at a given temperature, the ratio of toluene to m-xylene decreases with increasing rank. At a given rank this ratio increases with temperature, suggesting more severe thermal degradation.

Using an inductive furnace coupled with a GC/MS, Larter et al. (1978) pyrolysed samples of coal macerals at 600 degrees C. Monitoring m/z 141 (alkyl-naphthalene tropyllium analog), these authors found alkyl naphthalenes in alginite, sporinite and vitrinite. Sporinite and vitrinite contained more of the mono- and dimethyl naphthalenes than the alginite, but they contained no appreciable amounts of naphthalene with higher substitution. By monitoring tropyllium (m/z 91), more longer-chained alkyl benzenes were found in the alginite than in the other two macerals, although all three macerals contained alkyl benzenes with a wide range of chain sizes. The alkyl benzene side-chain length for all three macerals was apparently longer than for oil shale and marine kerogen samples (Solli et al., 1979). Comparisons between kerogens is only relative, however, since the fraction of the sample yielded as pyrolysate is different (less

than 40 percent for vitrinite at 600 degrees C) for different types of kerogen (Solli et al., 1980).

For vitrinites, exinites and sporinites, Meuzelaar and Harper (1984) confirmed the decrease in phenolic and dihydroxybenzene series and the increase in alkyl aromatic series with increasing rank. They also suggested that significant quantities of polyisoprenoid polyenes could be observed in the spectra of sporinites.

5. Role of Oxygen in Maturation and Liquefaction

As mentioned above, Rouxhet et al. (1979) and Villey et al. (1979) reported the importance of the loss of oxygen in producing the ordered state we associate with increasing rank among coals. A glance at the van Krevelen diagram (Figure 2, Tissot and Welte, 1978) indicates that the role of oxygen in diagenesis of coal is a very important one. The dry and mineral matter free (dmmf) oxygen content decreases almost linearly with increasing carbon content (Figure 3) throughout the range of coal ranks, from lignites to anthracites (Berkowitz, 1979). Apart from the simple correlation between liquefaction and rank, oxygen-containing functional groups may play an active role in

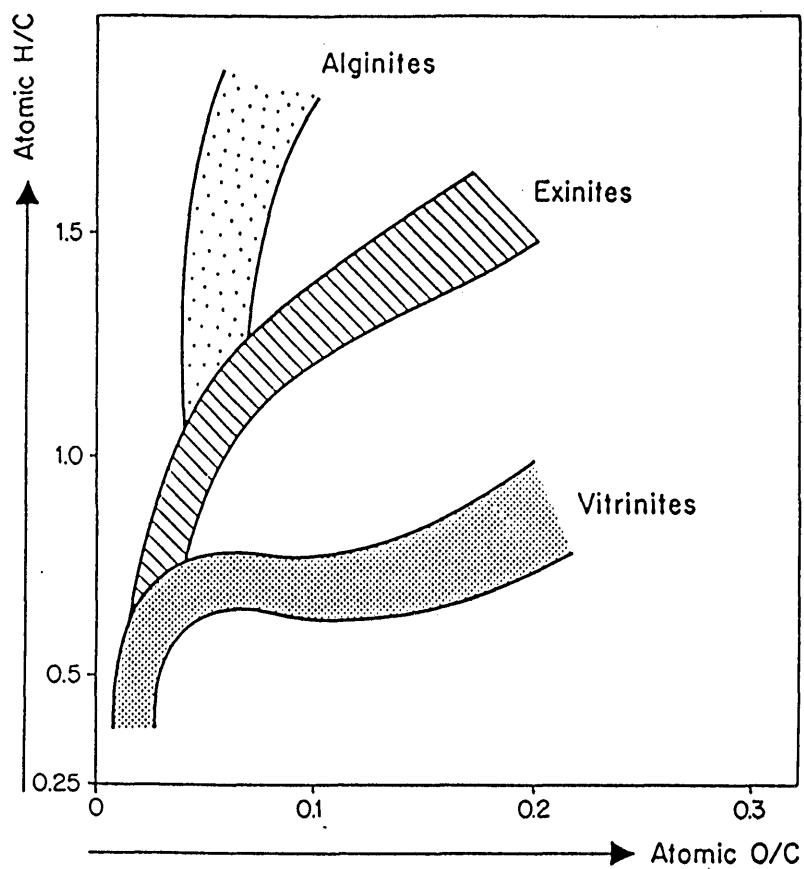


Figure 2. van Krevelen diagram (Tissot and Welte, 1978)

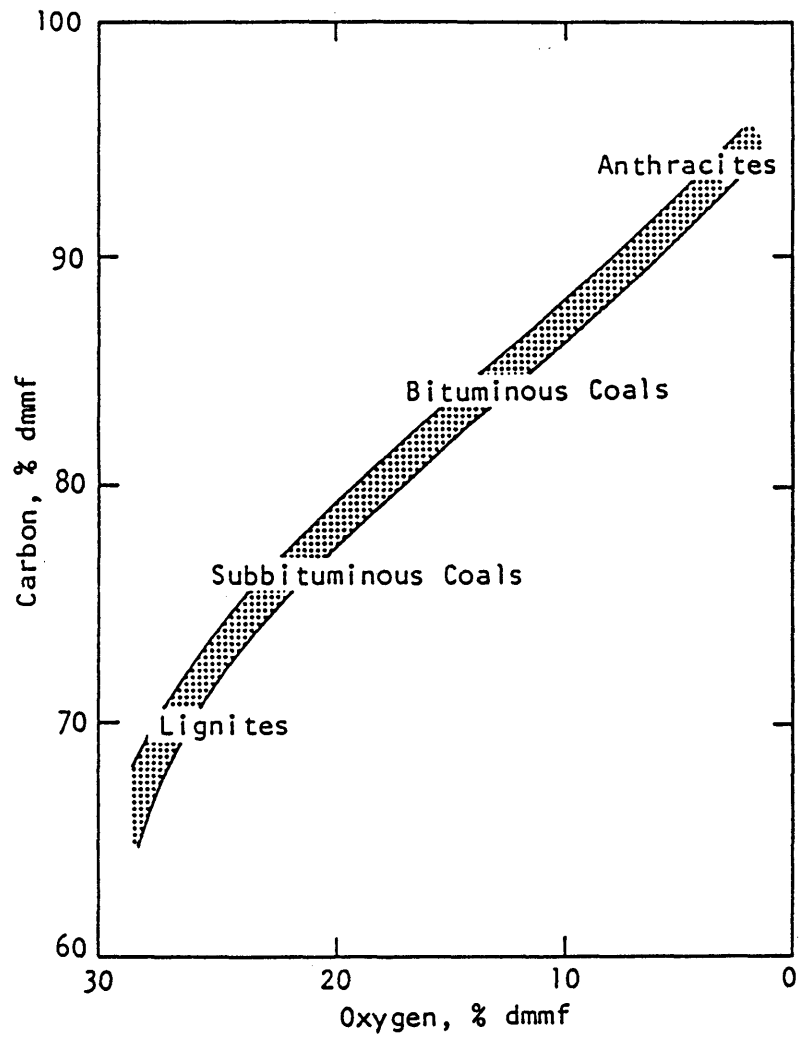


Figure 3. O content vs. rank (Berkowitz, 1979).

liquefaction processes. Villey et al. (1979) suggested that oxygen discourages liquefaction. Because of the relatively low C-O bonding energy, Z. Abdel-Baset et al. (1978) speculated that phenolic OH may dissociate to form free radicals, assisting hydrogenation reactions during liquefaction. The amount of oxygen as OH was found to be positively correlated with liquefaction conversion, but this effect could not be estimated because it was confounded with the negative correlation between conversion and rank.

Because of the high degree of chemical complexity of whole coal samples, some work on liquefaction mechanisms has been done using model compounds thought to be representative of certain components of coal. Work at Exxon has probed the mechanisms involved in the pyrolysis of dibenzyl ether (Schlosberg et al., 1981), benzyl phenyl ether (Schlosberg et al., 1980), and aryl alkyl ethers (Schlosberg et al., 1983) in tubing bombs. Whether the reactions produce low molecular weight (desirable) products or condensed high molecular weight products under liquefaction conditions depends on the availability of hydrogen and on residence time. Phenoxy radicals were found to be more efficient at

abstracting hydrogen than benzyl radicals, reflected in a higher yield of phenol than toluene in an experiment in which no hydrogen donor was added to benzyl phenyl ether. Benzyl radicals promoted termination reactions leading to heavier products.

Mastral et al. (1984) investigated the effect of various catalysts and of substitution of alkyl groups on anisole rings on the yield of bitumens after liquefaction. Activation of the ring from alkyl substitution improved the yield. Siskin and Aczel (1983) presented evidence based on liquefaction of model compounds and derivatized coal that most of the single ring phenols produced during 600 degree C pyrolysis are derived from alkyl-aryl ethers. Deprotonation of phenolic oxygen using concentrated KOH prior to liquefaction reduced the yield of liquid products, but increased the yield of gases. However, catalytic effects from the reagent may have been involved because of the pretreatment of the coal.

A detailed GC/MS analysis of atmospheric bottoms from the EDS process by McClennen et al. (1983) revealed relatively large concentrations of indanol and its alkyl homologs among the oxygen-containing aromatic

products. These authors suggested a further role of the oxygen-containing compounds as solvents.

6. Justification for Peak Assignments in Py/MS

By itself, Py/MS indicates which peaks (m/z values) are present in the pyrolysate, but does not provide enough information to specify unambiguously the chemical identity of the peaks. Since isomers exist with the same molecular weight, additional information, such as the separation step in GC/MS or the additional analysis step as in MS/MS, must be performed in order to confirm the identity of the species corresponding to peaks in the mass spectrum. But prior separation steps greatly increase the processing time of a sample, and would therefore increase the time and expense of a liquefaction prediction. Also, GC/MS data are much more difficult to process using computerized pattern recognition methods because of the continuous nature of the chromatogram, lack of reproducibility in the gas chromatography step, and the overabundance of information obtained from a GC/MS run. However, previous work involving GC/MS gives credence to the peak assignments in later chapters by identifying common pyrolysis products of coal.

Girling (1963) investigated both thermal desorption of volatiles and low temperature pyrolysis in the temperature range 100 - 500 degrees C using an inductive furnace and GC with a gas density balance detector. He measured exclusively hydrocarbons, and reported a wide variety of products (Table I). The majority of the hydrocarbon products were aromatic.

Table I. Compounds identified in coal pyrolysis products (Girling, 1963)

Paraffins	Olefins
n-Propane	Butene-1
n-Butane	cis- and trans-Butene-2
2-Methylbutane	Pentene-1
n-Pentane	cis- and trans-Pentene-2
2-Methylpentane	Hexene-1
3-Methylpentane	cis- and trans-Hexene-2
n-Hexane	Heptene-1
2-Methylhexane	cis- and trans-Heptene-2
3-Methylhexane	Octene-1
2,2-Dimethylpentane	Nonene-1
2,4-Dimethylpentane	
n-Heptane	Aromatics
n-Octane	Benzene
n-Nonane	Toluene
n-Decane	Ethylbenzene
Naphthenes	p-Xylene
Methylcyclopentane	m-Xylene
Cyclohexane	o-Xylene
Methylcyclohexane	1,3,5-Trimethylbenzene
Cyclohexene	n-Propylbenzene
	1,2,4-Trimethylbenzene

The papers by the Newcastle upon Tyne group (cited in the Macerals section above) identified diverse hydrocarbons in solvent extracts of coal (Allan and Douglas, 1977) and flotation-separated coal macerals (Larter and Douglas, 1978; and Larter et al., 1978). For Py/MS identification, the most important compounds identified were the lower molecular weight paraffins and alkyl-substituted benzenes (Solli et al., 1979) and naphthalenes (Solli et al., 1980).

Van Graas et al. (1980) looked at many of the oxygen-containing components using Curie-point pyrolysis, low voltage MS and GC/MS. A number of compound classes were identified, including methyl- and C₂-substituted phenols, indenenes, methyl- and C₂-substituted benzofurans, and methyl indanes. A summary of the compounds found is shown in Table II.

Table II. Compounds identified in the pyrolysis of GC/MS analysis of coal ILLS-23 (fixed carbon 55 percent). (van Graas *et al.*, 1979)

1. Benzene
2. Heptene
3. Heptane
4. Toluene
5. Methylthiophene
6. Octene
7. Octane
8. Ethylbenzene + ethylthiophene
9. m-Xylene + p-xylene + dimethylthiophene
10. Dimethylthiophene
11. o-Xylene + styrene
12. Nonene
13. Nonane
14. C₃-benzenes + C₃-thiophenes
15. Methylstyrene
16. Decene
17. Decane
18. Phenol + indene + indane
19. C₄-benzenes
20. o-Cresol
21. Undecene
22. m- & p-Cresol + undecane + methylbenzofuran + methylindanes
23. Methylindenenes
24. C₂-phenol
25. Naphthalene
26. Benzothiophene
27. C₂-phenols
28. C₂-benzofuran + dodecene
29. C₂-benzofuran + dodecane
30. C₂-indene + C₃-phenol
31. Methylbenzothiophenes
32. 2-Methylnaphthalene
33. 1-Methylnaphthalene
34. Tridecane
35. Biphenyl
36. C₂-benzothiophenes
37. Ethylnaphthalene
38. Dimethylnaphthalenes
39. C₃-naphthalenes + C₃-benzothiophenes
40. Dibenzofuran
41. Fluorene

42. Methyl dibenzofurans
43. Dibenzothiophene
44. Pristene
45. Phenanthrene
46. Anthracene
47. Methylphenanthrenes + methyl-anthracenes
48. C₂-phenanthrenes + C₂-anthracenes

McClennen et al. (1983) published a detailed study of aromatic oxygenated compounds (phenols and indanols) in liquefaction bottoms from the EDS and H-coal processes. Unfortunately, the severity of the liquefaction process and the non-pyrolytic nature of the experimental treatment make it difficult to extend the information to a direct Py/MS experiment. However, this set of experiments was significant because it suggested the importance of oxygenated compounds (particularly indanols, which may have been produced during the liquefaction process) to liquefaction.

7. Py/MS for Prediction of Liquefaction Reactivity

An early attempt to correlate Curie-point Py/MS data on Western coals with reactivity (Meuzelaar et al., 1980, Figure 4) failed to uncover gross trends which could distinguish intermediate-reactivity coals from low- or high-reactivity coals (measured in a laboratory-scale, ZnCl₂-catalyzed process). Harper et al. (1984) used a pattern recognition approach very

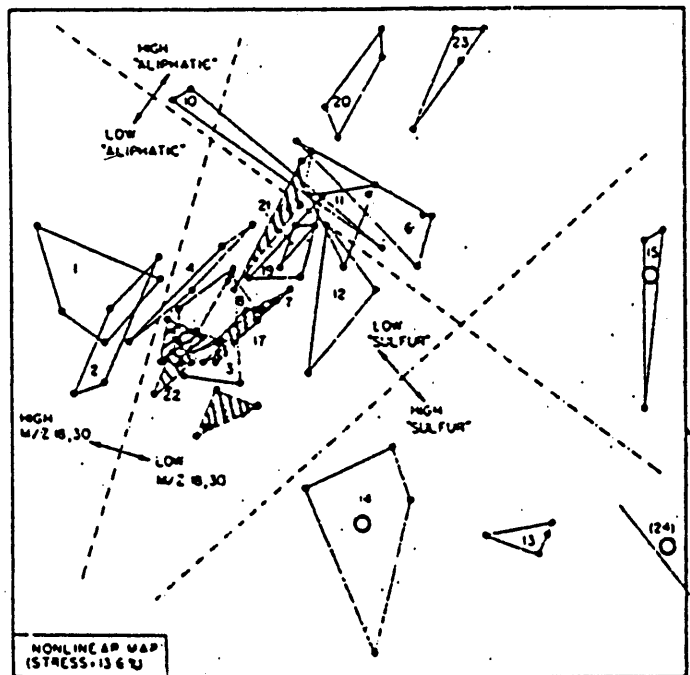


Figure 4. Early Py/MS - reactivity non-linear map
(Meuzelaar et al., 1980)

similar to the one described in this thesis to investigate correspondences between Py/MS and conventional coal parameters, including ultimate, proximate, petrographic and mineralogic measurements. Factor analyses of the spectral and conventional data matrices were performed separately. The Py/MS matrix was then analytically rotated to correspond with the matrix of conventional coal parameters. Although the procedure worked well for many of the conventional parameters, it was unsuccessful for predicting liquefaction reactivity because of the narrow reactivity range studied.

E. Pattern Recognition Background

1. Overview

The number of tools available for pattern recognition is enormous and is rapidly growing. Principal component analysis was chosen as the central technique for this study because of its relative simplicity, because it is a linear technique suitable for continuous properties, and because results can be easily related to the original data.

Much confusion in comparing discussions of factor analysis and related techniques comes from an ambiguous nomenclature and differing notation. Except for the more general terminology common in the chemometric literature (object, measurement, feature, etc.) described by Harper et al. (1977), the terminology and mathematical notation of Malinowski and Howery (1980) will be used to describe most of the pattern recognition methods. A partial summary of Harper's terminology (1977, p. 15 ff.) follows:

Continuous property analysis. The data are known to represent a continuous range of responses towards some given property(ies). The goals of such analysis are the identification of what parameters (if any) are functionally related to the property and (if possible) the selection of a rule which quantitatively predicts that property.

Unsupervised analysis. The data are not known to have any systematic characteristics. The goal of such analysis is the discovery of what systematic behavior the data exhibit (if any exists)...

Object. A compound, sample, individual or other entity for which a list of characterizing parameters is present in the data base.

Measurement. An experimentally available parameter (independent variable) used to characterize the objects.

Feature. Any transformation of one or more measurements used to characterize the objects. When referring to a parameter which can be either a measurement or a feature, the term "measurement/feature" is used.

Category. One of the groups of objects studied in the classification analysis algorithms...

Property. A quantitative parameter characteristic of the objects for which a functional representation is desired (dependent variable).

Some examples of these definitions applied to the coal analysis described in this thesis are: continuous property analysis - prediction of liquefaction; unsupervised analysis - exploration of data structure; object - coal sample; measurement - intensity of a mass spectral peak; feature - intensity of a mass spectral peak or a score from factor analysis (defined below); category - group of related coals; and, property - liquefaction reactivity or rank.

Many of the techniques currently popular in the chemical literature (Fuzzy C-varieties (Bezdek et al., 1981a and 1981b), SIMCA (Wold and Sjostrom, 1977), and discriminant or canonical variates analysis (Arunachalam and Gangadharan, 1984)) are closely related to principal component analysis, but these techniques are better suited to solving class

membership or component/mixture problems. Liquefaction reactivity and most other properties of coal are continuous properties. Therefore, class membership techniques were not used extensively in the work described in this thesis, since they are not expected to be as effective for predicting coal properties as is simple principal component factor analysis.

The pattern recognition procedures described here are used to accomplish several goals:

1. Reduce the influence of noise.
2. Limit variations which, although real, do not relate directly to the problem at hand.
3. Collect information which is relevant to the problem.
4. Reduce the data set to such a size that it can be interpreted, at the possible expense of some loss of information.
5. Predict a property or the class membership of unknown samples.
6. Increase understanding of the chemistry responsible for variation in the property being studied.

In the analysis of the coals, both unsupervised and supervised learning techniques will be used.

Although there is some overlap between the two, the goals of the two types of analysis are different.

a. Unsupervised learning. Unsupervised learning is used to explore the data for undefined or poorly-defined properties in the data. Because the structure of the data is not known or cannot be quantitatively measured beforehand, the analyst has considerable flexibility in using a variety of methods to investigate differences or trends in the data. These trends are detected by cluster analysis (methods which search for samples which are inherently similar), projections and mappings. The actual values of any properties are never used as data in the unsupervised analysis.

b. Supervised Learning. The goal of supervised learning is prediction. Therefore, there must be a quantitative understanding of the properties, such as rank or liquefaction reactivity, being sought.

The supervised approach requires the development of a model. Typical models are linear models, produced by discriminant or regression analysis, and classification models produced by cluster analysis. Once a model has been developed, it is tested using an

evaluation set consisting of samples with known properties which are treated as unknowns. If the prediction is successful for the evaluation set, the model is accepted and may be used on unknowns (the test set) provided that the unknowns belong to the same population as the one from which the evaluation set was drawn.

Flexibility is much more restricted in supervised learning than in unsupervised learning, and requires a well-chosen evaluation set. For example, linear regression is a common supervised technique in which the modeled property is used as the dependent variable in a linear equation. If no evaluation set is available, then the only way to estimate the validity of the final equation is by looking at statistics, such as the coefficient of variation, which apply to the equation itself. Assumptions required by these statistics, such as the normal distribution of error or the freedom from error of the dependent variable, are not true for most chemical data. If an adequate evaluation set is available, however, then the amount of error in the estimate can be directly measured and there is no concern about the validity of the model.

Unfortunately, the evaluation step is sometimes omitted in Py/MS work, and consequently the statistical demands on the data analysis are severe.

c. A valid approach for coals. In cluster analysis situations, the ratio (R) of patterns (samples) to features (peaks in the mass spectrum) should be greater than three in order for the error rate for the training set to be a meaningful indicator of the error rate on unknowns. Even at high values of R the error rate for the training set is less than the error rate for unknowns (Foley, 1972). Repeat runs may improve this figure by reducing apparent patterns due to pure error, but at the very low values of R common in the pyrolysis literature, structure will almost inevitably exist in the data which is not reproducible (Meisel, 1972). Therefore, in the use of unsupervised techniques, the mere existence of structure in the data does not imply that the structure is related to a meaningful property. In supervised techniques, an internally consistent model may sometimes be constructed which has little predictive power.

In order to avoid these potential pitfalls, the general approach proposed for the interpretation of

pyrolysis mass spectra of coals consists of the following:

1. Unsupervised learning techniques are used to determine whether the data contain information about the property being studied, whether this information is relatively prominent in the data, and whether extraction of the information using supervised learning techniques (such as discriminant analysis or linear regression) appears promising. The unsupervised learning techniques may also suggest which supervised techniques to use. Typically, principal component analysis followed by linear mappings of the factor scores or loadings two at a time (Karhunen-Loeve or KL projection), non-linear mapping, and hierarchical clustering show the overall relationships between the samples and make it possible to estimate the magnitude of the response relative to random error.

2. Supervised learning techniques are used based on what was uncovered by unsupervised learning and independent knowledge about the property being modeled. A successful model not only provides a means for predicting properties, but also provides insight into other properties or processes which cause the predicted

property to vary. The final coal reactivity model (reported below) was calculated using a sophisticated linear regression technique. Linear regression was chosen because reactivity is a continuous property, and might be linearly related to the concentration of components or sets of components in the coal.

Unsupervised learning methods supported this linear model. Confidence in the results comes from:

- a. Careful attention to assumptions and inadequacy of the data at each step.
- b. Periodic use of random dependent variables or entire random data sets to test the method of data analysis for false or misleading conclusions.
- c. The use of repeat samples in the training set in order to visually estimate the relative magnitude of experimental error.
- d. Validation of the model using an evaluation set, provided the data set is large enough. This step is useful even if the purpose of the supervised learning is not a prediction.

A promising alternative is to "Uncertainty perturb" the data (Duewer et al., 1976) by introducing random error scaled to the known variance in the data.

This approach has the advantage of dealing successfully with non-normal distributions if the true distribution is known. It also allows different variance in different features. It might be useful not only for comparing techniques as Duewer et al. did, but could be used during the course of a conventional analysis. Since a requirement of analytical analysis is an estimate of the error in the result, this method could be used to predict the error of the estimate for complex data when the propagation of error through the calculation is uncertain. Unfortunately, the individual errors cannot yet be estimated independently for peaks in pyrolysis mass spectrometry.

2. Data Pretreatment

In contradiction to usage in the social sciences, the term normalization will be used here to describe treatment of the elements (intensities of mass spectral peaks) of one data vector (mass spectrum) based on the known total ion current for that data vector or on the intensities of other elements in the same data vector. Scaling (one type of which is called normalization in the social sciences) will refer to the treatment of corresponding elements across the data vectors.

A mass spectrum i may be represented as a row vector, d_i with elements d_{ij} which are measurements of the intensities of mass spectral peaks for spectrum i . In this notation, j is an index for the mass, which is assumed to be an integer for low resolution mass spectra. Usually the masses are arranged in order of increasing mass. The data vectors are then combined to produce one large data matrix D with elements d_{ij} .

a. Normalization. The first step in the data analysis is pretreatment. The intensities of the peaks in a set of comparable pyrolysis mass spectra vary by up to a factor of 5 despite all attempts to maintain reproducibility. Possible reasons for these variations are the tendency of the coal particles to "clump" or spread out on the Curie-point wires (described below) as they dry so that the sample is not uniform in area or thickness, systematic changes in background and sensitivity of the instrument, slight inhomogeneities in the suspension, true differences in amount of pyrolysate produced by the same amount of different coals, inaccuracy in weighing and pipetting, and so on. On the other hand, experience has shown that these

changes in total ion current have little or no observable effect on the relative abundance of peaks in the mass spectra, provided none of the peaks in the normalization interval are so intense that they saturate the analog to digital converter.

Consequently, the essential first step in treating a set of Py/MS data is to normalize the data. Ideally, the resulting normalized data do not contain less meaningful information despite the closure of the data set as described by Johanssen et al. (1984).

Information on the amount of pyrolysate/amount of sample is not available from the technique, and the fluctuations in total ion current are due to effects which are not related to chemically useful differences between samples. In any case, the effect of closure in mass spectral data is arguably small, since it leads to the loss of one degree of freedom out of about 200 in the ideal limit.

The typical normalization technique for displaying and storing mass spectral data is normalization to the base peak (McLafferty, 1980):

$$d'_{ij} = \frac{d_{ij}}{\max(d_{i1}, d_{i2}, \dots, d_{ik})} \quad (2)$$

This method of normalization is highly undesirable for pattern recognition, since it is severely affected by closure (Johansson et al., 1984). In effect, all of the variance in the largest peak is added individually to the variances of each of the remaining peaks, in proportion to their size.

The second normalization technique commonly used is normalization to the sum of all the peaks.

$$d'_{ij} = \left(\frac{d_{ij}}{\sum_{k=1}^n d_{ik}} \right) K \quad (3)$$

K is an arbitrary constant to make the numbers conveniently large, but does not influence the relative magnitude of the data points. This technique has the advantage of averaging the contribution due to pure error of the individual peaks in the normalization

factor, but remains susceptible to several other effects described below.

A third normalization technique which has been used (van Graas et al., 1980; Eshuis et al., 1977) selects peaks based on the ratio of intraclass to interclass variance, Fisher weights (defined below), or other similar measures. The applicability of normalization of this sort depends very much on the experimental question. Given the large number of very small peaks, it becomes probable that chance correlation between variable peaks and category memberships will exist which are meaningless chemically. Ratio-determined normalizations select in favor of those peaks which are correlated by chance along with the meaningful ones. If the goal of the experiment is to demonstrate that Py/MS spectra are different for different sets of mixtures, then this form of normalization artificially weights the results in favor of that hypothesis. Another problem with such techniques is that they tend to select in favor of correlated features, leading to a data set which is more repetitive than before selection.

As suggested above, peaks which are saturated have an adverse effect on normalization. Once saturated, their intensity does not change as the TIC changes. These peaks typically occur at low masses and are due to atmospheric contamination, outgassing of adsorbed gases, alkane fragments and common fragments lost. Also, because their intensity is constant, they are not reproducible after normalization.

Another adverse effect on normalization is contamination due to background. The background peaks may be divided into two types. The first is a relatively constant (or at least random) background due to contamination in the instrument, leakage into the instrument of atmospheric gases, and diffusion pump oil products. Since these peaks do not vary, they will have a "damping" effect on normalization; the normalization will under-compensate for differences in the amount of sample. These peaks are found mostly at low molecular weights since atmospheric gases are the chief source. The second type of background is potentially more destructive. During the course of running high-molecular weight compounds, we observe a steady background increase in the high-molecular weight

region of the spectrum. For many types of samples, the high molecular weight region is the most diagnostic, since it contains the most unusual components (the lower molecular weight regions contain common fragments lost). The increase of background in this region is due to the higher molecular weight pyrolysate from previous samples adsorbing to the surfaces of the instrument and then slowly volatilizing. To some extent it can be controlled by instrument and experimental design. The inlet and source are heated to vaporize the compounds so they may be pumped away. At periods during the processing of a large number of samples, the instrument is left for a period of "pump down". The samples are run according to a balanced or random design so that increasing background will not be correlated with any class memberships or dependent variables in the data. Finally, normalization should not include regions most susceptible to background, i.e. the high molecular weights. For all of these reasons, a region of the spectrum excluding the lowest and highest masses is chosen for normalization:

$$d'_{ij} = \frac{d_{ij}}{\sum_{k=a}^b d_{ik}} K \quad (4)$$

in which a and b are the indices for the lower mass and upper mass normalization limits.

b. Scaling. Mass spectral data are collected in digitized form, are multidimensional and are related to structure. Therefore, they are ideally suited to computerized pattern recognition. Practically every commonly used scaling technique has been attempted. Jurs (1971) used the square root, fourth root, logarithmic transform and a binary coding scheme on a set of 600 mass spectra with 300 spectra in the training set and 300 spectra in the evaluation set. The logarithmic transformation provided the best predictive ability in a binary classifier technique.

Fourier, Walsh (Hadamard) and Haar transformations were tested by Domokos and Frank (1981). Transformations such as these are used to increase the dynamic range of a technique, decrease storage space, improve execution speed of digital algorithms, provide improved rejection of noise, or conform with some model

of the system external to the analysis. In the case of predicting equilibrium reactivity, a simple linear model (assuming no interactions) seems appropriate, and no transformation was used. Kinetic reactivities, on the other hand, are considerably more complex.

In order to use factor analytical methods, a similarity matrix must first be calculated:

$$Z = D^t D \quad (5)$$

where Z is an unspecified similarity matrix and t denotes the transpose. If raw, unscaled data are used, the similarity matrix is called covariance about the origin. If the mean is subtracted from the data, then the similarity matrix is termed covariance about the mean. If the standard deviation is then divided, either without or with mean subtraction, then the similarity matrix is called correlation about the origin or mean respectively. For example, correlation about the mean (autoscaling) is:

$$d'_{ij} = \frac{d_{ij} - \bar{d}_j}{s_j(m)^{1/2}} \quad (6)$$

where d_{ij} is a normalized data element. The mean value \bar{d}_j for peak j is given by

$$\bar{d}_j = \sum_{i=1}^m d_{ij} \quad (7)$$

for m spectra and s_j is the standard deviation for peak j :

$$s_j = \frac{1}{m} \sum_{i=1}^m (d_{ij} - \bar{d}_j)^2 \quad (8)$$

Use of the covariance matrix implies that the variance of the features is the same. The correlation matrix assumes the variance to be proportional to the magnitude of the features.

Rozett and Petersen (1975) argued that covariance about the origin is best for mass spectral data. Their experimental results showed that it appears to achieve a more rapid reduction in variance. They argued that it preserves information about the zero point of scale (which has meaning in pure mass spectra) and about the absolute scaling of the peaks. However, the zero point of scale is not as meaningful in pyrolysis mass spectra of complex mixtures, since the intensity of a single peak is usually the sum of the intensities of several components, all with the same mass. Therefore, the absence of one of the components with like masses does not produce an intensity of zero at that mass value. Furthermore, Rozett and Petersen did not justify a relationship between speed of reduction in the variance and accuracy of the solution. Using their highly complex data set, the variance decreased rapidly in the lowest factors from principal component analysis (described below). It will be shown below that the speed of reduction is a good indication of complexity, so that the results without autoscaling were in fact more misleading. Duewer et al. (1976) found that correlation about the mean was "by far" the most stable

of the four methods toward analytical uncertainty. Data to be treated with component analysis are very frequently autoscaled. Autoscaling is almost universally used in the pyrolysis literature, and has been used in these studies. However, it should be noted that, without justification from a model or some knowledge about the data external to the pretreatment, autoscaling (or other normalization transformation) alone does not necessarily yield a data space in which the distance between points is a measure of similarity (Meisel, 1972).

In some situations it is fruitful to examine the data to identify the peaks which have the greatest potential for revealing trends in the data before any formal pattern recognition has been done. The Fisher (Fisher, 1936) and variance weighting methods indicate the peaks which have the greatest ratio of between-class variance to within-class variance. Since these weighting techniques require foreknowledge of the class membership of the data, they are supervised techniques, and taint any future unsupervised data analysis. However, as described below, they may provide

information which guides the course of the later data analysis.

The Fisher weight f_{ij} is defined for each class in the data:

$$f_{jmn} = \frac{(x_{mj} - \bar{x}_{nj})^2}{\sum_{k=1}^{Nm} \frac{(x_{kmj} - \bar{x}_{mj})^2}{Nm} + \sum_{k=1}^{Nm} \frac{(x_{knj} - \bar{x}_{nj})^2}{Nm}} \quad (9)$$

The overall Fisher weight is defined as the mean of the individual category pair weights (Harper et al., 1977).

3. Pattern Recognition Techniques

a. Feature Selection: Apart from removing peaks which are saturated or which are independently known to be due to background, no other common method of feature selection appears justified prior to principal component analysis. In particular, methods which are designed to delete peaks which do not conform to the experimental hypothesis should be avoided in order to avoid artificially biasing the data. Feature selection is included as a pattern recognition technique because

of its profound effect on the data, and because the technique of principal component analysis, used first for feature selection, is intimately related to many unsupervised and supervised learning techniques.

Component analysis is one of a set of techniques which are related mathematically. These techniques are collectively referred to as factor analysis, although the term "factor analysis" is often used to refer specifically to component analysis in the American chemical literature.

A detailed derivation of the well-known identities and interpretation of the principal component model would obscure the application used on coals, and may be found elsewhere (Rummel, 1970; Malinowski and Howery, 1980). Instead, a brief justification of the approach will be used to introduce some of the relevant variables and matrices for later use; first from a geometrical perspective and then from a mathematical one.

The mathematical manipulations used in factor analysis are isomorphic (have a one-to-one correspondence) with geometrical manipulations in a space (called the data space) in which the distance

between points is a measure of their similarity. The isomorphism is so obvious that it is easily overlooked in two dimensions, where data mapped on Cartesian coordinates show trends which can be interpreted mathematically.

For instance, if two measurements are taken on an object, then the value of the two measurements may be plotted on two-coordinate (Cartesian) graph paper as a point. Note that no explicit calculation was involved. The plotting of the point on coordinates is just a manifestation of the one-to-one correspondence between ordered pairs of numbers and points on a plane. If the same two measurements taken on a different object are similar to those of the first object, then the point corresponding to this object will lie close to the point for the other object; otherwise, the two points will lie farther apart.

For a large number of objects, the similarity relationships can be easily summarized on a graph of this type. These similarity relationships would probably not be obvious if they were simply listed in a table.

To carry the two-dimensional analogy a step further, if the coordinate axes were both rotated through the same angle, the distance relationships between the data points would not change; only the projections of the data points onto the rotated axes would be different. In order to relate the two-dimensional analogy, consider changing the angle between the two axes (assumed to be orthogonal up until now). If this angle is changed, then the distance relationships between the data points are also changed. Such a change per se is not necessarily advantageous; factor analysis provides a solution which significantly improves the data because it changes the angle between the coordinates in a predefined way, using information contained in the data itself. By accounting for the correlation between the original measurements, factor analysis reproduces the same information with a minimum number of terms.

Since there are frequently more than two measurements in "real-world" data, there are frequently more than two axes necessary to account for all the information in the data space; but factor analysis can achieve a dramatic reduction in the number of axes

necessary. Spaces with more than two dimensions are difficult to plot on paper, but can often still be understood more easily in geometrical terms or by analogy to the two or three dimensional case. To summarize, component analysis achieves a dramatic reduction in the number of dimensions necessary by defining a set of orthogonal, uncorrelated axes (factors), each a linear combination of the original, correlated features. These factors form a basis (set of axes or coordinates) for the data space.

The measurements (peaks) in a mass spectrum of a complex mixture such as coal are highly correlated, and consequently the "intrinsic dimensionality" of the data space is usually much less than the number of measurements. This reflects the fact that homologous series such as methyl phenol, ethyl phenol, propyl phenol, etc. are correlated because they have similar chemical behavior, come from the same source, or are the product of the same geochemical conditions. Other peaks may be fragments produced by a common source. Factor analysis provides a set of independent factors which span the data space by factoring the data matrix

into a row matrix R of order $l \times m$ and a column matrix C of order $m \times n$ where m is the lesser of l and n :

$$D = R C . \quad (10)$$

The elements of R will be referred to here as scores and the elements of C as loadings. The rows of C are eigenvectors of the similarity matrix. The corresponding eigenvalues are proportional to the amount of variance. The eigenvectors calculated by the component factor model are derived such that the first factor describes the greatest amount of variance in the data matrix. The second and subsequent factors are each orthogonal to all the previous ones and each describes the maximum possible amount of remaining variance, until m factors have accounted for all of the variance. The amount of variance in successive factors rapidly decreases to the level of experimental error. These higher factors are removed, simplifying the data. A new column matrix C^* and row matrix R^* remain (Malinowski and Howery, 1980):

$$D = D^* = R^* C^* , \quad (11)$$

in which the new row matrix R^* is of order $l \times m'$ and the new column matrix is of order $m' \times n$, $m' < m$.

Numerous criteria have been put forward for determining how many factors to retain. For some research questions, the goal of the factor analysis is to estimate the number of independent factors controlling the data (Rummel, 1970). Therefore, the method used for feature selection is critical. The complicated nature of coal insures that there are a great number of such factors (independent ways in which the composition measured by Py/MS can vary). Therefore the use of factor analysis to predict this number would not be reasonable. Instead, factors are retained until a large amount (say 90 percent or more) of the variance has been defined, and the others are eliminated. Since the factors with the greatest amount of variance are not necessarily the most meaningful (Jolliffe, 1982), the number of factors retained from mass spectral data of coal will be relatively large, ranging from about 10 to 30. This results in a compressed data space which contains almost all of the information in the original data and which still retains linear relationships in

the data. Cross validation and stepwise linear regression will be used to determine which of the factors retained are statistically significant.

When the higher (low variance) factors are removed, the new matrix D^* in Equation 9 is a projection of the original m -dimensional data D onto an m' dimensional hyperplane defined by the first m' factors. Once this elimination of noise, irrelevant information and correlation in the data has been effected, the m' dimensional hyperplane may be thought of as a more refined, lower dimensional data space.

b. Unsupervised Learning: The most basic type of unsupervised learning is a close examination of the raw data in tabular form and in plots. Experience has shown that mysterious patterns in the data or failure to find any patterns after days of work may often be traced to such problems as misassignment of peaks (a series of peaks one mass unit off), formatting errors, misassignment of category numbers and low signal to noise, which are obscured by normalization and scaling.

Component analysis is fundamentally a feature selection technique. The treatment of the data following the component analysis determines the class

of pattern recognition the subsequent analysis belongs to; both unsupervised and supervised extensions of component analysis are commonly used. The solution from component analysis is unique because of the way component analysis accounts for variance in the data. However unless there is some scientific principle which dictates that a property of the data should be related to maximum variance, once the refined data space has been derived, the component solution no longer has any special meaning. Malinowski and Howery (1980) refer to the component factors as abstract factors. In order to find a correspondence between the factors and meaningful properties in the data, some form of rotation is needed.

Rotation is the geometric equivalent of rotating the abstract factor axes in the reduced data space so that they conform to real or meaningful factors. It is the mathematical equivalent of linearly transforming the factor solution vectors to a new set of vectors. If T is a transformation matrix, then a new row matrix

$$R^* = R T \quad (12)$$

may be calculated. If T is orthonormal then

$$D = R T T^t C \quad (13)$$

in which the superscript t denotes the transpose.

The rotation may be orthogonal or oblique. The most popular method of orthogonal rotation is varimax (Rummel, 1970). Varimax maximizes the variance of the squared (usually normalized) loadings under the assumption that a real factor will have large loadings for correlated features and small loadings for all other features.

Varimax rotation offers the security of providing a solution which is mathematically determinate, but assumptions are required to connect the varimax result to a set of real factors (properties). Real properties in coal are rarely orthogonal. For instance, rank and liquefaction reactivity are correlated. Maximizing loadings on one factor implies that the same loadings are minimized on all the other factors, but conceptually, a single mass in a spectrum may easily correlate with more than one property in the data. Finally, varimax provides one unique set of numbers, but does nothing to indicate the correspondence of the

rotated factors to real properties. On the other hand, varimax requires no prior knowledge of properties, and is therefore a valid unsupervised technique. Oblique methods which require knowledge of properties will be discussed below under supervised learning.

Graphical rotation is intermediate between orthogonal and oblique rotation. Scores on the raw, abstract factors may be plotted two at a time on Cartesian coordinates. Since the factor analysis solution is mathematically equivalent to the Karhunen-Loeve transformation (Meisel, 1972), these are called KL plots. If the normalization and scaling methods are such that they produce a data space in which the distance between points is a measure of their similarity, then distance related patterns on KL plots will correspond to chemical relationships between the spectra. The KL plots may be rapidly scanned, two factors at a time, for meaningful patterns. Alternatively, or at the same time, the loadings may be scanned for chemical information. Since the loadings are correlation coefficients between the intensities at each mass and the factor axes, the (positive and negative) loadings may be plotted and interpreted in a

manner similar to mass spectra, as a function of mass. The KL coordinates (the factors) can then be rotated, two at a time, through an angle α .

$$r'_x = r_x \cos \alpha + r_y \sin \alpha \quad (14)$$

$$r'_y = -r_x \sin \alpha + r_y \cos \alpha \quad (15)$$

where r is a column vector of factor scores from the row matrix R and r' is the new, rotated vector (Windig et al., 1981). The transformation in this case is the m -dimensional Givens matrix G , which is an identity matrix except for four elements, $g_{xx} = g_{yy} = \cos \alpha_{xy}$ and, for $x > y$, $-g_{xy} = g_{yx} = \sin \alpha_{xy}$. The Givens matrix is orthogonal, as are any number of products of Givens matrices (Searle, 1982), so that the rotated row and column matrices are orthogonal and reproduce the data:

$$D = (R \ T) \begin{matrix} & & & t \\ & & & \\ & & & \\ & & & \end{matrix} (T \ C) \quad (16)$$

Examination of the rotated scores gives an indication of patterns among the objects and examination of the rotated loadings gives information about the chemistry that the scores model. In this mode, graphical rotation is an orthogonal technique. The new rotated factors may be rotated further with other abstract factors until one or more optimum factors are found which relate to known or suspected structure in the data. Among its disadvantages, factor rotation is time-consuming and subjective, but it does enable a patient investigator to isolate patterns from data of high intrinsic dimensionality, and allows a much better understanding of the limitations and capabilities of the data than do tabular representations.

Hierarchical clustering is another commonly used unsupervised learning technique. Based on the Euclidian distance matrix, a matrix of similarity values is calculated according to the equation:

$$s_{ij} = 1 - \left(\frac{d_{ij}}{d_{\max}} \right) \quad (1)$$

where s_{ij} represents the similarity value between data vectors i and j , and d_{\max} is the maximum distance

between vectors in the data set. Therefore, the greatest similarity is the one with the least distance, and the values are scaled between 0 and 1. The hierarchical clustering results are typically plotted on a dendrogram, which group samples with high similarity together, and indicates clustering by the connectivity of the diagram. Several hierarchical clustering dendrograms are reported in subsequent chapters.

Another widely used unsupervised technique is non-linear mapping (NLM) (Sammon, 1969; Kowalski and Bender, 1972). A distance metric is used to construct a distance matrix S with elements s_{ij} which are the distances between object i and object j . The simple Euclidean distance

$$s_{ij} = \sum_{k=1}^n (d_{ik} - d_{jk})^2 \quad 1/2 \quad (17)$$

is often, but not exclusively, used. Then, starting either with a random mapping or with an educated guess such as the first two principal components, the data are mapped from the original m dimensions to m' dimensions ($m' < m$). A value of $m'=2$ is generally

selected so the map can be displayed on paper. The algorithm iterates to a minimum in the mapping error

$$E(\rho) = \sum_{i < j} \frac{(s_{ij} - s_{ij}^*)^2}{s_{ij}^\rho} \quad (18)$$

in which s_{ij}^* is the distance between point i and j on the m -dimensional map, ρ is an adjustable parameter which biases the routine toward larger or smaller distances, and E is the error or "stress" associated with the mapping. Note that E is a function of the number of samples, the normalization method, and scaling techniques. Therefore values of stress reported in the literature are difficult to compare, as are absolute measures of distance.

There are several disadvantages to non-linear mapping. For large data sets with high intrinsic dimensionality, the algorithm may be quite slow to converge, and therefore expensive. For data in which the variance decreases rapidly in factor analysis, the NLM solution looks very similar to the KL plot of the first two principal components. On the other hand, if the variance converges slowly, the error $E(\rho)$ will be high and the NLM will not show meaningful patterns

contained in higher factors. The NLM is non-linear, and therefore cannot be easily related to features in the original data. The NLM provides only one solution. In data with a large fraction of the variance related to a factor which is not the property being investigated, or which has many independent sources of variation, the desired property may be obscured by the other structure in the data. The NLM technique alone provides no method for removing the masking variance. An advantage to nonlinear mapping is that it provides the single best picture of the gross structure in the data. Structure apparent on the NLM is definitely structure which exists in the treated data.

c. Supervised Learning: In searching for patterns involving known properties relationships are rarely completely independent, and oblique rotations are required. Graphical rotation can be used in an oblique mode by searching for rotations of the factor axes that best display the known structure two at a time. It has been used successfully in a supervised mode by eliminating the test set from the display, building a model based on the training set, and seeing how the test set then plots. Focusing on the features rather

than on the objects, the loadings may be examined during rotation to see whether they correspond to known chemical properties (Windig et al., 1981). If the expected mathematical structure is well known, however, a mathematical technique provides a more rapid and more accurate solution than a graphical one.

Some properties of coal such as rank, S and O content and so forth are arguably linear with respect to mass spectral peaks, i.e.

$$\hat{p}_i = \beta_0 + \sum_{j=1}^k \beta_j d_{ij} \quad (19)$$

where the d's are intensities of peaks 1 to k in spectrum i, p is a predicted value of the property and the β are adjustable parameters. For a set of data with a number of spectra, Equation 17 may be expressed more economically as

$$\hat{p} = D \beta \quad (20)$$

in which p is a vector of predicted values. In order to accommodate β_0 , the first vector in matrix D is the column vector $d_0 = 1$. Although the β are

meaningful in themselves because they show the relative importance of each mass spectral peak in determining reactivity, calculating the β directly is not usually straightforward since k is usually much greater than i ; therefore i equations cannot be solved for k unknowns. However, the factor analytical solution reduces the data to a more manageable number of dimensions linearly, so that if Equations 18 and 9 hold, another set of coefficients β' exist satisfying

$$\hat{p} = R^* \beta' \quad (21)$$

If m' is small, the β' may be readily calculated using multiple linear regression. After scaling to length, the vector β' produces an orthogonal transformation:

$$D^* = R^* \beta \beta^t C^* \quad (22)$$

Once regression has been performed it would be possible to predict the value of the property p for an unknown sample from its mass spectrum, but the results of the analysis still contain some very useful information; the magnitude of the β in Equation 17.

From Equations 8 and 19, since C^* is orthonormal,

$$\hat{p}^* = D^* C^{*-1} \beta' = D^* C^{*t} \beta' \quad (23)$$

in which C^{*-1} and C^{*t} are the inverse and transpose of C^* , respectively. \hat{p}^* is a prediction of property p which differs from \hat{p} in Equation 18 because of the factor analysis. If the transformation from D to D^* has removed some of the experimental error, then \hat{p}^* will be a better prediction than p . Like Equation 18,

$$\hat{p} = D^* \beta \quad (24)$$

From Equations 21 and 22,

$$\beta = D^{*-1} D^* C^{*t} \beta' = C^{*t} \beta' \quad (25)$$

and β may be calculated since C^{*T} and β' are known. The coefficients β reflect the importance of each peak in the mass spectrum for determining reactivity. For unknown coal i , the reactivity p_i may be predicted from its mass spectrum, row vector a , using Equation 22:

$$\hat{p}_i = a \beta \quad (26)$$

In order to choose the factors for the model which are most related to the property under investigation, stepwise linear regression may be used (Draper and Smith, 1981). Starting with one variable, regression equations are calculated in succession, adding a new variable (factor) to the regression equation on each pass, based on its ability to improve the fit until the addition of a factor does not significantly improve the model.

For regression calculations with a large number of independent variables such as this, the traditional measures of fit such as the coefficient of variation and the standard error of the estimate must be interpreted cautiously, especially when the assumptions required by linear regression may be imperfectly satisfied, such as the assumption that the dependent variables are error-free (Draper and Smith, 1981). Adding variables to the predictive model will always improve these statistics, although the error in predicting unknowns may increase. Cross validation is a method which measures the ability of the model to

predict unknowns. A subset of the data vectors with known values of the property is selected to develop the model and then the value predicted is compared to the known values for the spectra which were not used in the model.

II. RESOURCES

A. Instrumentation

Since several groups of coals were investigated at different times over a period of years, no single set of pyrolysis conditions was used to collect all of the pyrolysis mass spectra. Instead, the precise experimental techniques were adjusted to reflect different sample size and characteristics, increasing experience and changes in the instrument. The general approach will be described here, but modifications of the experimental techniques or instrumental parameters for specific experiments will be described where pertinent. The coals were ground either by a machine such as a ball mill or by hand using an agate mortar and pestle to a fine powder (about 100 mesh or less). One or more portions of approximately 5 mg each were weighed on an electrobalance and sufficient distilled methanol was added to produce a suspension of 5.0 mg/mL. The suspensions were shaken and ultrasonicated if necessary until no appreciable settling occurred. If ultrasonication did not produce a suitable suspension, the coal was reground to a finer particle size. Five microliter aliquots were withdrawn from the

suspensions using a micropipette and placed on rotating ferromagnetic wires. Two, three or more drops were placed on each wire, for a total of between 50 and 100 micrograms per wire.

After evaporating the methanol, the wires coated with coal were placed axial to a high frequency coil located in the high vacuum inlet to an Extranuclear SpectrEL quadrupole mass spectrometer. The pyrolysis was conducted using a Fisher Curie-point pyrolyzer (1.5 kW, 1.1 MHz) power supply. When the radio frequency current was applied to the pyrolyzer coil, the temperature of the ferromagnetic wire rapidly rose to its stable Curie-point temperature. Both 510 and 610 degree C wires were used. The Curie point depends on the relative amounts of Fe, Ni and Co in the alloy used to make the commercially available wires. The pyrolysed sample expanded into a gold-plated expansion chamber from which it leaked into the electron ionization source to the mass spectrometer. The range of ionization energies 12 - 15 eV was used because it offered a reasonable compromise between ionization efficiency and limited fragmentation for hydrocarbons (Lumpkin and Aczel, 1964). The mass spectrometer was

scanned in a mass range which varied from experiment to experiment, but was typically from 10 to 240 or 300 amu. From 20 to 200 consecutive scans were recorded at a rate of approximately 1000 amu/sec. The data were collected and added in real time; then processed by a peak recognizing algorithm and stored as mass/intensity lists.

B. Computer resources

A variety of programs were necessary in order to collect, transfer and process the mass spectrometric data. The program SCAN determined the optimum instrumental conditions by plotting, storing and integrating the total amount or selected ranges of masses over time. The data were collected from the mass spectrometer, the peaks identified, previewed and stored as mass/intensity data by a program named ANALOG. This program also provides for retrieval of mass spectra for later viewing and tabulating. Normalization, concatenation and formatting were performed by a program named LAZY. SCAN, ANALOG and LAZY were written in FORTRAN and operate on a PDP-11/73 microcomputer. Transfer of the data to the CSM DEC-10

mainframe was accomplished using the NIH program CLINK prior to December, 1985. Subsequently, KERMIT was used to communicate with the CSM 8700 VAX system.

Preliminary processing of the data (including autoscaling and factor analysis) were performed using the chemometric package ARTHUR. The numerous pattern recognition routines in ARTHUR were also used for some of the unsupervised and supervised learning tasks. The output from ARTHUR was used as input for the program SPIN, which performs graphical rotations, targeted rotations, variance diagraming and factor spectra, all with graphical and tabular displays (described in Appendix I). The output from ARTHUR was also manipulated with an editor to serve as input to the program SPSS. Since SPSS is only capable of factor-analyzing 100 variables, the results of the factor analysis from ARTHUR were input along with the raw data matrix and matrices of dependent variables (properties) prepared using ad hoc programs. The linear regression coefficients and other relevant output from SPSS were transferred to IBM-PC and TI-PC personal computers, suitably edited using ad hoc programs written in BASIC and FORTRAN, and input into LOTUS 1-2-3 for matrix

multiplication, sensitivity analysis and graphical display. CLINK, ARTHUR, SPSS and LOTUS are available commercially. All of the other programs were written for the specific purpose of this thesis project.

C. Samples

Table III lists the coal samples used for the analyses. The groups of coals indicated in the table were obtained at different times, and in some cases were packaged and stored under different conditions. The large initial problems of data transfer, data formatting, debugging of programs and learning the use of ARTHUR and becoming familiar with the usage and output of a "working subset" of its many subprograms was accomplished using the smaller group 1 and group 2 data sets for reasons of economy and interpretability. Then larger and more varied groups were analyzed alone or in combination. Finally, all of the PSOC coals stored under comparable conditions and all of the EXXON coals were analyzed to produce a library of comparable pyrolysis mass spectra.

Table III. Coal samples available

Group 1: Original seven whole coals

Island Park, KY
 Consolidated Welling, WV
 Island Creek, KY
 University of Kentucky
 P&M, KY
 Ireland Mine, WV
 Eagle mine, CO

Group 2: Kinetic PSOC coals

PSOC 071	PSOC 107	PSOC 130	PSOC 151
370	437	444	FIES Mine

Group 3: Coking coals

PSOC 253	PSOC 254	PSOC 257	PSOC 258
259	260	320	323
822			

Group 4: "First" PSOC coals

PSOC 104	PSOC 265	PSOC 271	PSOC 276
278	302	305	307
308	320		

Group 5: "Last" PSOC coals

PSOC 217	PSOC 219	PSOC 268	PSOC 279
280	281	284	306
330	341	345	357
375	399	401	580
592	607		

Group 6: Exxon coals

EXCL 046AK	EXCL 033BM	EXCL 057DN	EXCL 060K9
027HB	031DM	002BB	022HN
086AV	007MK	098A4	105TD
021RH	097A3	041SJ	059SD
001YB	054MT	020CM	052SR
034SM			

The liquefaction reactivity of groups 1 and 2 was studied by Furlong (Furlong, 1981; Furlong et al., 1982) in a stirred batch autoclave reactor and in

tubing bombs. They were received as < 200 mesh powder in small glass bottles closed under the atmosphere.

Groups 3, 4 and 5 were obtained in separate shipments in #2 cans, sealed under N_2 , from the Pennsylvania State coal bank. On receipt the cans were opened in an Ar atmosphere and transferred to several small glass jars, so that in each experiment a sample which had not previously been exposed to the atmosphere could be withdrawn. The PSOC coal sample bank is a well-maintained widely-used set of coals. Therefore, the results obtained in these studies may be compared to extensive data in the literature.

Group 6 consisted of virtually pristine samples obtained from R.C. Neavel of Exxon. These samples were collected and prepared at the Baytown Coal Research laboratory. Close attention was given to representative sampling and protection from atmospheric oxygen, although collection of the samples was deliberately biased toward vitrinite-rich layers. These samples were ground, transferred and sealed under nitrogen into 1.5 ml Eppendorf containers. Nine of the samples were removed from the nitrogen atmosphere and deliberately exposed to severe weathering. The

remaining sealed containers were frozen until shortly before the pyrolysis.

III. UNSUPERVISED METHODS

A. Tubing Bomb Reactivity

The coals listed in group 1 of Table III were used in this set of experiments. Liquefaction results were obtained from Yarzab et al. (1979). Liquefaction reactivity in that paper was defined as the total yield of benzene soluble products after approximately 1 hour of reaction time. Table IV summarizes the characterization properties for this suite of coals.

Table IV. Summary of Coal Samples Studied in Tubing Bomb Reactor (Group 1 coals)

Mine	Observed ^a		Rank
	Liquefaction Reactivity (%)	Sulfur Content	
Island Creek, KY	>80	3.85	HVAB
Consolidated, WV	>80	3.50	HVAB
Island Creek-2, KY	>80	3.85	HVAB
Univ. of Kentucky, KY	>80	3.46	HVAB
P & M, KY	>80	3.81	HVAB
Ireland Mine, WV	>80	4.45	HVAB
Eagle mine, CO	<50-65	0.32	LVB

reactivity measured as solubility in benzene after reaction for approximately 60 min. at 2000 psi (H₂), 400 deg. C, with tetralin as solvent.

The general preparation and mass spectrometric analysis methods were those described in the introduction. The analyses used approximately 50

micrograms total material, a Curie-point temperature of 510 degrees C, and an electron ionization voltage of 14 eV. The spectra were obtained at the University of Utah; consequently, the mass spectral data were collected as summed spectra on a Hewlett Packard 2100 S computer and printed on paper. These data were hand-entered by the author using a text editor on the Colorado School of Mines DEC 10 computer system for subsequent statistical calculations.

Typical Py/MS spectra for two of the coals are shown in Figure 5. The Euclidean distance table, based on three replicate pyrolyses is illustrated in Table V. Values on the diagonal are the average distance for the three replicates. Therefore, a small value on the diagonal indicates good reproducibility.

As previously mentioned, it is difficult to visualize the distance table totally. The nonlinear map of the data in Figure 6 allows for better comprehension. The triplicate samples for each coal result in a triangle on the nonlinear map. Also, the size of the area enclosed by the connected data points indicates the reproducibility of the data for a selected coal.

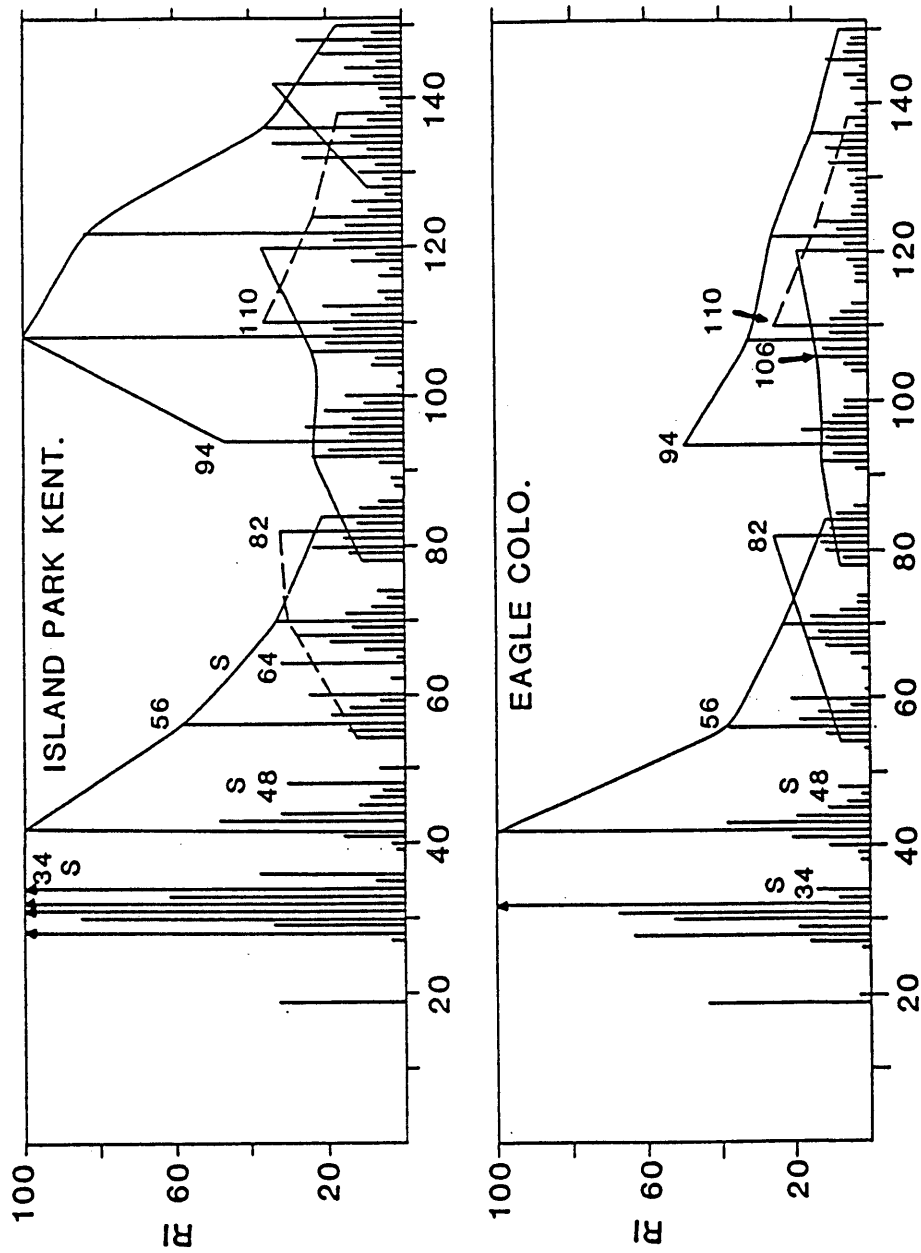


Figure 5. Py-MS of two selected coals. Key: first, Island Park, KY; and second, Eagle mine, CO.

Table V. Distance Table for Various Coals Studied
in Tubing Bomb Reactor (Group 1 coals)

	IC	Con	IC 2	UK	PM	IM	EC
Island Creek, KY (IC)	7.5	46.4	14.2	17.2	29.8	48.5	49.6
Consolidated, WV (Con)		15.7	45.2	37.1	53.0	23.8	53.7
Island Creek-2, KY (IC 2)			5.3	15.2	30.4	47.1	47.1
University of Kentucky, KY (UK)				8.1	29.1	39.8	41.8
P & M, KY (PM)					7.2	50.9	45.2
Ireland Mine, WV (IM)						11.1	58.4
Eagle Mine, CO (EC)							6.1

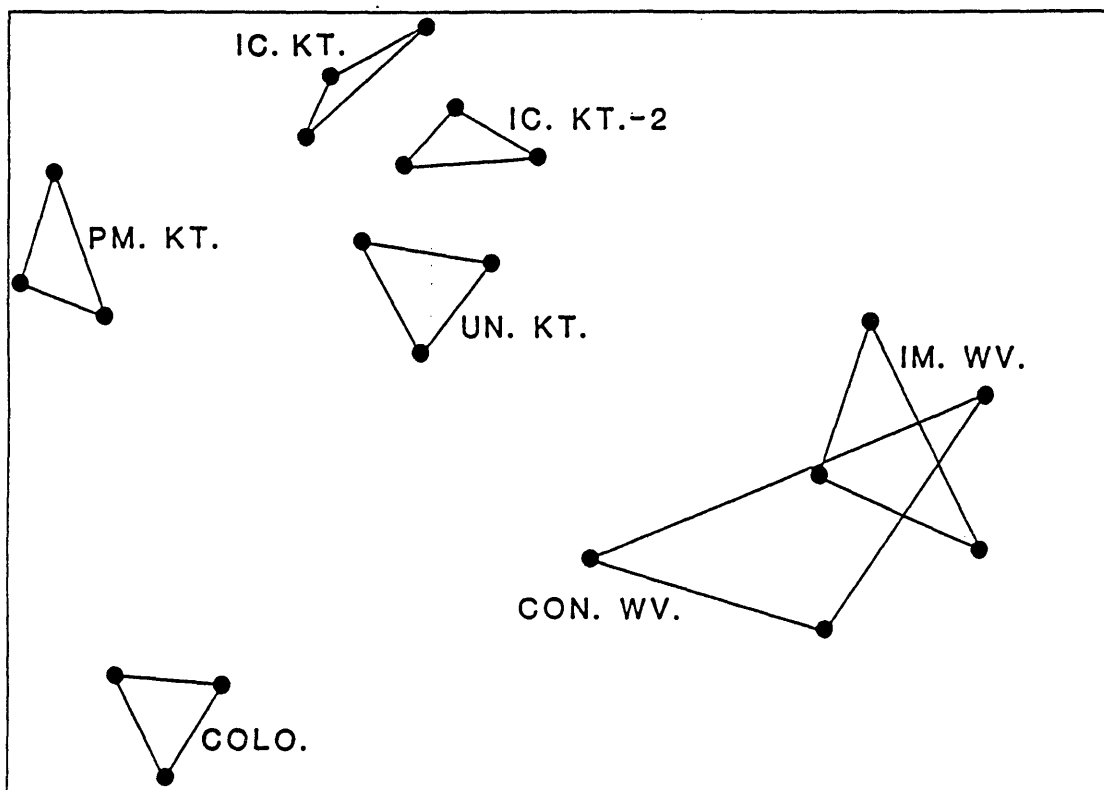


Figure 6. Nonlinear map of Py-MS data from tubing bomb coals.

The points on the nonlinear map clearly cluster into three groups. Comparison of the nonlinear map with the data in Table IV shows that the clustering results from coals that represent three different geographical locations. The conversion data in Table IV indicate that the eastern coals have better conversion characteristics than the single Colorado coal. Based solely on the implications of the geographical differences, the data suggest the possibility of detecting variations in coal reactivity by Py/MS.

The correlation for the various groups represented in the nonlinear map can also be shown by the hierarchical clustering dendrogram in Figure 7. Based on the similarity coefficients, the coals are clearly separated into three groups. The spectra in Figure 5 show the actual mass spectral differences between the best and the worst coal.

The most important observation that results from these data is the fact that the Py/MS/pattern recognition approach is capable of detecting and differentiating organic structural differences in coals. However, based on the limited number of poor

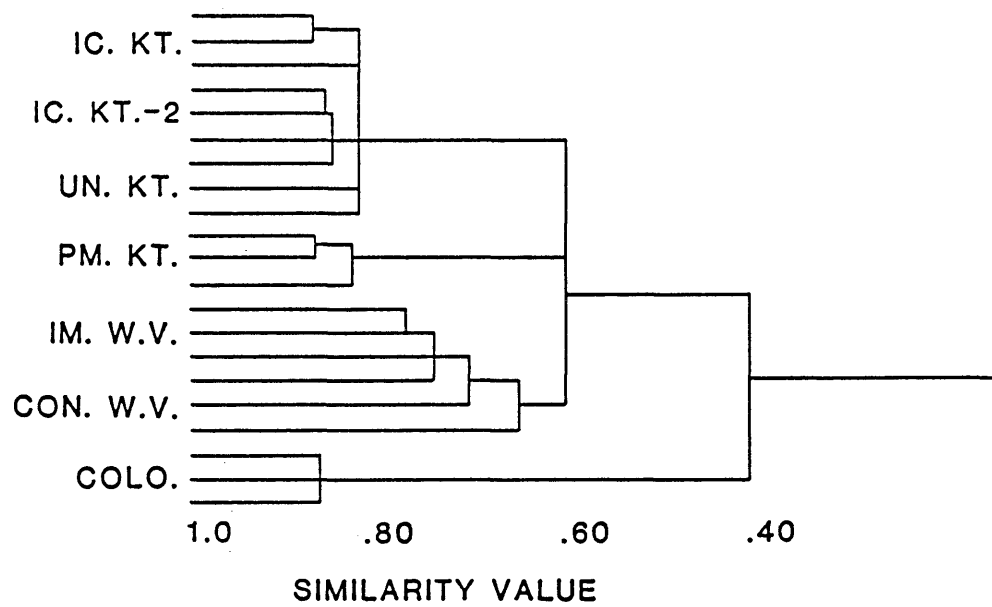


Figure 7. Hierarchical clustering dendrogram of Py-MS data from tubing bomb coals.

reacting coals, it is not possible to state categorically that reactivity prediction has been demonstrated in this suite.

Information concerning the factors that affect the differences between samples within the set of coals can be obtained by comparing the magnitude of the Fisher ratios. Table VI lists the 13 ions (plus possible compound classes) that have the highest Fisher ratios. The Fisher ratio reflects partially the intensity differences between ions. Therefore, each peak with a high Fisher ratio can be assessed as to its positive or negative influence, within the suite, with regard to conversion yields. The group of ions m/z 128, 142, and 156 are probably alkylnaphthalenes and seem to have a positive correlation with reactivity. The ions that appear to correspond to the sulfur species - S_2 (m/z 64) and CH_3SH (m/z 48) - are produced most strongly in the coals with the best conversion performance. The total sulfur analysis for this suite (Table VII) showed a weak relationship between sulfur content and conversion yields, with the highest sulfur content coal exhibiting the highest reactivity.

Table VI. Fisher Weights for Tubing Bomb Coals (Group 3 coals)

Mass	Fisher Weights	Probable Compound Class	Contribution to Conversion
36	57.6	Sulfur compounds	?
156	53.6	C ₂ Naphthalenes	+
60	39.4	Carboxylic acid	?
142	32.9	C ₁ Naphthalenes	+
97	28.0	Alkene	+
128	27.0	Naphthalenes	+
146	25.6	?	+
48	24.4	Sulfur compounds	+
106	23.2	Benzenes	?
157	16.5	?	?
64	14.0	S ₂ Sulfur compounds	+
57	12.8	Alkanes	?

B. GCF Reactivity

Because of the large number of coals studied by Given, the GCF process coals (group 4 in Table III) could be selected more carefully to allow for a wider range of rank and characterization properties as well as reactivity differences. The conversion yield in the continuous flow experiments was defined as solubility in ethyl acetate after a 1 hour residence time (a pseudo-equilibrium reactivity). Table VII summarizes pertinent data for the selected coals.

Table VII. Characterization Data for Gulf Continuous Flow Reactor Coals (Group 4 coals)

PSOC Number	Rank	Conversion		%S
		(%)	%C	(dry)
265	HVA	44	86.7	0.6
271	HVA	49	86.5	---
276	HVA	72	83.5	3.5
278	HVA	83	80.5	5.6
302	HVA	29	88.9	0.7
305	HVB	77	82.3	4.1
307	HVA	77 ^a	82.3	2.7
308	HVB	74	80.3	4.3
320	MV	15	90.4	1.2

^a Measured in a tubing bomb reactor.

The distance table and nonlinear map for the pyrolysis data are shown in Table VIII and Figure 8, respectively. The reactivity for each coal is included, for reference, on the nonlinear map.

The points on the nonlinear map are clustered into two main groups. A comparison between the relative positions and the liquefaction conversion data suggests that a hypothetical line can be constructed through the center of the map to divide the coals into two major conversion classes. The coals above the dividing line exhibited conversion yields less than 50 percent, while those below the dividing line had yields greater than 50 percent. In addition, a very striking feature of the map is the fact that Coal 320 had the lowest conversion reactivity, while Coal 278 had the best

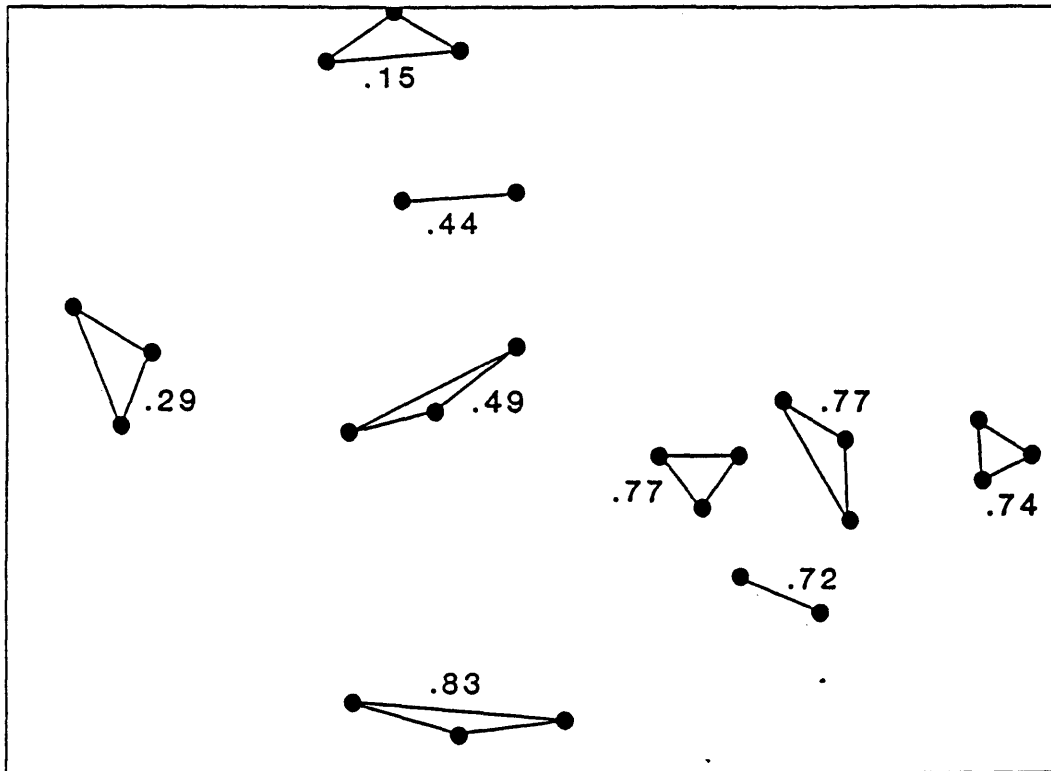


Figure 8. Nonlinear map of Pennsylvania State University coals. Numbers indicate percent conversion yields.

liquefaction properties. The positions of these coals represent the extremes on the nonlinear map. The same type of clustering, based on 50 percent reactivity, is shown in the hierarchical clustering dendrogram in Figure 9.

The catalytic effect of mineral matter is an important factor in controlling liquefaction properties (Berkowitz, 1979). The selection of the inorganic catalysts employed in liquefaction processes has always been the most effective chemical method for controlling the product. Often, the contribution of the organic structure to liquefaction conversion has been ignored. The Py/MS results for the Gulf continuous flow reactor coals present evidence that more emphasis should be placed on the organic portion.

Fisher ratios can be used within a single suite of coals to determine positive or negative effects of homologous ion series on conversion yields. Table IX is a list of the 13 greatest Fisher weights for this set of coals. For the coals liquefied using the Gulf continuous flow reactor, a series of bivariate plots using homologous series that were identified in the top 50 percent of the Fisher ratios was constructed. For

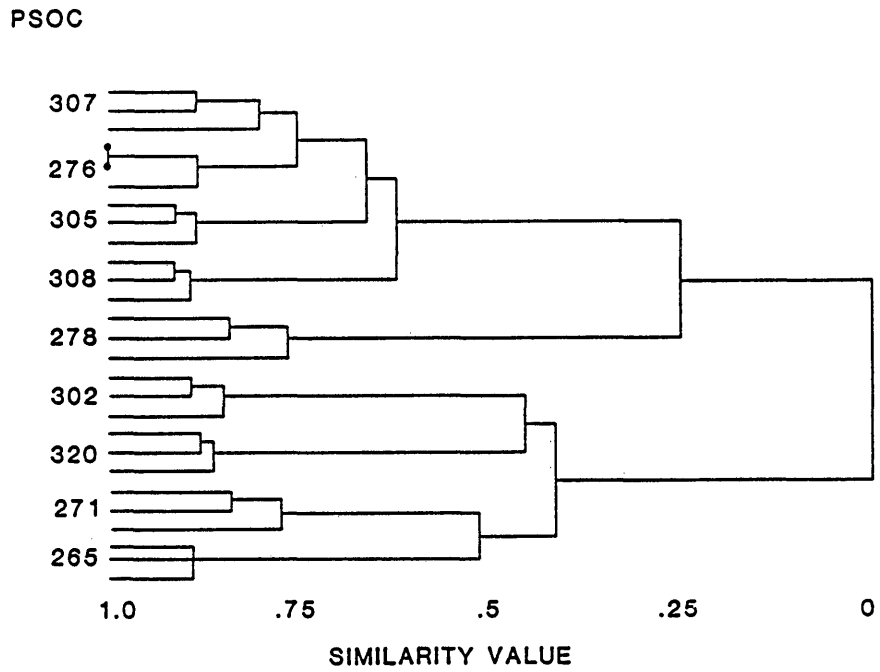


Figure 9. Hierarchical clustering dendrogram of Py-MS data from Pennsylvania State University coals.

example, a plot showing the intensity of phenol vs. methylphenol, m/z 94 vs. 108, is shown in Figure 10. In this case, the highest liquefaction reactivity was observed in the coals with highest intensity for these homologs. This observation is consistent with the reported variation of liquefaction reactivity with total oxygen content of coal (Furlong, 1981). Other series such as the sulfur species - m/z 34, 48, 64, and 76 - and the series at 148, 162, and 176 - may be plotted to give similar trends. In addition, a bivariate plot can be made from members of two different homologous series. Figure 11 illustrates a plot of m/z 83 vs. m/z 85 that represents the intensities of $C_6H_{11}^+$ (alkene fragment ion) vs. $C_6H_{13}^+$ (alkane fragment ion). From this plot, coals with the highest concentration of these ions showed the poorest liquefaction performance. The good correlation of the alkene:alkane signals is somewhat surprising because a broad spectrum of alkane yields had been reported for the liquefaction of various ranked coals (Whitehurst et al., 1980). Once the trends have been established by a multivariate statistical calculation, the complexity of the data analysis and conversion correlations often can

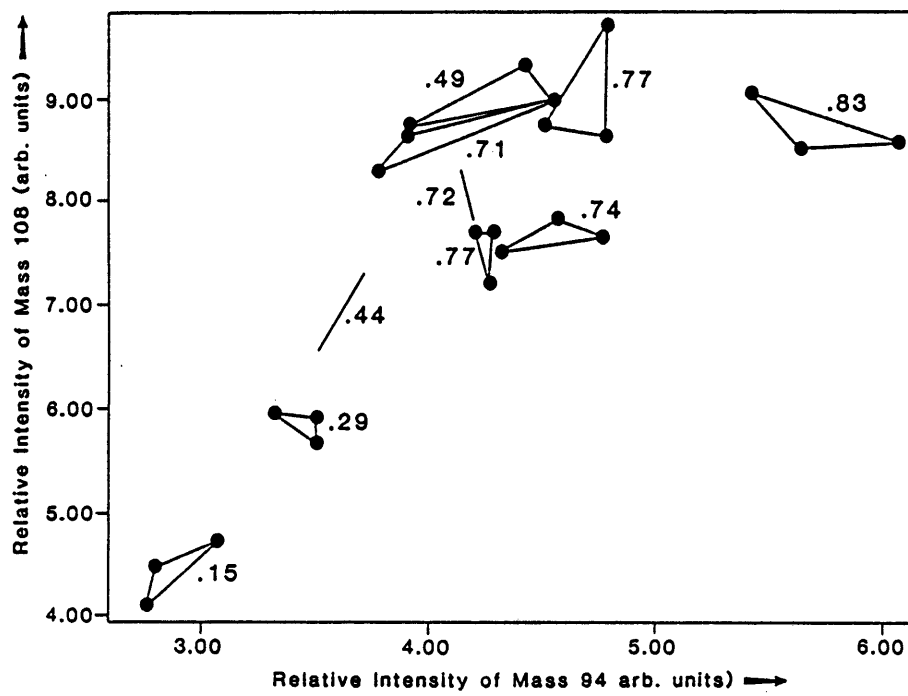


Figure 10. Bivariate plot of m/z 94 vs. 108.

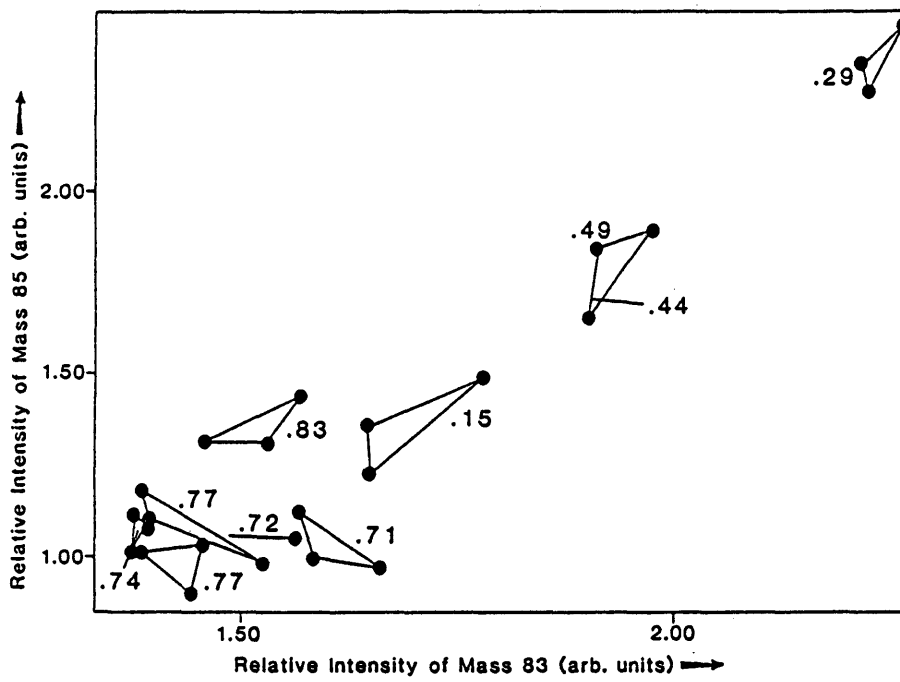


Figure 11. Bivariate plot of m/z 83 vs. 85.

be reduced by the use of a bivariate statistical approach. However, bivariate techniques are no substitute for multivariate analysis.

Table IX. Fisher Weights for Gulf Continuous Flow Reactor (Group 4 coals)

Mass	Fisher Weights	Probable Compound Class	Contribution to Conversion
176	213.1	?	+
92	144.2	Benzenes	-
36	143.8	Sulfur compounds	?
85	139.1	Alkanes	-
41	115.9	Alkenes	?
64	84.4	S ₂ Sulfur compounds	+
124	83.2	Lignin	?
83	79.7	Alkenes	-
128	73.0	Naphthalenes	?
207	72.3	?	?
156	61.3	C ₂ Naphthalenes	+
162	55.3	?	?

C. Stirred Bomb Reactivity

Another group of coals investigated represented a suite that had been studied at Colorado School of Mines in a stirred batch reactor. Reactivity data for the coals were based on tetrahydrofuran (THF) solubility after a 60 minute residence time in a quick charge batch stirred autoclave. Table X lists the characterization properties of these coals.

Table X. Coals Used in Stirred Batch Reactor (Group 2 coals)

PSOC Number	Rank	Conversion ^a (%)	%C	%S (dry)
071	HVCB	83.7	78.0	0.58
107	HVBB	81.0	82.0	0.53
130	MVB	<60.0	91.4	0.56
151	HVCB	80.0	78.4	0.47
370	HVAB	75.2	84.0	0.67
437	HVAB	85.0	80.3	0.51
444	HVBB	85.8	78.9	0.46
Fies (KY 9)	HVCB	87.5	66.9	3.85

^a Conversion defined by product solubility in THF after 1 h reaction time at 400 degrees C, 2000 psi (H₂). Liquefaction solvent was tetralin.

The data analysis results for the Py/MS data are summarized in Table XI and Figure 12. The points on the nonlinear map (Figure 12) basically cluster into one major group with two outlying sets of points. Comparing the position of the data points on the nonlinear map to the reactivity data in Table X presents a picture consistent with that derived from the two previous coal suites. In general, the major cluster comprises coals that show conversion yields greater than 75 percent. The first set of outlying points (Fies Mine) above the major group shows good reactivity (87.5 percent). However, a review of the history of the sample indicated the possibility of

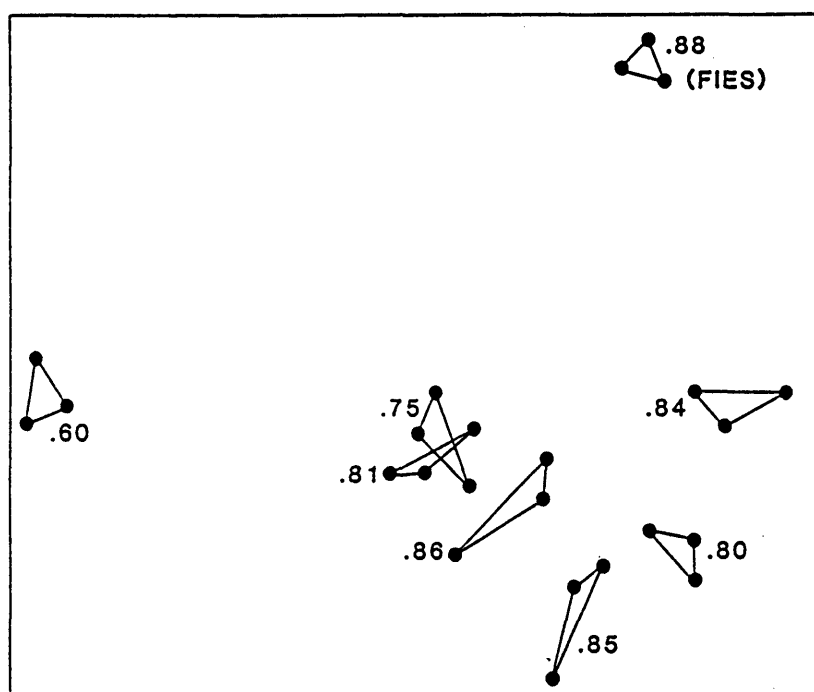


Figure 12. Nonlinear map of Py-MS data of CSM coals.
Numbers indicate percent conversion yields.

weathering effects. No special precautions had been taken for storage of this sample; therefore, it had been in contact with air for several years.

Table XI. Distance Table for the Coals Studied in the Stirred Tank Reactor (Group 2 coals)

PSOC	071	437	444	151	370	Fies	130	107
071	26.2	72.6	79.1	62.9	84.7	139.9	177.9	83.8
437		34.3	53.5	42.7	66.4	187.3	163.5	70.4
444			33.5	56.2	43.5	176.2	150.5	40.9
151				20.5	70.9	168.4	167.3	68.8
370					30.3	175.8	127.0	31.9
Fies						15.5	239.3	175.8
130							20.0	137.1
107								22.0

Examination of the data in Table X shows an approximate conversion yield was measured for the PSOC 130 coal. This coal, when subjected to the standard conditions of the stirred batch liquefaction procedure, agglomerated to such an extent that reproducible data of the quality of the other measured reactivities in the suite could not be obtained. The Py/MS results suggest significant structural differences in this particular coal. The role of weathering is an uncertain one, with possible significance for industrial-scale liquefaction. Therefore, the EXXON

coal experiments described later included a test of the effect of weathering on the liquefaction prediction equations.

The data from the Py/MS strongly indicate that liquefaction conversion information, relating to tubing bomb reactivity, GCF reactivity and stirred batch reactor reactivity, is present within the pyrolysis spectra. The Py/MS/pattern recognition results also generate information that can correlate structural features with the conversion data. Subsequent chapters will describe quantitative methods which enable the prediction of reactivity in these processes with high accuracy, cost effectiveness and rapidity.

IV. SUPERVISED METHODS

The experiments and data analysis methods of the previous chapters demonstrated the fact that a prominent feature of the Py/MS spectra of coals is information relating to liquefaction reactivity. Fisher ratioing indicated peaks which might be related to reactivity. With the use of a much larger data set and more sophisticated data analysis methods, two goals will be accomplished for the tubing bomb and GCF liquefaction processes. 1) Accurate equations will be developed for the prediction of reactivity in both processes, and 2) the identity and amount of the pyrolysis products which positively or negatively correlate with reactivity will be determined.

A large set of coals were available for this set of experiments. They were pyrolysed in a carefully designed study so that a library of spectra, many of which were well-characterized in terms of conventional characteristics, could be constructed.

A. Experimental Method

1. Samples

Forty-seven PSOC coals, some from the preliminary experiments, were analyzed. Four of the coals were received in duplicate in separate shipments, and were treated as separate coals in the analysis and calculations which follow. Thus, a total of 51 coal samples were studied. The coals in the study were selected based on availability from Pennsylvania State and the fact that Given et al (1979) and Yarzab et al (1980) had reported a wide range of reactivities in both the tubing bomb and Gulf continuous flow reactors. In addition, a variety of analytical data for these coals has been reported. Kinetic reactivity data were available for twelve of the coals. It was desirable to obtain comparable mass spectra for all of the coals; however, reactivity and other data are only available for subsets of the 51 samples. For instance, Yarzab et al reported tubing bomb reactivities for 26 of the coals, while Given et al reported Gulf Continuous Flow reactivities for only 11. Percent carbon was available from various sources for all but 1 of the coals, and so on.

The samples, received from Pennsylvania State in sealed cans, were opened in a dry-box under an Ar atmosphere and transferred to labeled bottles. Since one of the goals of this study was to provide a rapid, practical technique for predicting reactivity during liquefaction, elaborate precautions were not taken to prevent mild air oxidation of the coals by the atmosphere once the sealed bottles were opened.

As in the previous experiments, suspensions with a concentration of 5 mg/mL of the coals in methanol were prepared. No appreciable settling of the coal particles was observed in any of the suspensions within the length of time required to withdraw a representative sample. On the second and subsequent days, however, two of the coals, PSOC 284 and PSOC 399, exhibited a pronounced tendency to flocculate. This was relieved by ultrasonication. These two coals were reground at two day intervals during the course of the study. Four other coals also tended to flocculate to a lesser extent (PSOC 580, PSOC 217, PSOC 306 and PSOC 280). In addition, suspensions of PSOC 056, PSOC 577 and PSOC 130 developed a brown supernatant during the course of the study. Since it was not known whether

these changes were purely physical or chemical, new suspensions of all the coals were prepared from the previously ground samples midway through the study. Eight of the coals, PSOC 217, PSOC 280, PSOC 284, PSOC 306, PSOC 399, PSOC 401, PSOC 580 AND PSOC 607, were reground every two days in order to determine whether detectable chemical changes were occurring in the coals as a result of being suspended in the methanol.

2. Collection of Data.

Curie point wires were loaded with approximately 50 micrograms of coal and pyrolyzed at 610 degrees C. The 610 degree C wires were chosen because they provided the most significant contribution from higher masses. In order to minimize fragmentation, the pyrolysed molecular fragments were ionized by 14 eV electron ionization. The spectra used for data analysis were the sum of 100 repetitive scans in the mass range 30-240.

In order to control both systematic and random changes in the instrument's tuning, the following pyrolysis schedule was adopted: Each of the 51 coals was pyrolysed once each day for eight days. The coals

were pyrolysed in a different, previously determined random order each day. To minimize short-term increases in background, and possible changes in sensitivity due to effects of the large number of samples being processed, the coals were pyrolysed in small sets (typically six at a time) with an approximately 15 minute pump-down period between sets. At about four hours into each run, the instrument was allowed to pump down for at least 1.5 hr. The mass spectra were processed using the statistical packages ARTHUR and SPSS. Supplementary programs, described in Appendix 1, performed the factor rotations and graphic displays.

Every possible care was taken to apply a uniform and reproducible quantity of suspension to the Curie-point wires, yet the total ion intensity was seen to vary by an order of magnitude. Possible sources of this variation were mentioned in the discussion of normalization in the INTRODUCTION. Eight replicates were collected with the idea of casting out samples with extreme values of total ion intensity. Yet, except for a few peaks known to be prominent in the background, plots of total ion intensity versus

selected masses, factors produced by factor analysis and values for chemical and physical properties of the samples showed no trends related to total ion current. Apparently, after normalization, variation due to quantity of pyrolysate is negligible relative to other factors (such as true differences between coals and order within a set), so none of the spectra were deleted.

The mass spectra were reduced to mass/intensity data and normalized to the sum of the peaks in the mass range 72 - 160 (for a justification for selecting this mass interval, see INTRODUCTION). After normalization, the spectra were autoscaled.

B. Prediction and Evaluation

Factor analysis was performed using ARTHUR. Of the 30 factors extracted, 20 factors, all having eigenvalues greater than 1.3, were retained for further analysis using SPSS. These 20 factors contain 72 percent of the variance present in the original data set. The question of how many factors to retain has been discussed at length (Duewer et al., 1976 and Harper et al., 1977). In this study a relatively large

number of factors was kept in order to reduce the risk of discarding meaningful information at the possible expense of retaining some of the purely random error. The factors in the range from 10 to 20 were shown to be meaningful from the fact that plots of these higher factors discriminated between coals. The large number of factors retained is reasonable, since highly complex mixtures such as coals can vary in a large number of statistically independent ways. On the other hand, 20 independent factors is a small enough number that statistical techniques are appropriate for a large sample size.

1. Tubing Bomb Reactivity. Since the most abundant data were available from the tubing bomb reactor study (Yarzac et al., 1980), it was chosen to test the predictive reliability of linear modeling. Various methods have been proposed for selecting the specimens which are to be used in a training set (Geisser, 1975, Snee, 1977 and McCarthy, 1976). In this case, half of the eastern and interior coals were selected at random, since random selection requires no prior knowledge of the quantity being predicted for the

test set. Therefore, the test set measures the behavior of true unknown samples. A randomly selected subset consisting of all eight replicates of PSOC numbers 217, 271, 281, 284, 305, 341, 345, 357, 375, 399, 401 and 580 was extracted from the 26 coals for which reactivity data were available. These 96 spectra were used to generate a predictive model for reactivity, using the SPSS multiple linear regression routine NEW REGRESSION.

The regression technique used was stepwise regression (Draper and Smith, 1981). Starting with one variable, regression equations are calculated in succession. Each succeeding regression equation includes an additional variable (factor), which is added based on its ability to improve the fit, until the addition of a variable does not significantly improve the model. This occurred when the twelfth factor was added. For interest, simple multiple regression was also performed using all 20 factors in the equation. Figure 13 shows the behavior of the sum of the squared residuals. As expected, the error decreases as more factors are added for the coals used in the model (the "knowns"). At first the decrease in

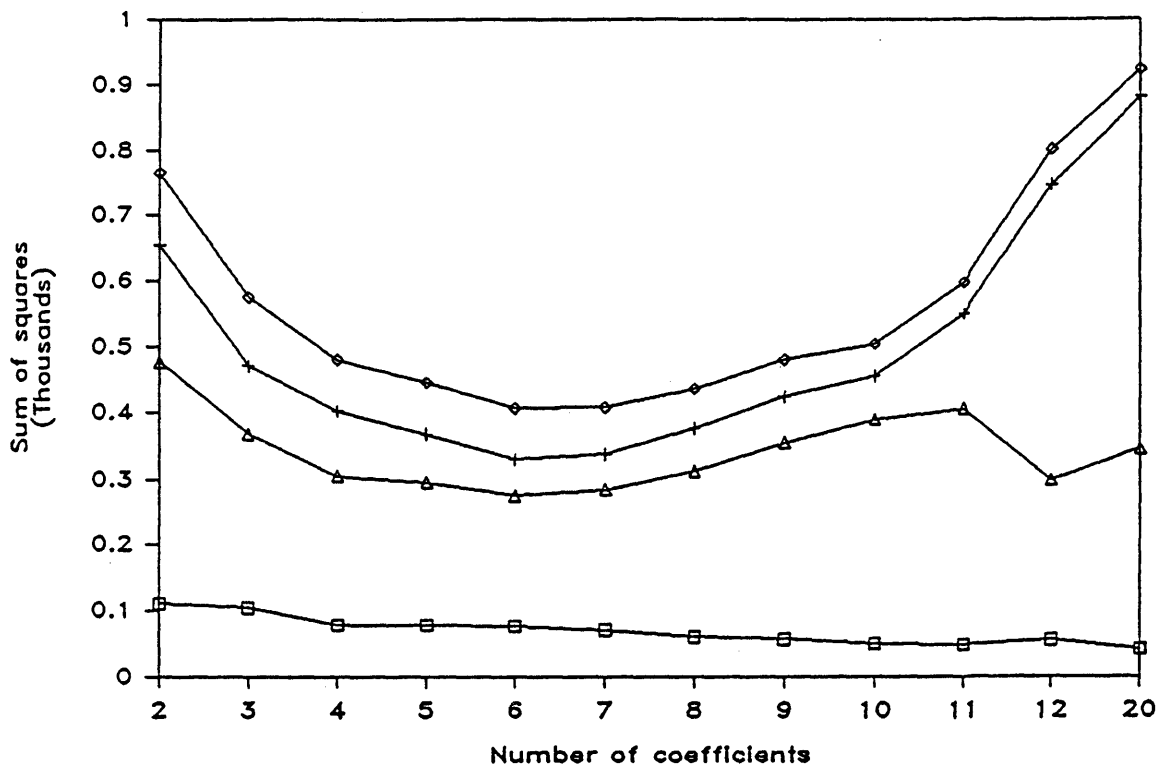


Figure 13. Accuracy of predicting "unknown" reactivities as coefficients are added through stepwise linear regression: (□) knowns, (+) unknowns, (◇) total; (Δ) unknowns without western coals.

error is caused by genuine improvement in the model by adding more factors. Eventually, however, additional factors are correlated by chance, and including them in the model actually degrades the prediction of unknowns. As Figure 13 indicates, the optimum number of factors is about 6, after which the model becomes fitted to the error in the training set. Based on the model which includes 6 factors, a predicted reactivity, the average of the predictions for each of the eight replicates, was calculated for all of the PSOC coals included in this study. The predictions are recoded in Table XII, along with the measured values, when available. Figure 14 shows the relationship between this predicted reactivity and the reactivities reported by Yarzab et al. (1980). Interestingly, two coals, which were described as outliers in the paper of Yarzab et al., PSOC 281 and PSOC 592, were predicted within 4 percent by this model.

Several studies have described chemical similarities between coals which correspond for the most part with coal province. Yarzab et al found that the coals they studied separated naturally into groups 1, 2 and 4, (labeled groups Y1, Y2, and Y4,

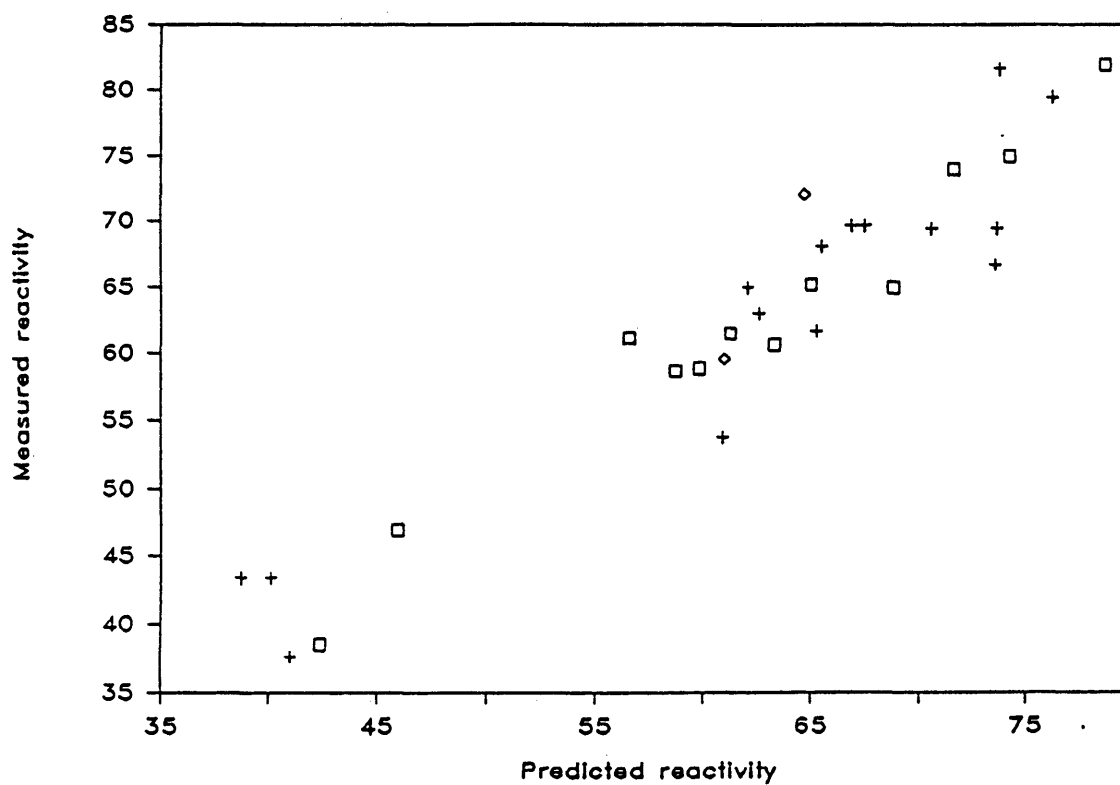


Figure 14. Prediction of reactivities with six coefficients: (\square) knowns, ($+$) unknowns, (\diamond) unknown western coals.

respectively, to distinguish them from the groups of coals described in this thesis). The groups Y1, Y2 and Y4 corresponded primarily to Eastern, Interior and Rocky Mountain coals, respectively. In this set of experiments, group Y1 coals and group Y2 coals were isolated and used to develop separate predictive models.

Figure 15 shows the effect when group Y1 coals are used to predict group Y2 coals. The prediction is acceptable, but not as accurate as the model derived from the random sample. Figure 16 displays the prediction of group Y1 coals based on a model derived from group Y2 coals. Again, the prediction is not as accurate as that using the random sample. The linear trend successfully extrapolates to several group Y1 coals which fall well outside the range of reactivities used to develop the model. Oddly the models based on only group Y1 and only group Y2 coals fail to predict the reactivity of western (group Y4) coals, but the model based on both is successful. A possible explanation is that a model based on only one group of coals is more accurate, but lacks generality. The

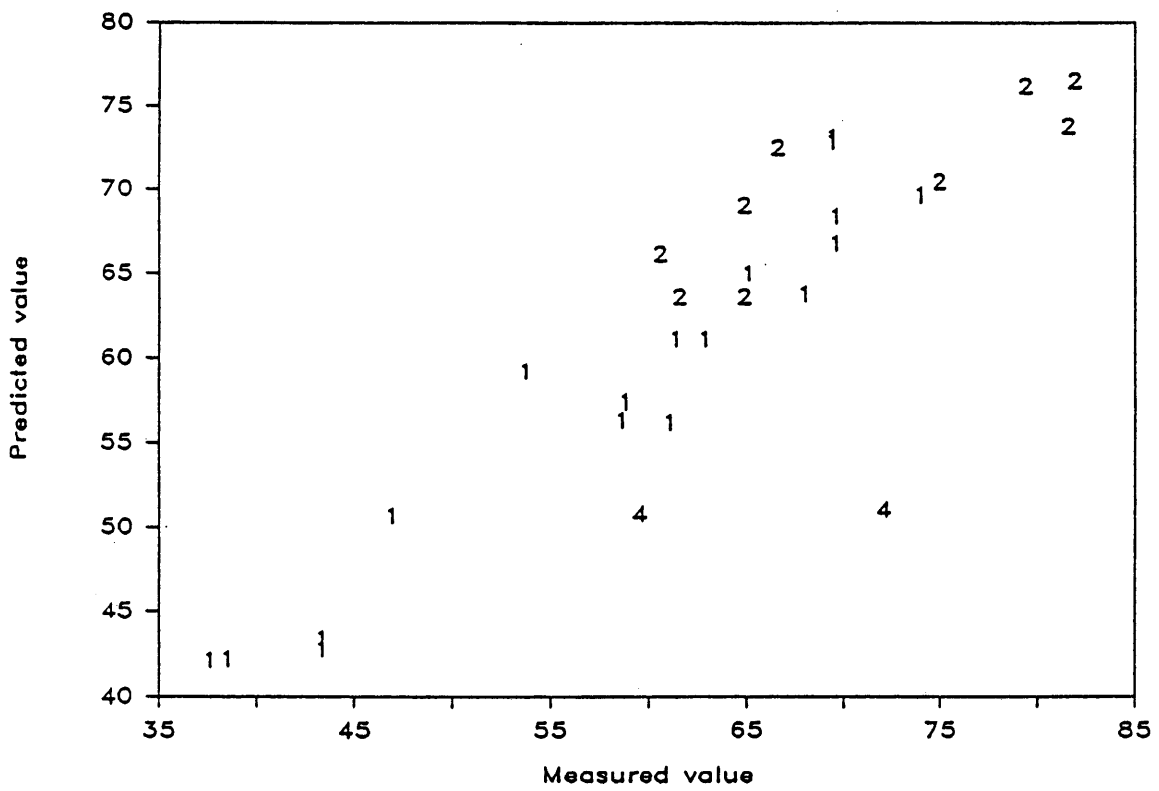


Figure 15. Predicted reactivity of group Y2 coals from a model based on group Y1 coals (group numbers shown on plot).

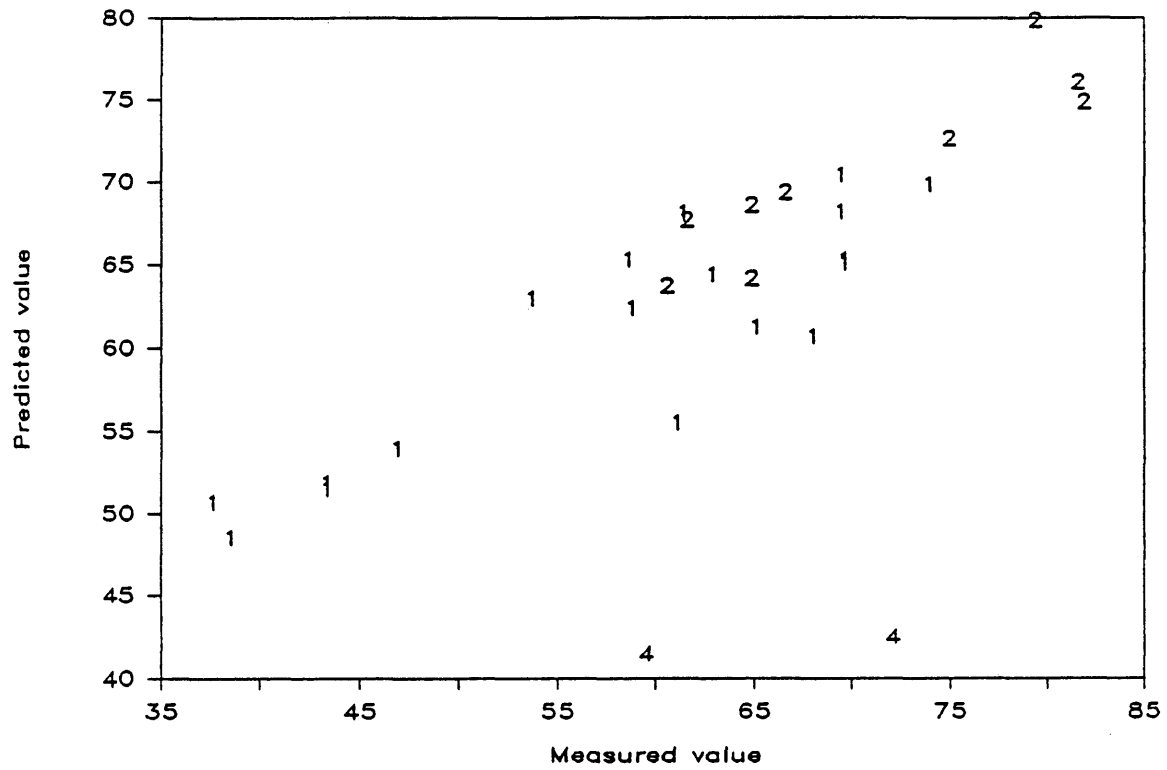


Figure 16. Predicted reactivity of group Y1 coals from a model based on group Y2 coals (group numbers shown on plot).

small number of western coals with known reactivities makes this hypothesis impossible to test.

Figure 17 is a loading spectrum for percent conversion in tubing bombs (it differs from a factor spectrum because the loadings were not multiplied by their corresponding standard deviations). Although this plot was generated based on the randomly selected subset of coals described earlier, it differs in only minor respects from plots based on only group Y1 and on only group Y2 coals. In other words, the correlations with mass spectral peaks are quite similar for group Y1 coals, group Y2 coals, or a random selection from both. The same is true of the other loading and factor spectra displayed in this chapter; therefore, a high degree of confidence may be placed in the correlations, since nearly identical conclusions may be drawn from completely different subsets of the coals.

From Figure 17 it is clear that liquefaction is very complex. Homologous series differing by one CH_2 unit (m/z 14) tend to be similarly correlated. In a few cases, these homologous series can be identified with a fair degree of confidence, as in the case of masses 94, 108, 122, ..., probably due predominantly to

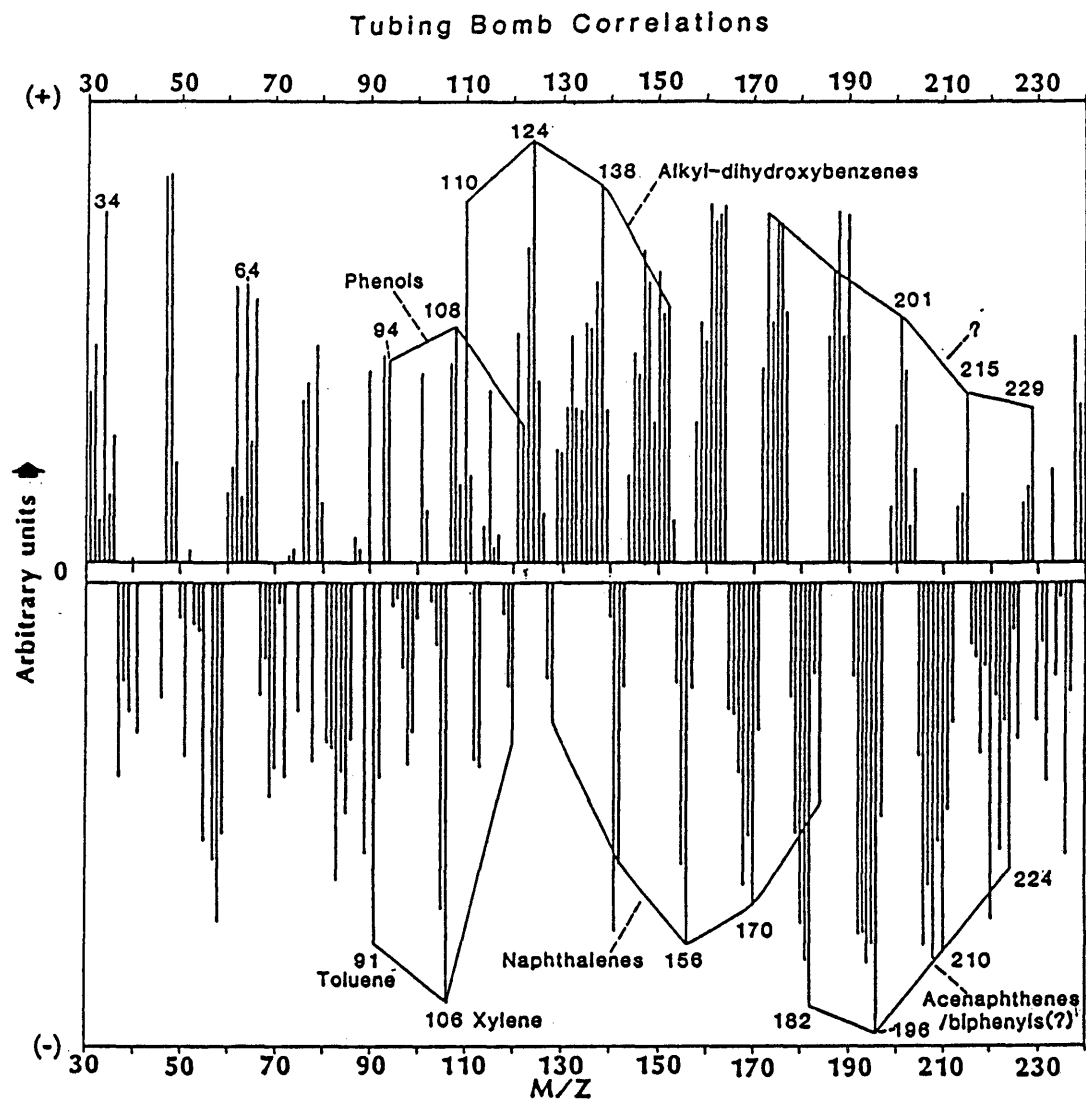


Figure 17. Correlation of mass spectral peaks with tubing bomb conversion.

alkyl-substituted phenols or alkyl-aryl ethers. Other series are ambiguous, such as the prominent series 182, 196, ..., which may be due to alkyl substituted acenaphthenes or biphenyls (Meuzelaar et al., 1984b), tetrahydrophenanthrenes, unidentified components, or a combination of them all. Still other sequences of correlated peaks cannot be identified confidently. The identification of the homologous series was obtained by selecting components which seem to be reasonable structures in a coal. Most of these masses have been similarly assigned in the literature. (Larter and Douglas, 1982; Meuzelaar et al., 1984, Larter and Douglas, 1978; and Larter and Douglas, 1980). At this point, the identities of the peaks must be considered tentative, since numerous other compounds may also occur at the same masses. Of course, the peaks do not in general represent whole molecules in the original coal, but are mostly pyrolysis fragments of larger coal molecules. As such, they must be interpreted as structural components of the original coal molecules. It is these original, generally high molecular weight components which are directly correlated. On the other hand, the pyrolysis process resembles the

liquefaction process enough that many of the pyrolysis fragments observed are probably the same species which would be present during liquefaction. Peaks in the positive direction on the loading spectrum may be interpreted as having a positive correlation with reactivity, and negative peaks as having a negative correlation.

In most cases, it is desirable to know not only whether a given component is correlated with the property of interest, but also whether it is present in sufficient abundance to be important. Part of the difficulty in interpreting Figure 17 is that the sensitivity of Py/MS is sufficient that correlations may be found between peaks which are quite small. Ultimately, it is desirable not only to model reactivity, but to reveal the importance of the peaks in the mass spectrum with respect to reactivity. But the smallest peaks may be highly correlated without significantly affecting reactivity. In an attempt to include information about both size and correlation, the loadings, (coefficients in Equation 17 of Chapter I. INTRODUCTION, section D. Pattern Recognition Background), are often multiplied by the standard

deviations which were removed in autoscaling (Windig et al., 1981). The resulting "factor spectra" resemble conventional mass spectra, which are, after all, plots of abundance. Figure 18 is a plot of the data in Figure 15 multiplied by the standard deviations. Masses 34 and 64, not shown on Figure 18, illustrate one of the difficulties in interpreting this sort of plot. H_2S and S_2 were present in such great abundance that they would have been many times larger than any of the peaks in Figure 18, although they were only moderately correlated with reactivity, as Figure 17 shows. Therefore, loading spectra and factor spectra offer complementary types of information, and the best interpretation requires an understanding and comparison of both.

Multiplication by the standard deviation creates the danger that the linear model is no longer valid for the property it models. One way to demonstrate validity is to plot the scores for each of the samples used in the model without standard deviation multiplication versus the scores after multiplication. Then, if the standard deviation multiplied factor scores are linearly related to the untreated factor

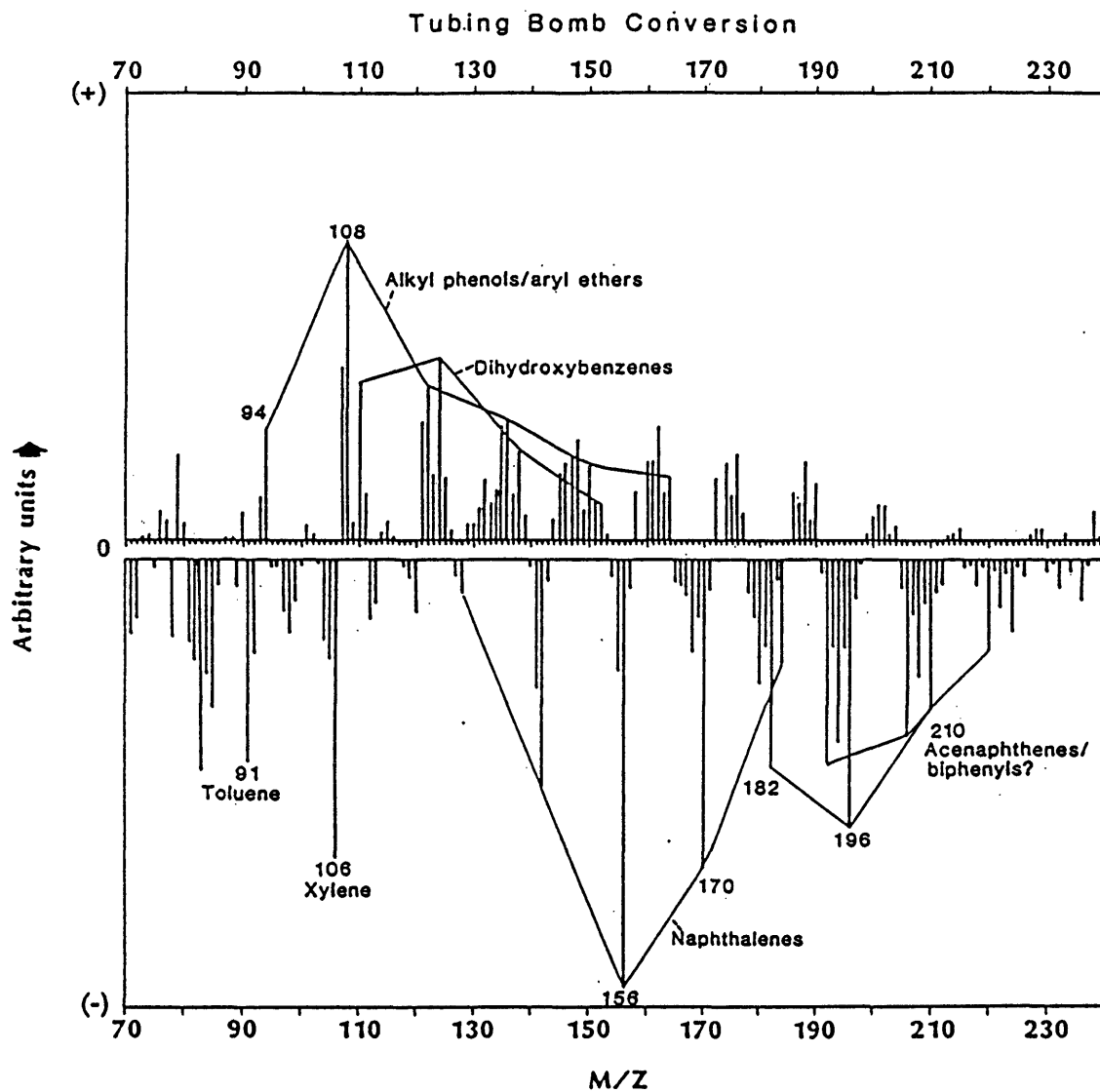


Figure 18. Factor spectrum for tubing bomb conversion.

scores, they are also linearly related to the property they are used to model. Figure 19 shows that this is the case for the liquefaction reactivity prediction.

Figure 20 shows the factor spectrum for vitrinite reflectance (rank) in the tubing bomb reactor coals (Yarzac, et al., 1980). Figure 21 shows that the scores without standard deviation multiplication have a linear relationship with the scores after standard deviation multiplication. The large peaks due to alkyl phenols, alkyl naphthols and dihydroxybenzenes in the negative direction correspond with the familiar loss of oxygen as water and CO₂ with increasing geochemical age (Tissot and Welte, 1978). The positive peaks due to naphthalenes correspond to increasing condensation and aromatization. Similarly, the higher molecular weight series 192, 206, 220 probably corresponds to alkyl phenanthrenes, and the series 182, 196, 210, 224 probably corresponds to acenaphthenes, biphenyls and/or tetrahydrophenanthrenes. Unfortunately, the quadrupole mass spectrometer used was not capable of scanning higher masses with sufficient sensitivity to reveal the higher molecular weight, polycyclic aromatic products also expected in the positive direction. This factor

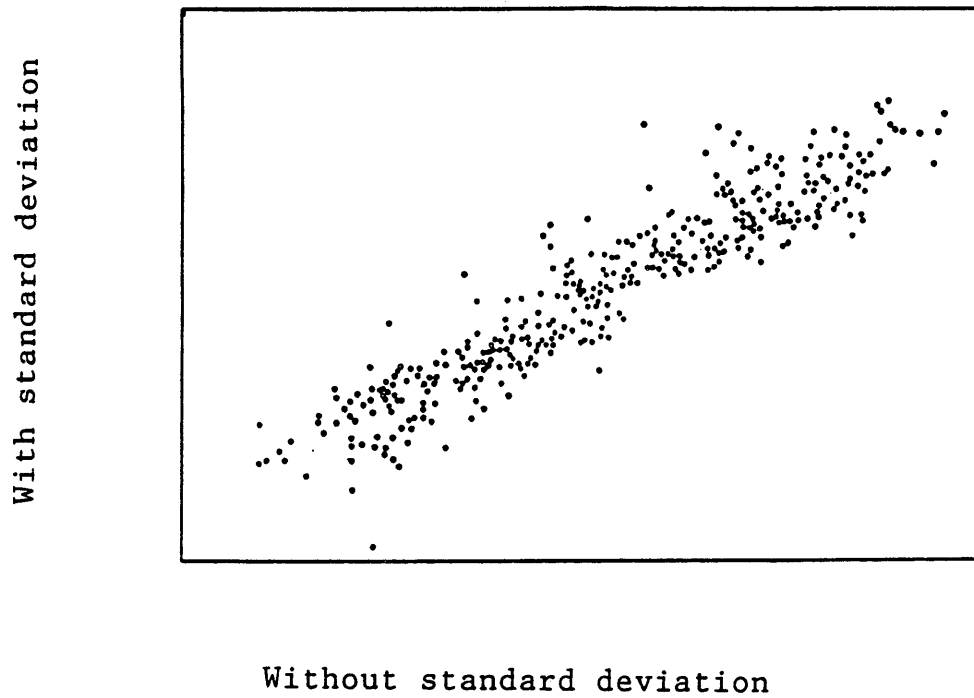


Figure 19. Comparison of predictions for tubing bomb reactor reactivity with and without standard deviation multiplication.

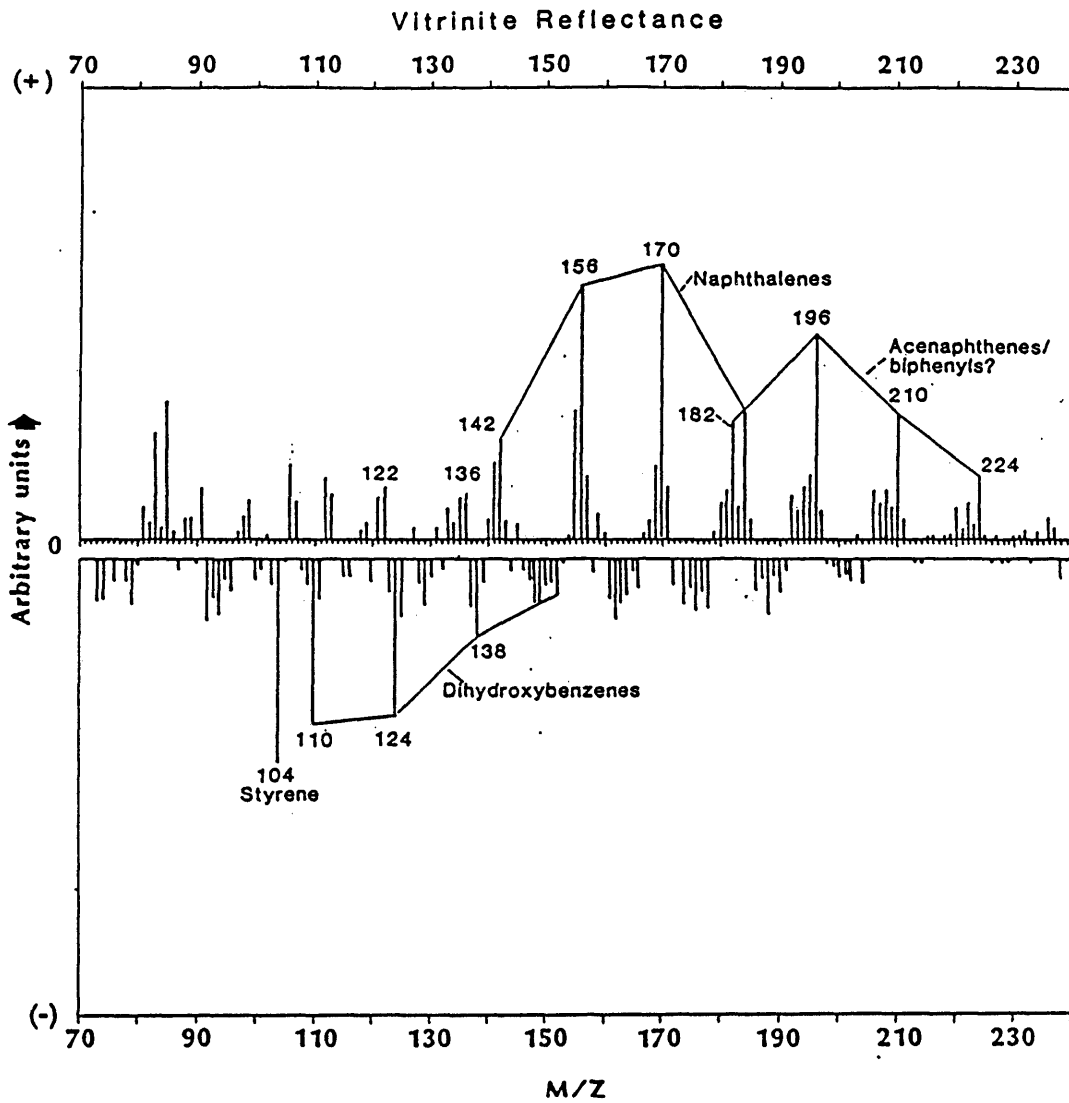


Figure 20. Factor spectrum for vitrinite reflectance.

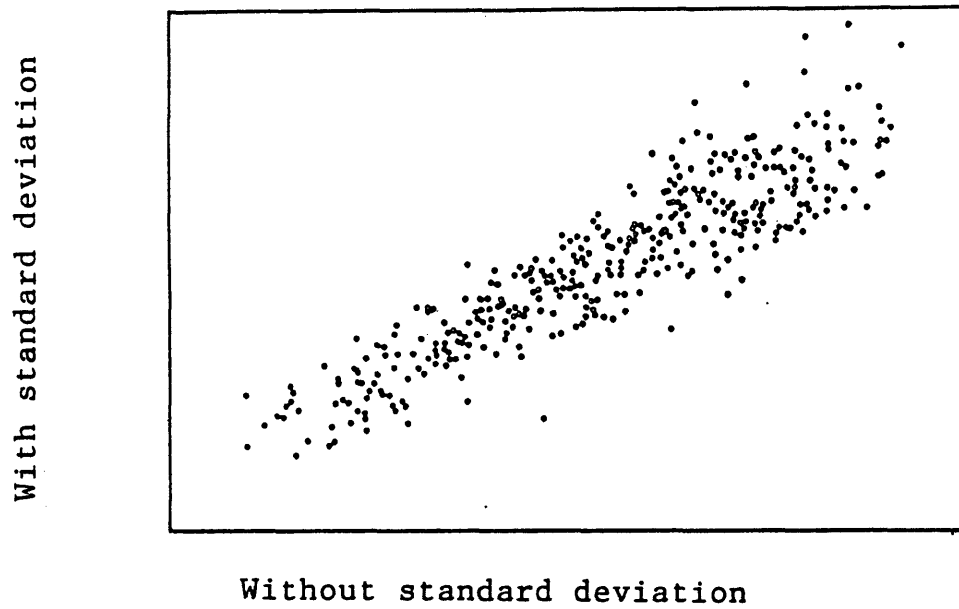


Figure 21. Comparison of predictions of rank with and without standard deviation multiplication.

spectrum is consistent in general respects with similar plots for other suites of coals (Meuzelaar et al., 1984 and Voorhees and Durfee, 1983).

A comparison of Figures 18 and 20 reveals that they are opposite in many ways. Most large peaks in the vitrinite reflectance spectrum are also large, but opposite in sign from peaks in the reactivity spectrum. This reflects the well-established negative correlation between rank and reactivity (Figure 22). Correlation between two measurements does not imply that one causes the other, and therefore it cannot be concluded from Figure 18 that, for instance, large quantities of alkyl naphthalenes are detrimental to liquefaction reactivity. The factor spectrum is an accurate reflection of the pyrolysis products from the coals selected for this study which may be used to estimate reactivities. It contains a large component due to rank, but includes other components as well.

An interesting result was found when the alkyl phenol/aryl ether series were compared. In this group of coals, the entire sequence of alkyl phenols/aryl ethers was not found to correlate uniformly with rank. Peak 94 was negatively correlated, 122 and 136 were

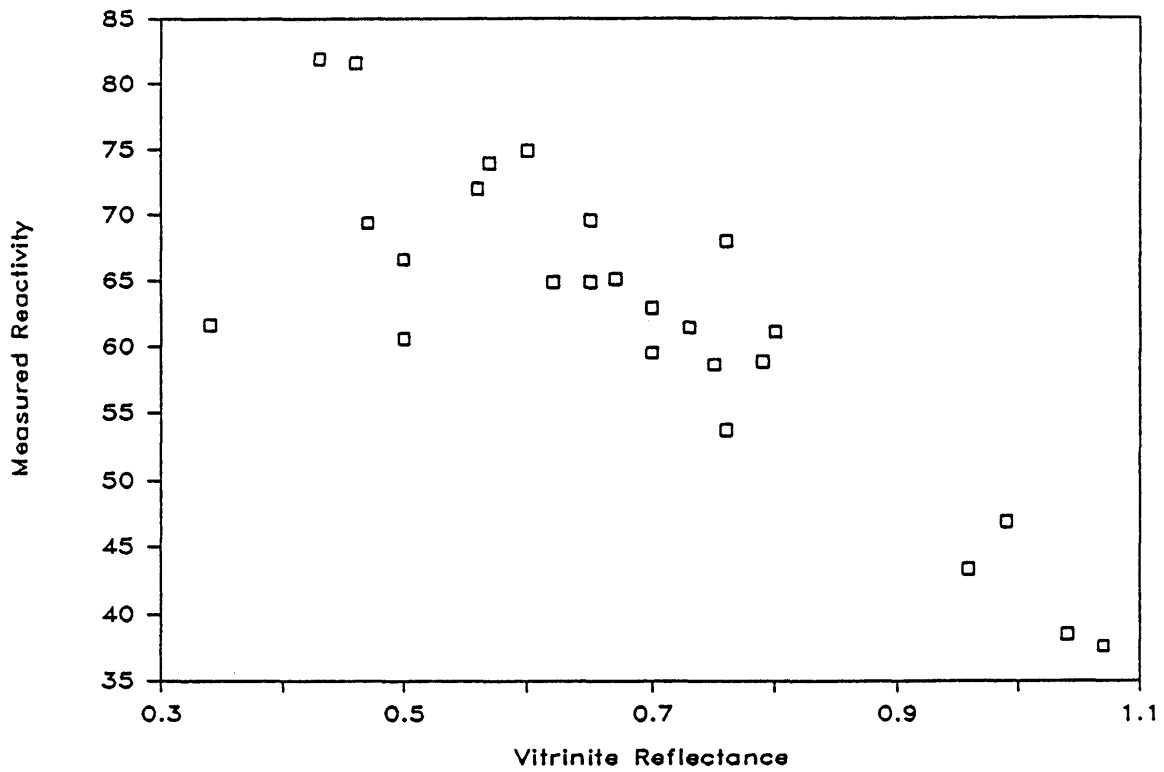


Figure 22. Correlation between rank and conversion in tubing bombs.

positively correlated, and the rest were not prominently correlated in either direction. This series was, however, correlated positively with liquefaction reactivity. This implies that alkyl phenols and/or aryl ethers are related to liquefaction yield independent of rank, assuming that these pyrolysis fragments are also the active agents during liquefaction. Some evidence suggests that aryl ethers or phenols are directly involved (Abdel-Baset, M. et al. 1978; Schlosberg et al., 1983; and McClennen et al., 1983). In the Gulf continuous flow reactor (Given et al., 1979) the four coals with the lowest conversions also had the lowest evolution of CO, CO₂ and H₂S. This relationship is not unique to the coals selected for this study, but is consistent throughout the coals studied by Given et al.

The complex behavior of coals liquefied in tubing bombs can be accounted for using a surprisingly small number of coals. Although there are a large number of statistically independent ways in which the coals may vary, there are evidently only a few ways which relate to reactivity for this set of coals. It was found unnecessary to distinguish between Eastern and Interior

coals for the purpose of prediction. The two western coals were accurately predicted from a model derived from a set of eastern and interior coals. This is surprising, since western coals are dramatically different in age, source material, and geochemistry. Since only two western coals with known reactivity were available, it is possible that these coals were accurately predicted by chance. If the two coals are representative of western coals, however, then liquefaction behavior may be controlled by factors which are much simpler than one would expect.

2. GCF Reactivity. Due to the paucity of samples with reported reactivities, it was not possible to test the models based on Gulf Continuous Flow reactivities by cross validation. A total of 15 coals were available from the study of Given et al. (1979); eleven individual coals with four duplicates received in separate shipments. Since there were 8 replicates for each coal, a total of 120 coal spectra were used to develop the predictive model for the Gulf continuous flow reactor. The resulting regression equation which included 5 factors had an r^2 of 88.8 percent and

standard error of the estimate of 5.7. Using all 20 factor scores, an r^2 of 93.1 percent and standard error of the estimate of 5.7 were obtained. These values compare with $r^2 = 83.2$ percent and standard error of the estimate 4.2 in the overall regression equation of Given et al, which included total sulfur, total reactive macerals and volatile matter. Because of differences in sample size, number of variables in the equation, imperfectly satisfied assumptions about the distribution and variance in the measured data, and conceptual differences between the two approaches, these results cannot be compared statistically.

Figure 23 shows the factor spectrum for reactivity in the Gulf continuous flow process. The coals used to develop this model were different from the ones used to develop the tubing bomb reactivity factor spectrum, yet both factor spectra are remarkably similar. They are so similar, in fact, that it is not possible to say whether the differences observed are due to actual differences between the two processes or are simply due to the fact that the models were based on different samples. In either case, the conclusions from the tubing bomb factor spectrum are reinforced: components

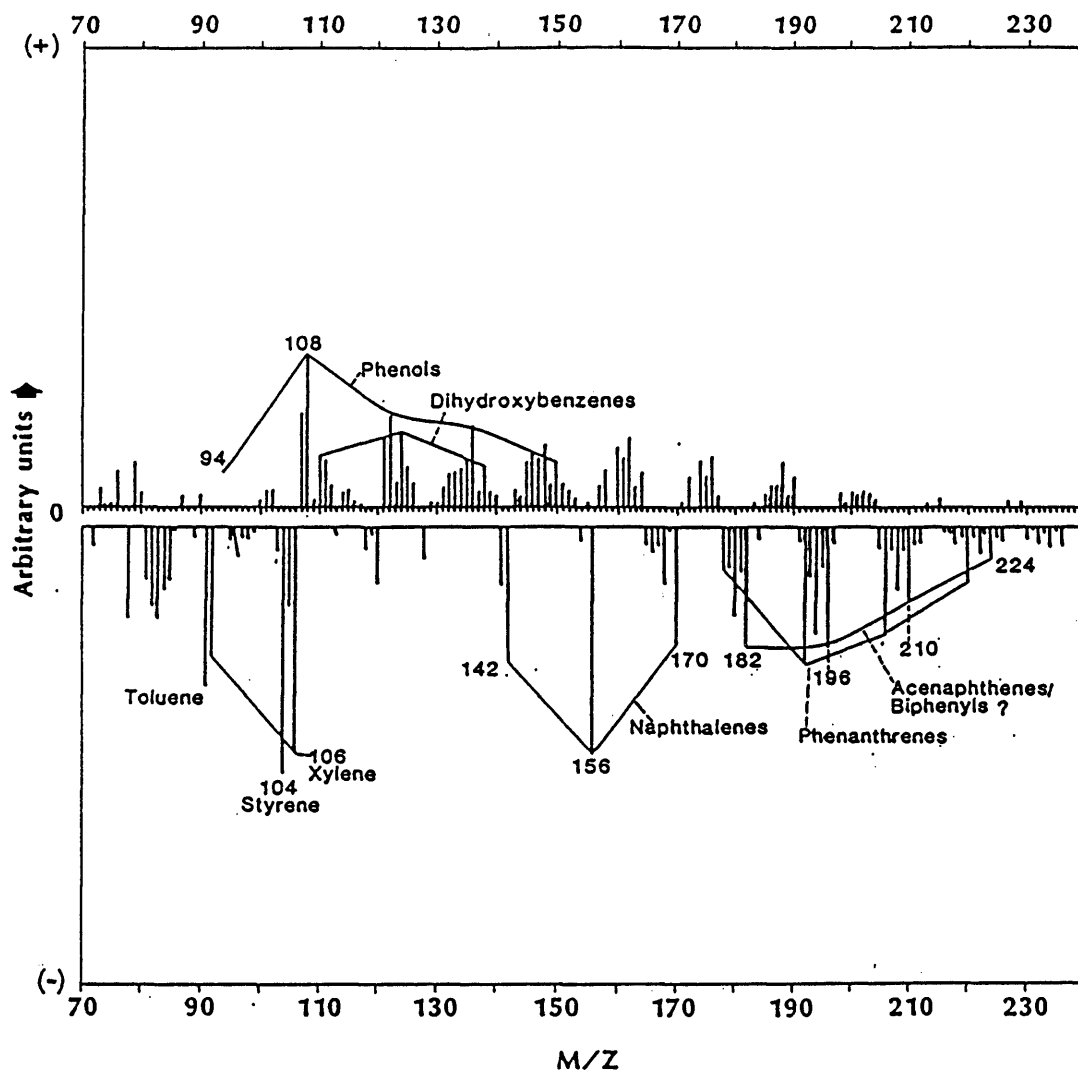


Figure 23. Factor spectrum for reactivity in the Gulf continuous flow process.

which characterize increasing rank are correlated negatively with reactivity, and phenols/aryl ethers have a positive correlation separate from rank.

D. Discussion of Results.

Most of the peaks with large correlations, in the factor plots which are not multiplied by the standard deviation, cannot be identified with any certainty and may represent several components each. Fortunately, most of the peaks which are ambiguous are also present in small abundances. The factor spectra produced after standard deviation multiplication contain less information but are easier to interpret. They show the familiar negative correlation between reactivity and rank and the positive correlation with sulfur content. They also reveal the fact that, for both tubing bomb and Gulf continuous flow conditions, alkylphenols/aryl ethers have a positive correlation apart from sulfur content and rank. The data indicate that aromatic, oxygen-containing compounds are important for increasing yield in both liquefaction processes.

If data for several types of reactivity were available for more of the coals, much more could be

learned about structural components which relate to reactivity and differences between liquefaction processes. The accuracy of the prediction could only be improved by obtaining samples for pyrolysis from the same coal which was liquefied. In this study, variance was introduced due to sampling differences, the different storage and handling of the specimens for liquefaction from the ones for pyrolysis, and the difference in time between the processing of the specimens for liquefaction and those for pyrolysis. The known heterogeneity of the coal and the possible effects of weathering affected the accuracy of prediction adversely. With a larger data set, it would be possible to do a more planned study in which one parameter such as rank or amount of the structural components was held nearly constant. It would be very interesting to see which of the oxygen- and sulfur-containing components retained a correlation after the effect of rank was removed.

For purposes of predicting liquefaction reactivity, the method could be improved. The data for this study were collected over an 11-day period. Variation due to background and changes in the

instrument's performance are apparently small compared to true differences between coals which may be used to predict reactivity. Therefore, it would be reasonable to compile a library of spectra from different coals for later use in developing models. From this library, coals with similar physical properties to an unknown coal could be used to develop a more accurate estimate of the known coal's reactivity. This would produce a quick, inexpensive test for purchasing of feedstocks and blending of coals in an industrial-scale liquefaction process.

V. EXXON COALS

A. Experimental

1. Samples

The processing of the Exxon coals was more rigorous in avoiding atmospheric contamination than that for the coals in previous chapters. Therefore the preparation of the samples will be described in somewhat more detail. Nine of the Exxon coal samples (Group 6 in Table III), including the seven coals for which reactivity data were available, were mixed with solid CO_2 in order to maintain a relatively inert atmosphere and ground in a ball mill. They were then stored in sealed bottles under N_2 gas in a dry box. From each of these sealed bottles, six separate samples, about 1 g each, were withdrawn, two of which were for purposes of studying weathering. The remaining thirteen coals were sampled four times. Each of these four samples was ground individually in an agate mortar and pestle to approximately 100 mesh-size under N_2 gas. All of the ground samples were placed in small sealed containers. The 18 weathering samples were removed from the drybox, spread in a thin layer on Petri dishes and covered loosely with a thin tissue.

Of the two samples for each coal, one was placed on a window ledge with a southern exposure, so that it was exposed to the diffuse summer sunlight and the atmosphere for a period of 68 days. The other sample was placed in an inside room with only fluorescent overhead light for the same 68 days.

At the end of the 68 days, the remaining samples; four each of the 22 coals available, were removed from the nitrogen environment. All of the weathered and unweathered samples were weighed and suspended in distilled methanol at a concentration of 5 mg/mL. Except during the weighing process and during the actual analysis, the unweathered opened samples and suspensions were stored at -20 degrees C. Fifty micrograms were applied to 610 degree ferromagnetic wires.

2. Analysis

In order to control for possible drift in the tuning of the instrument, the samples were pyrolysed in a previously determined random order (randomized complete block design, Montgomery, 1976). They were pyrolysed in sets of eight. Bracketing each set, two background samples were collected, consisting of a wire

which was not coated with coal. Although the background increased both within a set and between sets on the average, these increases were small relative to the amount of sample. Because of the random order in which the samples were run, background was not correlated with any of the meaningful factors in the coal, and was removed during the data treatment by principal component factor analysis when the higher factors were removed.

Since there were six samples each of nine coals plus four samples each of thirteen coals, 106 pyrolysis mass spectra were obtained altogether. Although every care was taken to insure uniformity in coating the coals on the wire, a significant but unavoidable random variation in the amount of pyrolysate produced was found. As in the previous studies, this was found to have little effect on the relative abundances of the peaks in the mass spectra. In order to remove this influence, the mass spectra were normalized to the sum of the peaks in the mass range 72-160 amu. Following normalization, the mass spectra were merged into one large file and factor analyzed using the statistical package ARTHUR.

3. Pattern Recognition

Since the peaks in a mass spectrum represent chemical components or derivatives of chemical components in the original coal, linear models work well for properties which are proportional to concentration. This was shown to be the case for the pseudo-equilibrium tubing bomb and GCF reactivities in the previous chapters. It is also justified by Equation 17 and subsequent arguments in chapter I. For measurements which are not linear with respect to concentration, such as kinetic measures of reactivity, a linear model is not completely justified. Over a narrow enough range, any function found in nature is linear. Unfortunately, for a small range, the magnitude of experimental error is large relative to the true response so a linear model is inaccurate. Therefore, there must be a trade-off between the range of the behavior modeled and the accuracy of the model.

B. Equilibrium and Kinetic Reactivities

The 106 coal spectra containing 209 peaks each were first autoscaled. The spectra were then factor analyzed in ARTHUR. The first 20 factors, which

accounted for 83 percent of the total variance, were retained for regression analysis using the program SPSS. The row matrix from factor analysis was entered into SPSS as well as the original mass spectra and all of the reactivity data available for this set of coals, yielding a large matrix containing all of the information available for each coal. Since reactivity data were only available for seven of the coals, no more than seven factors were justified for linear regression analysis. However, the seven factors with the highest percentage of variance explained are not necessarily the ones with the best predicting power (Jolliffe, 1982). In order to select the factors best able to predict reactivity, the technique of stepwise linear regression was employed.

Figures 24 to 26 show predictions of the rate constants for production of tetrahydrofuran-soluble (THF), toluene-soluble and pentane-soluble liquefaction products assuming a linear model. The number of factors used and some of the statistical parameters for each of the regression equations are summarized in Table XIII. Figures 24 to 26 strongly suggest that kinetic data may be modeled by Py/MS factor scores, but that a linear

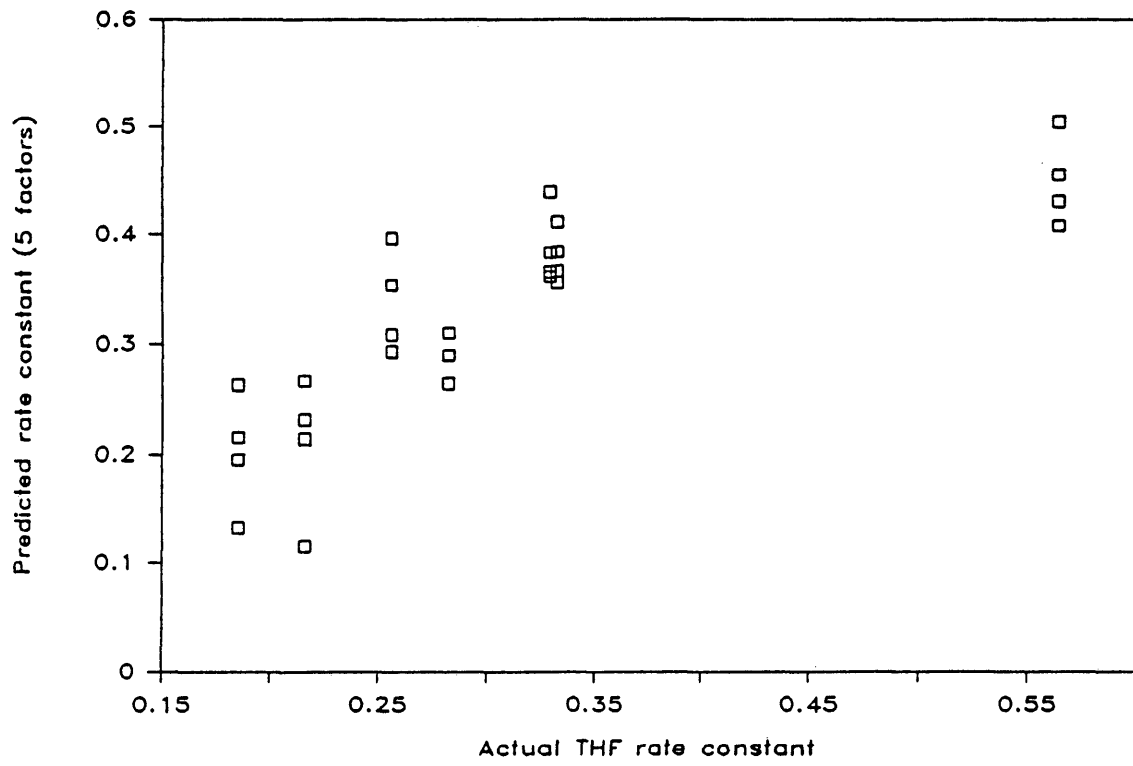


Figure 24. Predicted THF rate constant.

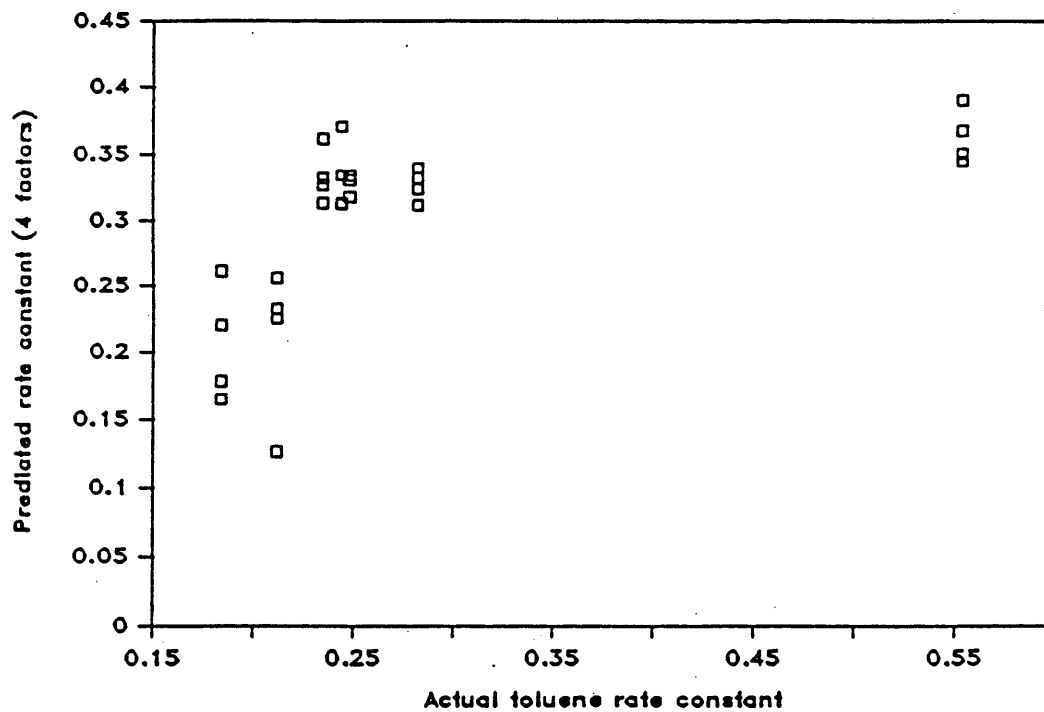


Figure 25. Predicted toluene rate constant.

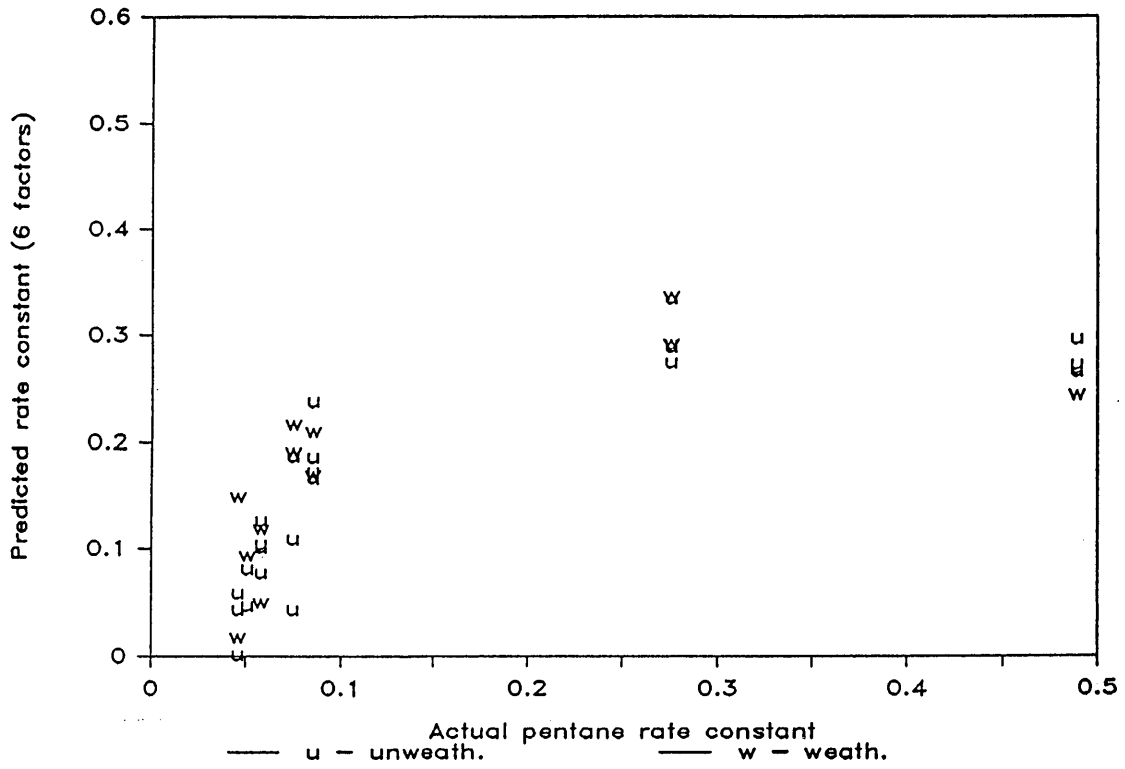


Figure 26. Predicted pentane rate constant.

	a	a	a	k	k	k	k	X of 60
	THf	toluene	pentane	THf	toluene	pentane	THf	
Number of factors in "best" equation:	3	7	6	5	3	6	5	
n of samples	7	7	7	7	7	7	7	
Coeff. of var. (r squared)	0.81	0.90	0.96	0.89	0.74	0.93	0.82	
Std. err. of estimate	0.046	0.020	0.014	0.044	0.065	0.049	2.33	

Table XIII. Regression statistics for most reasonable equation calculated from factor analysis and stepwise linear regression.

model is not ideal. Figure 26 also shows the reactivity prediction for the extremely weathered coals. The weathered coals were not used to develop the model, but were treated as unknowns. Their reactivities were predicted using Equation 20 in chapter I. Considering the severity of the weathering treatment and the nonlinearity of the kinetic data, the similarity between the predictions of kinetic reactivities of the unweathered and the weathered coal is remarkable.

Consideration of the original factor scores of the coals offers some insight into the reasons that kinetic reactivity prediction is not as good as equilibrium reactivity prediction. Figure 27 is a KL plot of the two factors most correlated with the kinetic reactivity (k) for pentane solubles. In order to simplify the plot, the samples were ordered according to the magnitude of k . Sample 7 had the greatest k and sample 1 the lowest. Samples with the same number are replicates. The two samples with the lowest k (1 and 2) are found in one region, and 5, 6 and 7 are radially distributed. One possible interpretation of the figure is that coals with a low k are similar, but higher rate

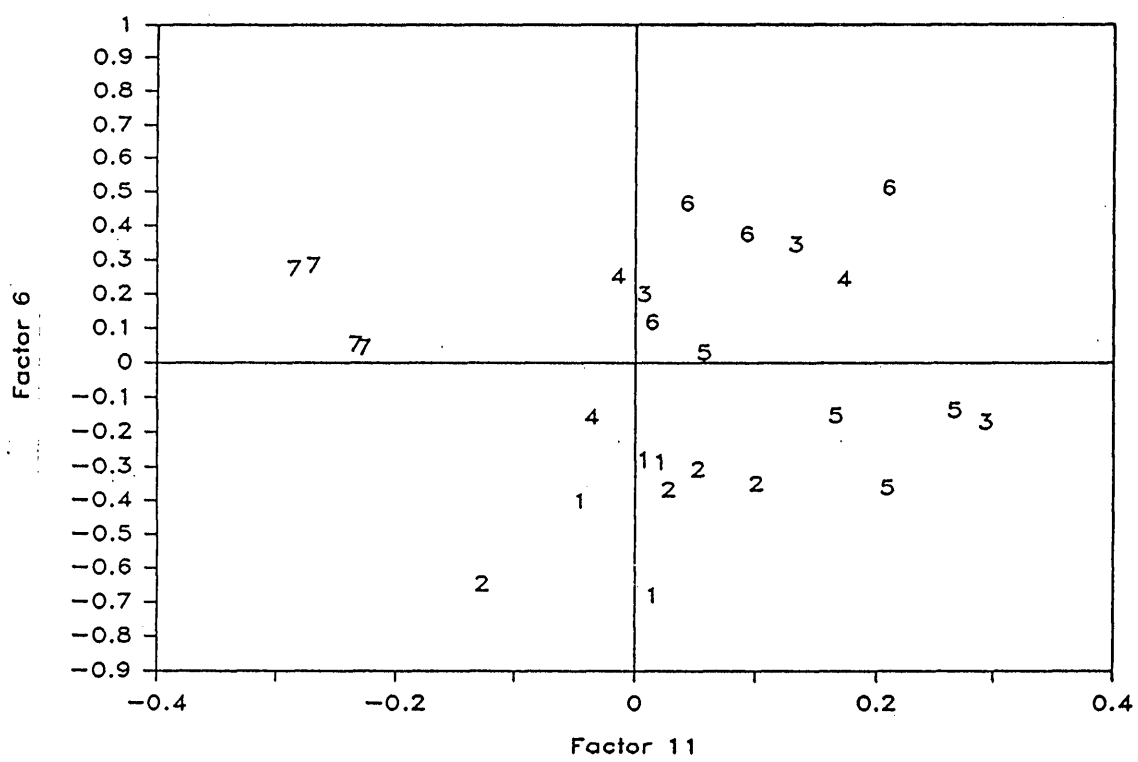


Figure 27. KL projection for pentane k.

constants arise from different sources. No linear trend is apparent, but the plot contains information which could be used to qualitatively estimate the value of k . Figure 28 is a plot of two orthogonal axes in the data space, one a combination of factors 5 and 11 and the other a combination of factors 6 and 12. The coefficients for the regression calculations were used as the cosines for the angle of rotation of the principal components. This is justified by the similarity between Equations (20) and (21) in the INTRODUCTION and the close relationship between regression and factor analysis (Malinowski and Howery, 1980). Again, no single straight line could be drawn which predicts k , but the lowest and highest reactivity coals are set apart from the coals with intermediate kinetic reactivity.

Figure 29 is a KL plot of the two factors most correlated with the rate constant for toluene-solubles (toluene k). The samples were renumbered in order of increasing toluene k . Again, the sample with the greatest k is set well apart from the others, and the samples with the lowest correlation are grouped together. Figure 30 is a plot similar to Figure 28 for

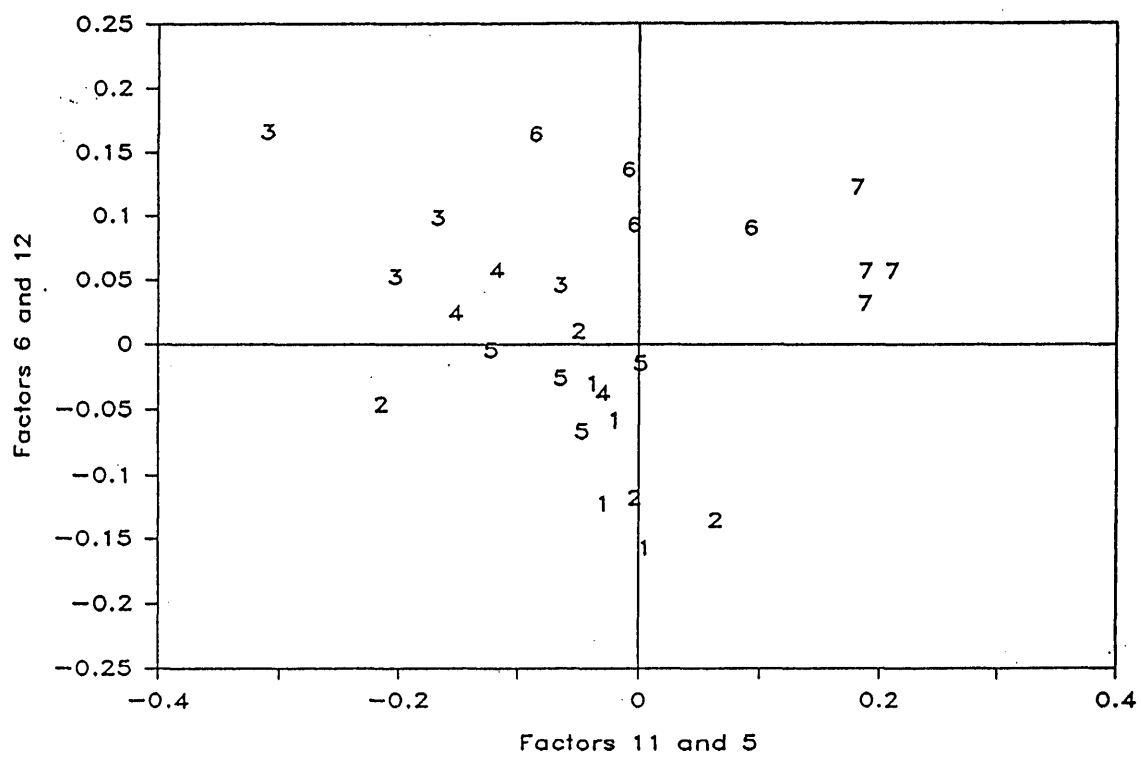


Figure 28. KL projection for pentane k with rotation.

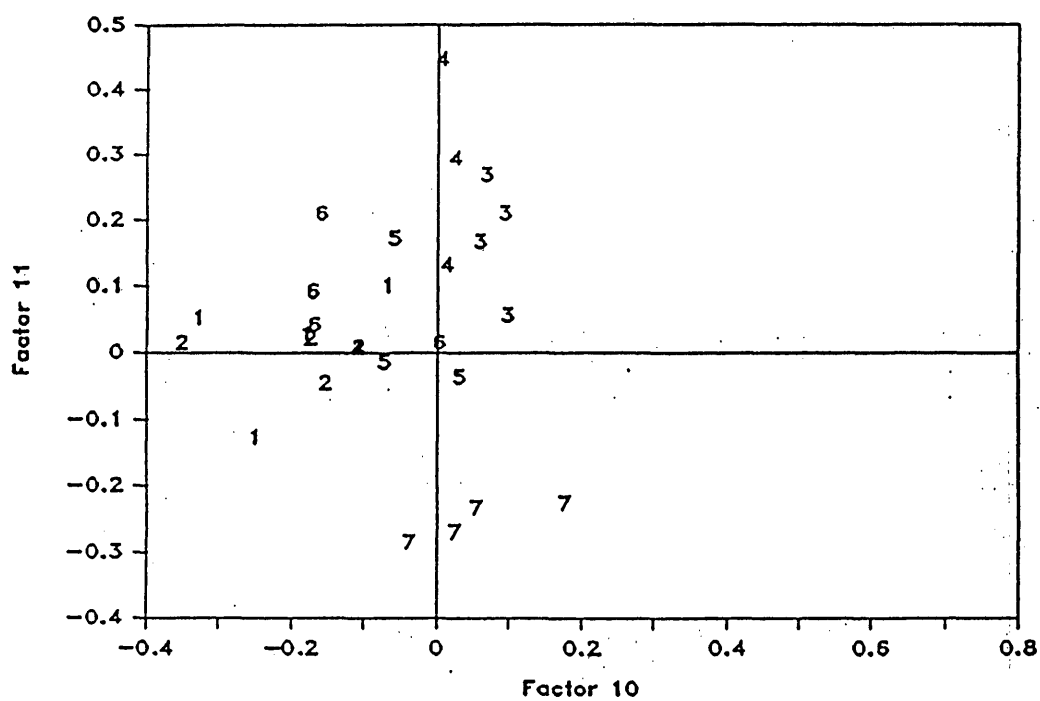


Figure 29. KL projection for toluene k.

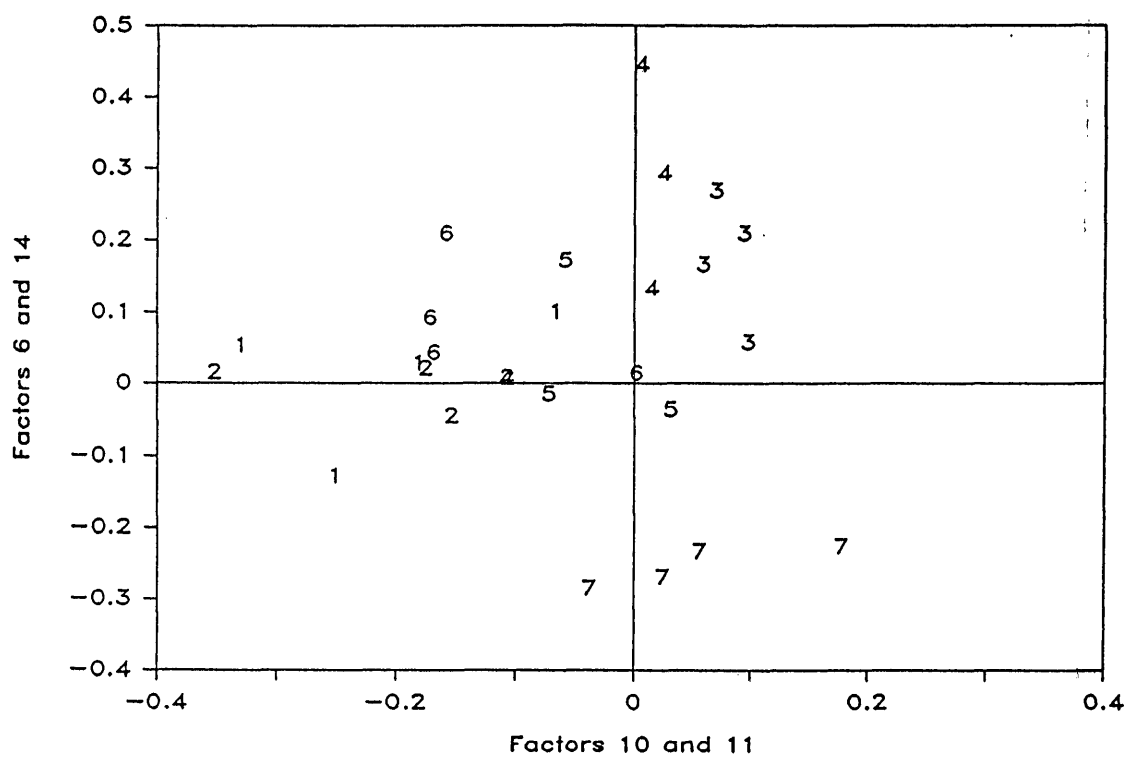


Figure 30. KL projection for toluene k with rotation.

toluene conversion. As in that plot, there seem to be trends related to reactivity, but they are poor for prediction purposes except for the coal with the highest reactivity. In the 20-dimensional data space there might be other factors or combinations of factors which better model reactivity, but the space is so enormous that without a technique such as the linear regression technique used here it is impossible to effectively scan the space visually for trends. In future work, more sophisticated techniques of cluster analysis might greatly improve the ability to estimate kinetic reactivity.

The most promising KL plot for predicting kinetic reactivity is shown in Figure 31. A reasonably straight line may be drawn which separates the lowest from the highest reactivity coals. All of the plots imply that the highest k reactivities are controlled by completely different factors than the ones which differentiate lower k coals.

Figures 32, 33, 34, and 35 show the performance of the best equations found for predicting the parameter a by Walberg (1984). Parameter a is a pseudo-equilibrium reactivity extrapolated to infinite residence time from

kinetic data. It is the fraction of the coal which would be converted to products soluble in the specified solvent after infinite reaction time. A pseudo-equilibrium reactivity, X of 60, indicates the amount which was converted to toluene soluble products after 60 minutes. All the equations model the data well.

As more parameters are added to the model, the accuracy of predicting the reactivity of a coal which was used in the model (standard error of the estimate) will always improve. The large data set study reported earlier showed that the accuracy of predicting unknowns will improve as more factors are added to the equation, but only up to a point. After that point has been reached, the accuracy of predicting unknowns will actually become worse as chance correlations are entered into the equation and the model becomes increasingly "fitted to the error". Except by chance, the ability to predict unknowns is never quite as good as Figures 32 through 35 indicate (Draper and Smith, 1981).

In the large data set study, Chapter IV, the approximate range of 5 to 7 factors was found to provide the best predictive equation. In this data

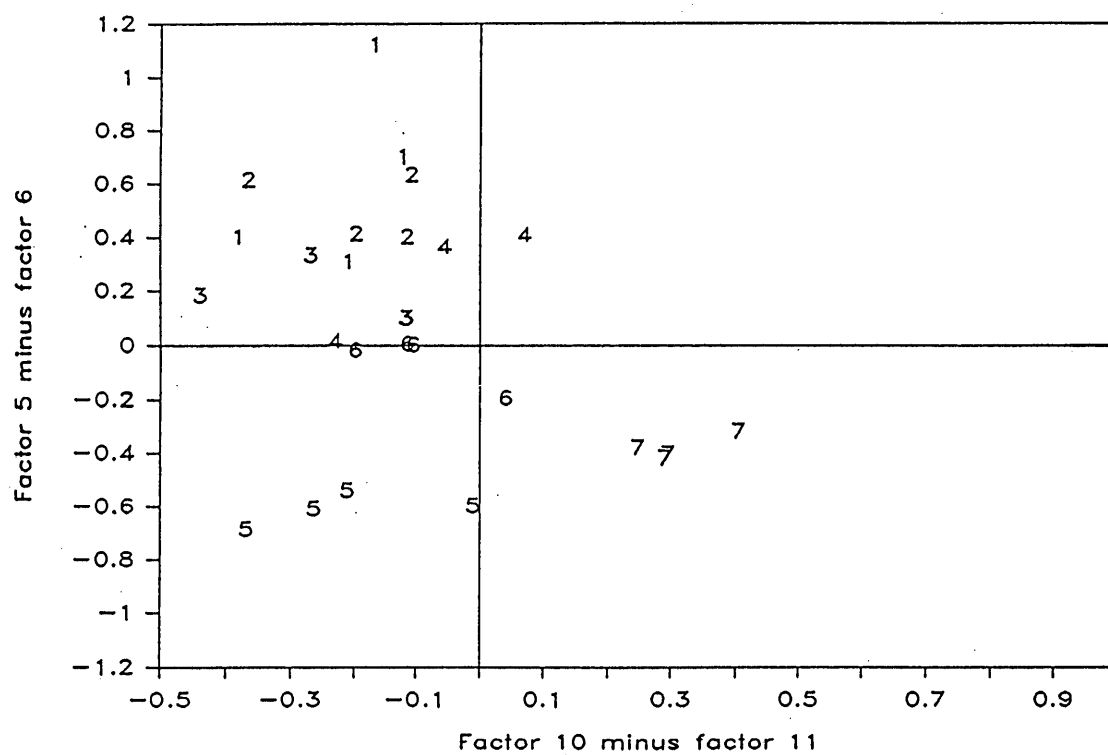


Figure 31. KL projection for THF k with rotation.

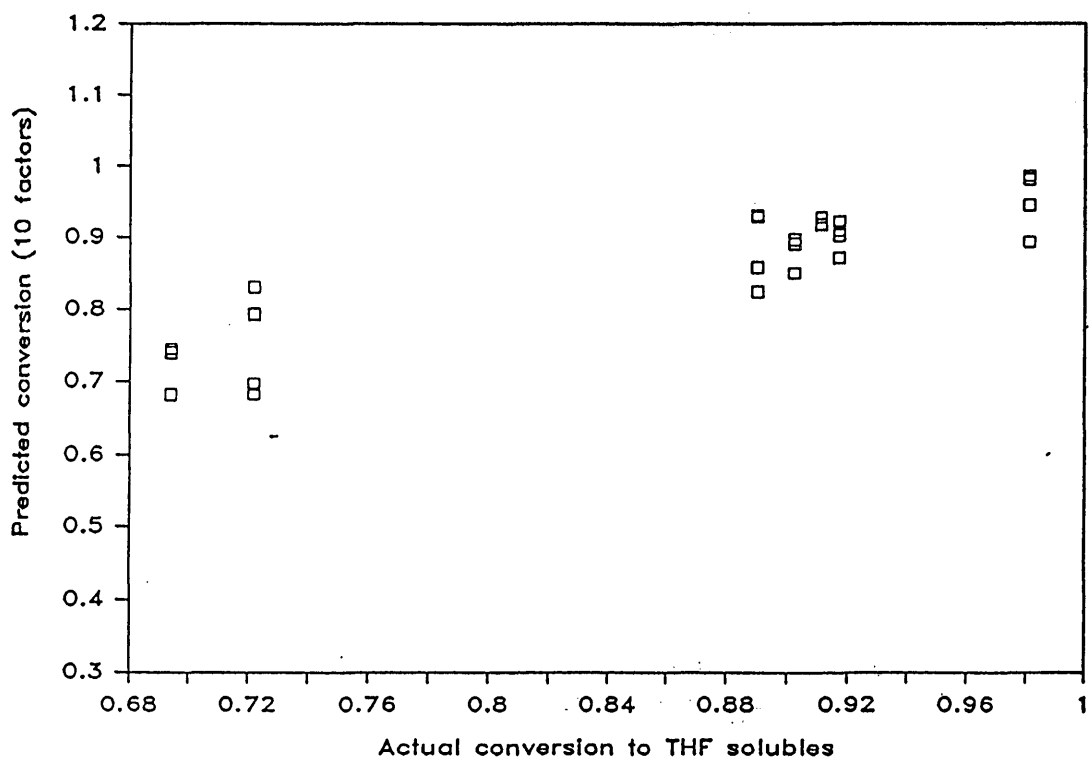


Figure 32. Predicted THF conversion (three factor model).

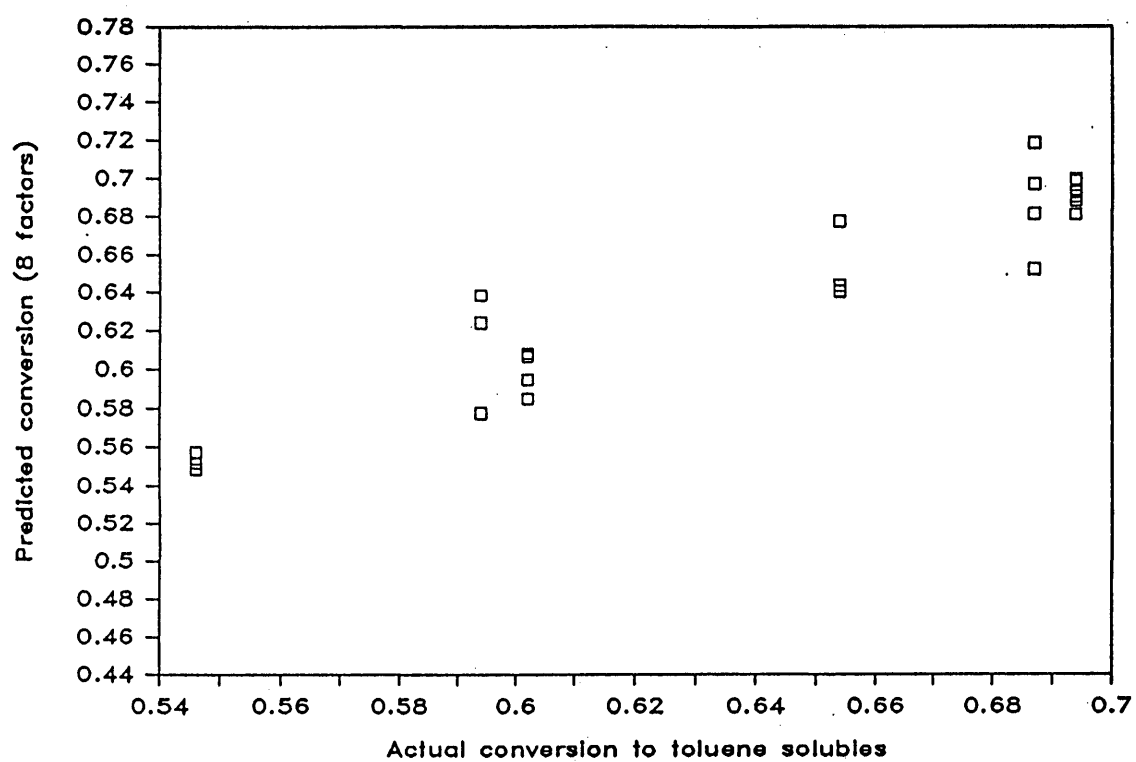


Figure 33. Predicted toluene conversion (eight factor model).

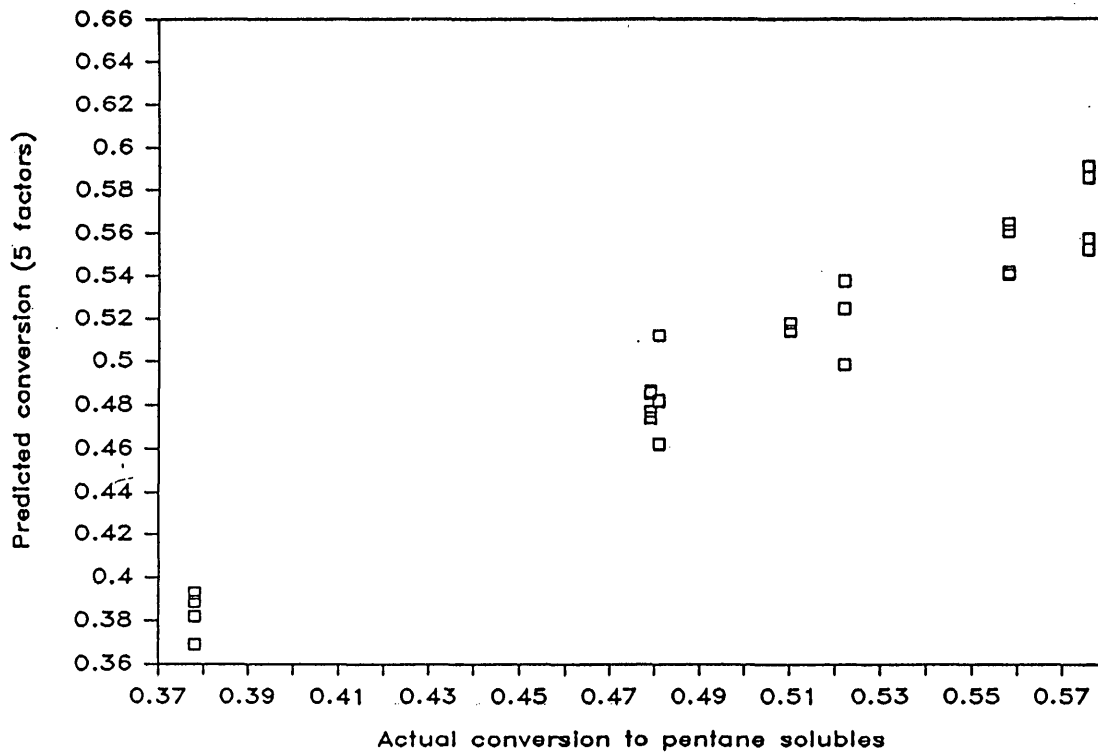


Figure 34. Predicted pentane conversion (five factor model).

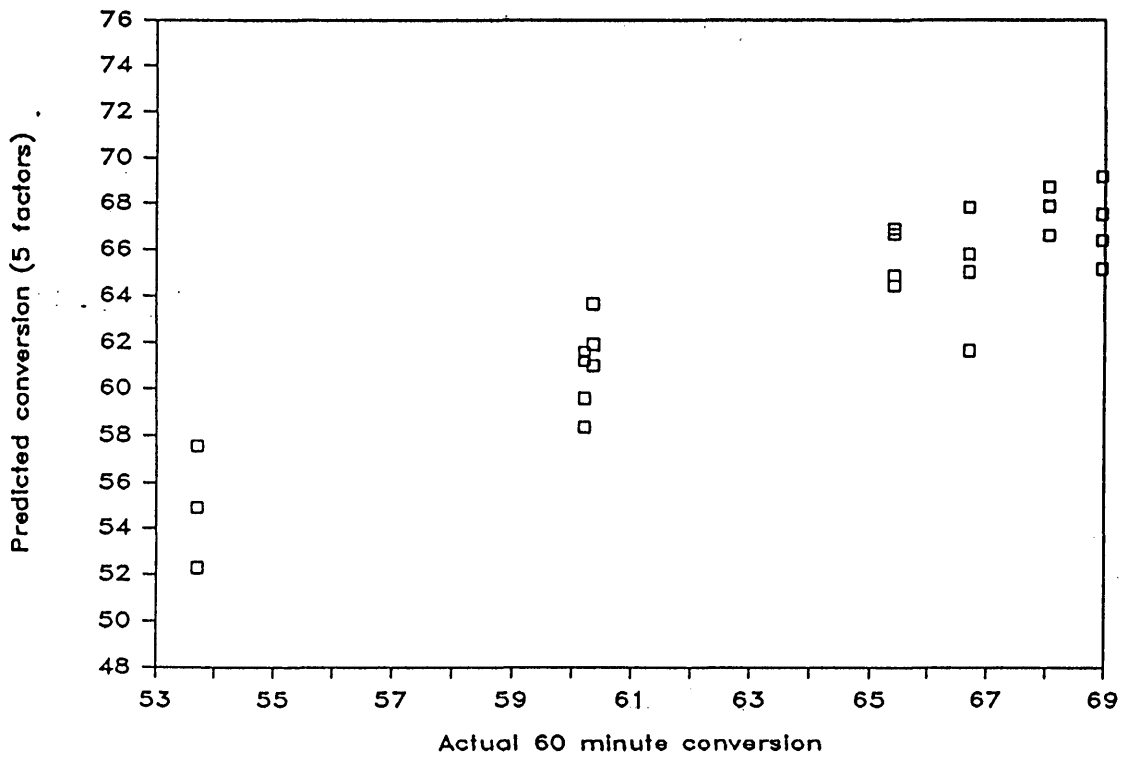


Figure 35. Predicted pseudoequilibrium reactivity (five factor model).

set, there were not enough samples with known reactivity to assess the optimum number of factors directly by cross validation. Approximately 5 to 7 factors were used in the equations for this set of experiments under the assumption that the data were similar in complexity.

At this point, the data presented in Figures 32 through 35 and the reactivity predictions in Table XIV are the best possible from the available data, taking the results from the large data set into account. For routine, accurate prediction of the reactivity of unknown coals, the model should be tested on true unknowns.

As in the large data set studies, an additional useful piece of information from the factor analysis/linear regression technique is the correlation between the original mass spectral peaks and the property modeled by regression. The peaks identified on the factor spectra to follow are those for which the identity of the major component responsible is known with reasonable confidence, and which have been similarly identified in the literature reviewed in the INTRODUCTION.

Coal	k THF		k toluene		k pentane		a THF		a toluene		a pentane		X of 60	
	meas.	pred.	meas.	pred.	meas.	pred.	meas.	pred.	meas.	pred.	meas.	pred.	meas.	pred.
046 ak		0.383		0.334		0.114		0.987		0.633		0.423		65.4
033 bm		0.348		0.272		0.098		0.833		0.643		0.559		63.1
057 dn	0.256	0.265	0.244	0.333	0.058	0.103	0.911	0.922	0.602	0.582	0.481	0.482	60.2	61.4
060 k9	0.332	0.322	0.235	0.334	0.085	0.192	0.917	0.903	0.694	0.686	0.479	0.481	68.1	67.1
027 nb	0.185	0.197	0.184	0.206	0.051	0.051	0.981	0.952	0.694	0.679	0.558	0.555	69.0	66.9
031 dm	0.282	0.274	0.248	0.328	0.075	0.114	0.694	0.723	0.546	0.550	0.522	0.521	53.7	57.4
002 bb	0.564	0.533	0.554	0.364	0.489	0.276	0.722	0.751	0.654	0.649	0.576	0.572	65.4	63.3
022 hn	0.216	0.206	0.212	0.210	0.046	0.017	0.902	0.870	0.594	0.618	0.510	0.582	60.4	61.8
FIES #11	0.329	0.387	0.282	0.327	0.276	0.297	0.890	0.886	0.687	0.680	0.378	0.382	66.7	65.1
086 av		0.342		0.331		0.185		0.899		0.619		0.441		63.1
007 mk		0.412		0.309		0.211		0.958		0.746		0.488		70.5
098 a4		0.385		0.361		0.111		1.037		0.623		0.406		65.5
105 td		0.430		0.317		0.235		0.456		0.599		0.638		55.7
021 rh		0.388		0.351		0.242		0.632		0.591		0.539		58.4
097 a3		0.328		0.320		0.122		1.015		0.644		0.429		66.2
041 sj		0.268		0.230		0.130		0.872		0.686		0.566		65.8
059 sd		0.270		0.265		0.096		0.795		0.617		0.567		61.6
001 yb		0.323		0.287		0.043		0.895		0.645		0.539		65.0
054 mt		0.422		0.346		0.384		0.817		0.679		0.390		63.8
020 cm		0.443		0.336		0.420		0.974		0.794		0.470		72.3
052 sr		0.466		0.382		0.331		1.042		0.750		0.425		72.2
034 sm		0.235		0.292		0.027		0.632		0.496		0.565		53.4

Table XIV. Prediction of reactivities for Exxon coals.

The factor spectrum for conversion to pentane solubles is shown in Figure 36. H_2S and S_2 are both strongly negatively correlated. Aromatic hydrocarbons including alkyl benzenes, alkyl naphthalenes and phenanthrenes are all negatively correlated. Interestingly, the dihydroxybenzenes are positively correlated, but the phenols are negatively correlated.

Figure 37 is a factor spectrum for conversion to toluene solubles. Peak 64, due to S_2 , is prominent. Two other peaks typically associated with S at 34 and 48 amu are, however, absent. The use of these peaks to indicate the form of the sulfur in the coal deserves further investigation. In Figure 37 the oxygenated aromatic components are negatively correlated and the non-oxygenated aromatic components (except the naphthalene series) are positively correlated. The naphthalene series, which is not prominently correlated in either direction, is a good indicator of rank. It is also negatively correlated with oxygen-containing components in rank-dependent properties. If the factor spectrum were simply a reflection of rank, one would expect oxygen-containing sequences and the naphthalene sequence to be strongly and oppositely correlated.

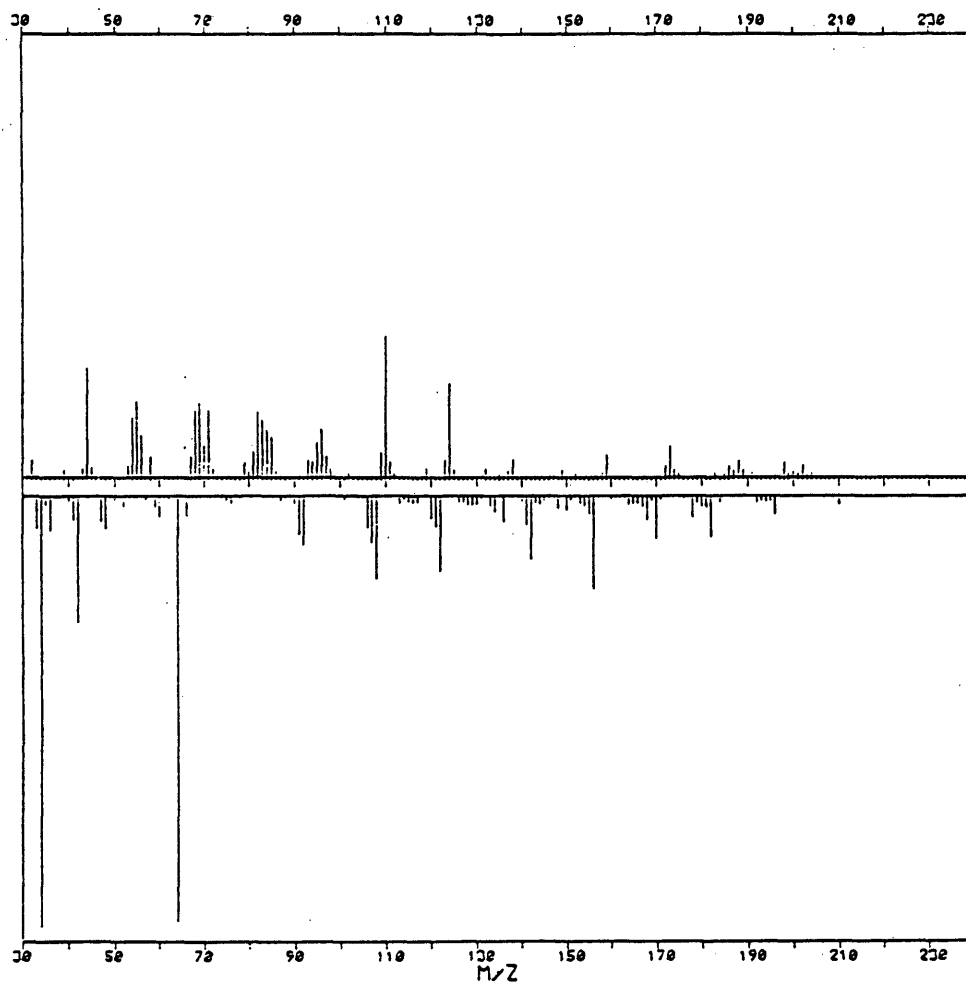


Figure 36. Factor spectrum for pentane conversion.

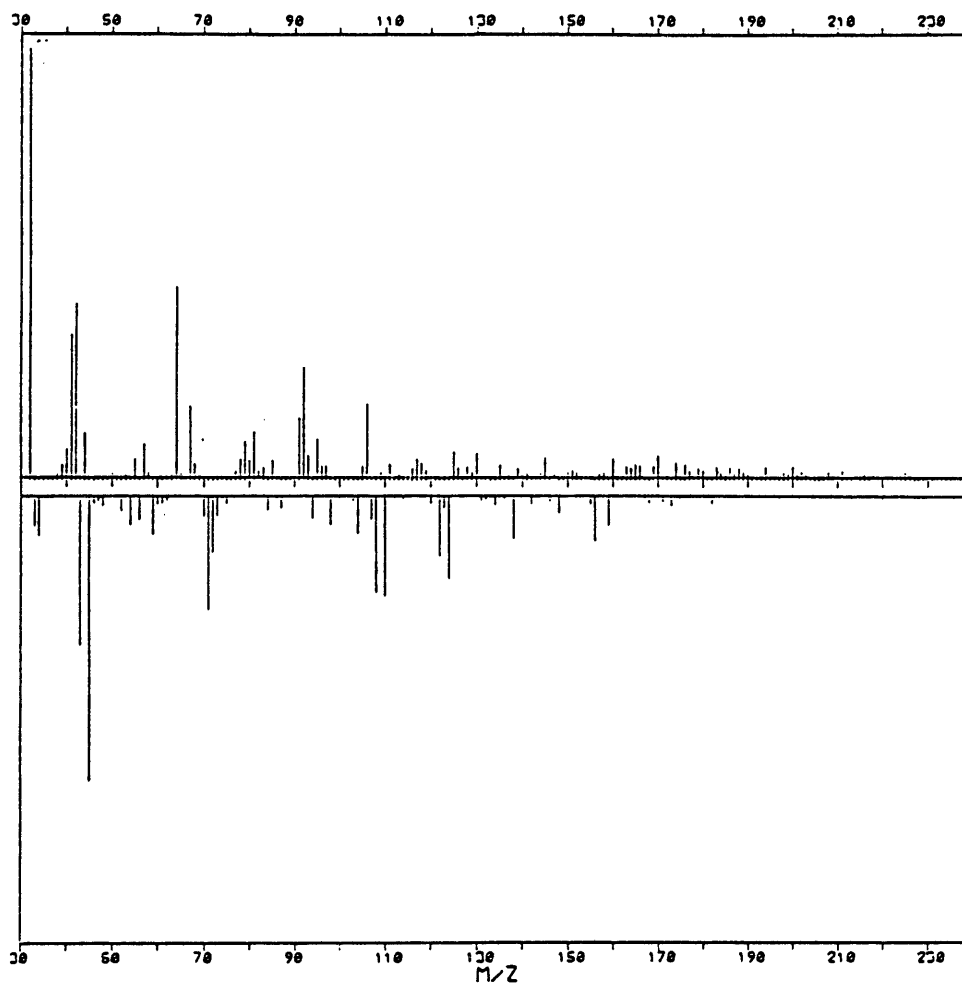


Figure 37. Factor spectrum for toluene conversion.

Since the naphthalene series is not correlated in this spectrum, the negative correlation of oxygen indicates correlation apart from rank.

Figure 38 is a factor spectrum for conversion to THF solubles. As with pentane conversion, it reveals a negative contribution due to methyl phenol, dihydroxybenzene and CO_2 , all oxygenated species, probably lignin-derived. Toluene, xylene and the methyl naphthalenes are positively correlated with THF conversion. These patterns are similar to those observed in a typical factor spectrum for rank. They may indicate a correlation between rank and THF conversion for this set of coals or they may reflect actual involvement of oxygenated aromatic species detrimental to conversion.

The pentane, toluene and THF solubles were definitions for oils, asphaltenes and preasphaltenes, respectively in Walberg (1984). Since toluene, xylene and the alkyl naphthalene series are probably indicators (after pyrolysis) of aromaticity and condensation in the original coal, the negative correlation of these components with oils, small positive correlation with asphaltenes and strong

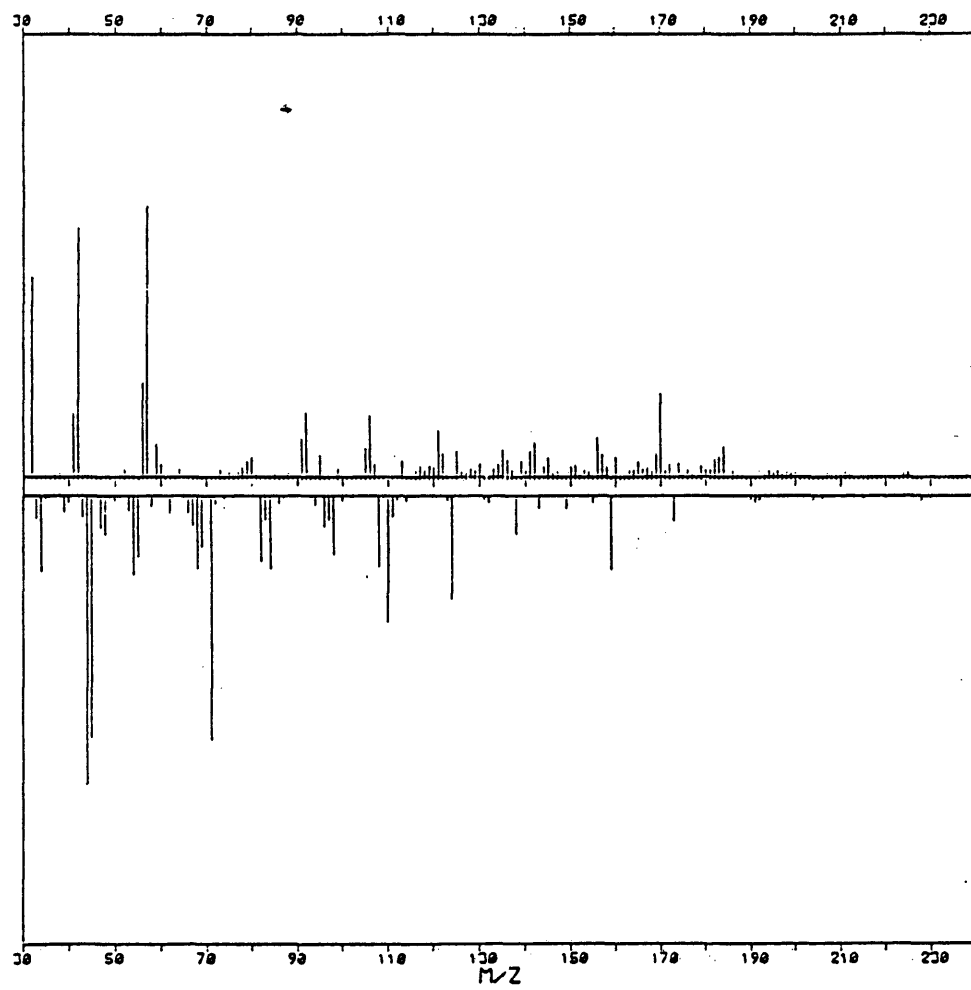


Figure 38. Factor spectrum for THF conversion.

positive correlation with preasphaltenes are expected. The positive correlation between the dihydroxybenzene series and yield of oils which is supported by a negative correlation between this series and preasphaltene yield might seem unusual since heteroatom content would normally be higher in the latter. But factor spectra do not necessarily indicate the composition of the product. They show which peaks are correlated with yield of the product. These results are consistent with the possibility that aryl ethers are involved as catalysts in liquefaction (Abdel-Baset, M. et al., 1978; Schlosberg et al., 1983; and McClennen et al., 1983)

The negative correlation between masses 34 (H_2S) and 64 (S_2) with conversion to pentane solubles is peculiar. Total sulfur content, when correlated, is consistently positively correlated with reactivity (Yarzac et al., 1980 and Given et al., 1979). These sulfur peaks probably do not reflect the content of pyritic sulfur.

VI. CONCLUSIONS

The general and specific goals set forth in Chapter I section B, Statement of Objectives were accomplished within the scope of this thesis research. Accurate predictive models were developed for liquefaction reactivity in a variety of industrial- and laboratory-scale processes according to several definitions of reactivity. At the same time, the correlation of specific pyrolysis products with reactivity was determined. These goals were accomplished in the sequence outlined in the Statement of Objectives.

1. Nonlinear mapping of pyrolysis mass spectra of triplicate samples of eight coals with known tubing bomb reactivity showed that there were three clusters. The clusters related to geographical location and also to reactivity. Hierarchical clustering analysis revealed the same clustering relationships. Fisher ratios suggested peaks which might be qualitatively related to the clustering trends. From these results, it was apparent that the pyrolysis spectra contained

information which could be used to predict liquefaction reactivity.

2. Triplicate samples of nine coals with known reactivity in the Gulf Continuous Flow reactor were examined using the same techniques. Two main groups were evident on the nonlinear map, corresponding to coals with greater than 50 percent conversion and coals with less than 50 percent conversion. As with the tubing bomb coals, reactivity information was present in the mass spectra, and Fisher ratios provided qualitative clues about the ions which were related to reactivity.
3. Triplicate pyrolysis spectra of eight coals with known reactivity in the stirred batch reactor formed one central cluster of coals with intermediate reactivity. Using simple, unsupervised pattern recognition methods, the connection with reactivity was not as obvious in this set of coals as in the tubing bomb and GCF reactor experiments.
4. Using a large data set with eight replicates each of 26 coal samples, equations were developed which

accurately predicted the liquefaction reactivity of coals treated as unknowns. It was not necessary to discriminate between Eastern and Interior coals; in fact, a model based on Interior coals alone successfully predicted the reactivity of Eastern coals and vice versa. The amount of success in predicting reactivity values for Western coals could not be assessed because of the small number of Western coals available.

5. Several homologous series were found to be uniformly correlated with reactivity. Alkyl-substituted phenols and/or alkyl aryl ethers as well as alkyl-substituted dihydroxybenzenes or diethers were the most prominent series which were positively correlated. This result is in agreement with several suggestions in the literature that aromatic oxygen-containing components are actively involved in liquefaction. References and descriptions of these papers may be found in the INTRODUCTION. Toluene, xylene, alkyl naphthalenes and acenaphthenes/biphenyls possessed a strong negative correlation. A comparison of the liquefaction factor spectrum to the vitrinite

reflectance spectrum revealed that the peaks related to liquefaction could be accounted for in part by rank effects, since rank and reactivity are negatively correlated. This is not to say that those peaks which are prominent in both factor spectra are not actively involved in liquefaction, since the higher reactivity of lower-ranked coals might be due to these components. The alkyl phenol/aryl ether series was particularly interesting because it was strongly correlated with reactivity but was not strongly correlated with rank in this set of coals. This may indicate some involvement of compounds in this series apart from the influence of rank.

6. A predictive model was developed for the Gulf Continuous Flow reactor (GCF), based on 15 coals, with eight replicates each, of known reactivity in this reactor. Cross validation was not possible because of the small number of samples, but regression statistics were calculated. The regression statistics with 5 factors in the model

were $r^2 = 88.8$ percent and standard error of the estimate = 5.7 percent conversion.

7. The factor spectrum for reactivity in the GCF reactor was very similar to that for the tubing bomb reactor. Therefore, the conclusions relative to the chemical structural characteristics which applied in conclusion 5 above also apply to the GCF results. The derivation of similar chemical information from both reactivity measures strongly reinforces the conclusions.
8. Predictive equations were developed for reactivity in the Exxon Donor Solvent (EDS) process, using a variety of definitions of reactivity. Unfortunately the training set was rather small, consisting of six samples each of seven coals with known reactivities for each of the definitions. Values for seven definitions of reactivity were measured by Walberg (1984) for this series of coals. The definitions of reactivity used were kinetic reactivity for production of THF-soluble material; kinetic reactivity for production of toluene-soluble, THF-insoluble material; kinetic reactivity for production of pentane-soluble,

toluene-insoluble material; total conversion to toluene-soluble material (pseudo-equilibrium reactivity); and true equilibrium reactivity estimates from the three kinetic definitions. The kinetic reactivity definitions were poor dependent variables for linear modeling because of the inherently non-linear nature of the presumed second-order kinetic data. The four equilibrium and pseudo-equilibrium reactivity measurements were modeled acceptably.

9. The factor spectrum for equilibrium conversion to pentane solubles (oils) was quite similar to the factor spectra for total conversion in the tubing bomb reactor and in the GCF reactor described earlier. Alkyl dihydroxybenzenes or their isomers were positively correlated, the alkyl phenols/aryl ethers were weakly but negatively correlated, and the alkyl aromatics were all negatively correlated. The factor spectrum for conversion to toluene samples (asphaltenes) showed a positive correlation with the non-oxygenated polyaromatic components. The factor spectrum for conversion to THF solubles (preasphaltenes) was very similar to

a typical factor spectrum for rank. Therefore, a general conclusion from this spectrum is that yield of preasphaltenes relative to other products should be greater for higher-ranked coals.

A number of difficulties in predicting liquefaction reactivity were cited in the opening paragraphs of this thesis. The Py/MS/pattern recognition methods which have been described in this thesis have overcome many of these problems: 1) The technique measures pyrolytic fragments of the whole coal and consequently relates more directly to liquefaction than measurement of the liquefaction products. 2) In Chapter IV it was demonstrated that liquefaction is a more general property than expected. A model based on Eastern coals successfully predicted the reactivity of Interior coals and vice versa. 3) Representative sampling did not appear to be a problem with any of the coals examined. Reproducibility was high and the predictions involving cross validation demonstrated that sampling was adequate. 4) A broad range of reactivities was investigated in the large data set study of Chapter IV, so that the results extend to most coals likely to be found, assuming that

the similarities found between reactivity of Eastern and Interior coals generalizes to coals from other locales. In one case, the modeling technique successfully extrapolated well outside the range of reactivities used to construct the model (Figure 16).

5) Although there is no generally accepted definition of reactivity, a separate model may be constructed for the pseudo-equilibrium yield of products for tubing bomb reactivity, GCF reactivity, and EDS reactivity. The method is also successful for modeling the amount of specific products expected. The method was unable to quantitatively model kinetic reactivity, however.

6) Since predictions of true unknowns are not available in papers describing laboratory-scale prototypes, their ability to predict liquefaction reactivity in large-scale processes cannot be compared directly to the Py/MS approach. The tubing bomb reactor and the stirred batch reactor are much more similar to the industrial-scale processes than analytical pyrolysis is, but they do not provide detailed information about the structure of the coal.

VII. REFERENCES

- Abdel-Baset, M.B., R.F. Yarzab and P.H. Given, 1978, Dependence of coal liquefaction behaviour on coal characteristics. 3. Statistical correlations of conversion in coal-tetralin interactions: *Fuel*, v. 57, p.89-94.
- Abdel-Baset, Z., P.H. Given and R.F. Yarzab, 1978, Re-examination of the phenolic hydroxyl contents of coals: *Fuel*, v. 57, pp. 95-99.
- Allan, J., 1975, Ph.D. Thesis, University of Newcastle upon Tyne, England.
- Allan, J. and A.G. Douglas, 1977, Variations in the content and distribution of n-alkanes in a series of carboniferous vitrinites and sporinites of bituminous rank: *Geochimica et Cosmochimica Acta*, v. 41, pp. 1223-30.
- Arunachalam, J. and S. Gangadharan, 1984, Feature extraction from spectral and other data by the principal components and discriminant function techniques: *Analytica Chimica Acta*, v. 157, pp. 246-259.
- Barker, C., 1974a, Programmed temperature pyrolysis of vitrinites of various rank: *Fuel*, v. 53, pp. 176-7.
- Barker, C., 1974b, Pyrolysis Techniques for Source-Rock Evaluation: *AAPG Bulletin*, v. 58, pp. 2349-61.
- Berkowitz, N., 1979, An Introduction to Coal Technology: New York, (Academic Press), 345 p.
- Bezdek, J.C., C. Coray, R. Gunderson and J. Watson, 1981a, Detection and characterization of Cluster Substructure I. Linear structure: Fuzzy c-Lines: *SIAM J. Appl. Math.*, v. 40, pp. 339-57.

- Bezdek, J.C., C. Coray, R. Gunderson and J. Watson, 1981b, Detection and characterization of Cluster Substructure II. Fuzzy c-varieties and convex combinations thereof: SIAM J. Appl. Math., v. 40, pp. 339-57.
- Domokos, L. and I. Frank, 1981, Orthogonal transformations for feature extraction in chemical pattern recognition: Analytica Chimica Acta, v. 133, pp. 261-270.
- Draper, N.R. and H. Smith, 1981, Applied Regression Analysis, Second Edition: New York, John Wiley and Sons, 709 p.
- Duewer, D.L., B.R. Kowalski and J.L. Fasching, 1976, Improving the reliability of factor analysis of chemical data by utilizing the measured analytical uncertainty: Anal. Chem., v. 48, pp. 2002-2010.
- Durfee, S.L., K.J. Voorhees, S. Larter, and J. Allan, 1982, Pyrolysis/Mass Spectrometry/ Pattern Recognition Studies of Sporinite Macerals: Paper presented at ACS Div. Geochem., National ACS meeting, Kansas City.
- Eshuis, W., P.G. Kistemaker and H.L.C. Meuzelaar, 1977, Some numerical aspects of reproducibility and specificity: in: C.E.R. Jones and C.A. Cramers (Eds.), Analytical Pyrolysis, Amsterdam, (Elsevier) pp. 151-66.
- Fisher, R.A., 1936, The use of multiple measurements in taxonomic problems: Ann. Eugen., v. 7, pp. 179-188.
- Foley, D.H., 1972, Considerations of Sample and Feature Size: IEEE Transactions on Information Theory, v. IT18, pp. 618- 626.
- Furlong, M.W., 1981, Correlation of parent coal properties with a kinetically-defined donor solvent liquefaction reactivity, Ph.D. Thesis, Colorado School of Mines, Golden, CO.

- Furlong, M.W., R.M. Baldwin and R.L. Bain, 1982, Reactivity of coal towards hydrogenation - ranking by kinetic measurements: Fuel, v. 61, pp. 116-120.
- Geisser, S., 1975, The predictive sample reuse method with applications: J. Am. Statist. Assoc., v. 70, pp. 320-328.
- Girling, G.W., 1963, Evolution of volatile hydrocarbons from coal: J. Appl. Chem., v. 13, pp. 77-91
- Given, P.H., W. Spackman, A. Davis, P.L. Walker, H.L. Lovell, M. Coleman and P.C. Painter, 1979, The relation of coal characteristics to liquefaction behavior: USDOE Quarterly Tech. Progress Reports, Jan-Jun 1978, FE-2494- 7/8, 1979, 68p.
- Harper, A.M., 1982, Personal communication.
- Harper, A.M., D.L. Duewer, B.R. Kowalski and J.L. Fasching, 1977, ARTHUR and experimental data analysis: the heuristic use of a polyalgorithm: in B.R. Kowalski (Ed.), Chemometrics: Theory and Applications, ACS Symp. Ser. no. 52, Wash. D.C., (Amer. Chem. Soc.) pp. 14-52.
- Harper, A.M., H.L.C. Meuzelaar and P.H. Given, 1984, Correlations between pyrolysis mass spectra of Rocky Mountain coals and conventional coal parameters: Fuel, v. 63, 1984, pp. 793-802.
- Hirota, K., Y. Yoshikawa and Y. Kobayashi, 1954, Gases evolved by heating coals in vacuo at low temperatures: Kagaku, v. 24, p. 524.
- Holden, H.W. and J.C. Robb, 1958, Mass Spectrometry of substances of low volatility: Nature (London), v. 182, p. 340.
- Holden, H.W. and J.C. Robb, 1960, A study of coal by mass spectrometry: Fuel, v. 39, pp. 39-46.

- Johansson, E., S. Wold and K. Sjodin, K., 1984, Minimizing effects of closure on analytical data: Anal. Chem., v. 56, pp. 1685-8.
- Jolliffe, Ian T., 1982, A note on the use of Principal Components in Regression: Appl. Statist., v. 31, pp. 300-303.
- Jurs, P.C., 1971, Machine intelligence applied to chemical systems. Prediction and reliability improvement in classification of low resolution mass spectral data: Anal. Chem., v. 43, pp. 22-26.
- Kowalski, B.R. and C.F. Bender, 1972, Pattern recognition. A powerful approach to interpreting chemical data: J. Amer. Chem. Soc., v. 94, pp. 5632-5639.
- Larter, S.R., 1978, Ph.D. Thesis, University of Newcastle upon Tyne, England.
- Larter, S.R. and A.G. Douglas, 1978, Low-molecular weight aromatic hydrocarbons in coal maceral pyrolysates as indicators of diagenesis and organic matter type: in W.E. Krumbein, (Ed.), Environmental Biogeochemistry and Geomicrobiology, Vol. 1, Ann Arbor, MI, (Ann Arbor Sci. Publ.), pp. 373-386.
- Larter, S.R. and A.G. Douglas, 1980, A pyrolysis-gas chromatographic method for kerogen typing: Phys. Chem. Earth, v. 12, pp. 579-583.
- Larter, S.R. and A.G. Douglas, 1982, Pyrolysis methods in organic geochemistry: J. Anal. App. Pyrol., v. 4, p. 1-19.
- Larter, S. R., B. Horsfield and A.G. Douglas, 1977, Pyrolysis as a possible means of determining the petroleum generating potential of sedimentary organic matter: in C.E.R. Jones and C.A. Cramers, (Eds.), Analytical Pyrolysis, Amsterdam, (Elsevier), pp. 189-202.

- Larter, S.R., H. Solli and A.G. Douglas, 1978, Analysis of kerogens by pyrolysis-gas chromatography-mass spectrometry using selective ion detection: *J. Chromatogr.*, v. 167, p. 421-431.
- Lumpkin, H.E. and T. Aczel, 1964, Low-voltage sensitivities of aromatic hydrocarbons: *Anal. Chem.*, v. 36, pp. 181-184.
- Malinowski, E.R. and D.G. Howery, 1980, *Factor Analysis in Chemistry*: New York (Wiley-Interscience), 251p.
- Mastral, A.M., V.C. Cebolla and J.M. Gavilan, 1984, Alkyl aryl ethers in lignite solubilization. 1. Study of the process: *Fuel*, v. 63, pp. 1422-1426.
- McCarthy, P.J., 1976, The use of balanced half-sample replicates in cross-validation studies: *J. Am. Statist. Assoc.*, v. 71, pp. 596-604.
- McClennen, W.H., H.L.C. Meuzelaar, G.S. Metcalf and G.R. Hill, 1983, Characterization of phenols and indanols in coal-derived liquids: *Fuel*, v. 62, pp. 1422-1429.
- McLafferty, F.W., 1980, *Interpretation of Mass Spectra*: Mill Valley, Calif. (University Science Books) 303p.
- Meisel, W.S., 1972, *Computer-oriented approaches to pattern recognition*, New York (Academic Press), 250p.
- Meuzelaar, H.L.C., A.M. Harper, R.J. Pugmire and J. Karas, 1984, Characterization of coal maceral concentrates by Curie-point pyrolysis mass spectrometry: *Int. J. Coal Geol.*, v. 4, pp. 143-171.
- Meuzelaar, H.L.C., A.M. Harper, G.R. Hill and P.H. Given, 1984b, Characterization and classification of Rocky Mountain coals by Curie-point pyrolysis mass spectrometry: *Fuel*, v. 63, pp. 640-52.

- Meuzelaar, H.L.C., J. Haverkamp and F.D. Hileman, 1982, Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials, Compendium and Atlas: Amsterdam, (Elsevier), 293 p.
- Meuzelaar, H.L.C., G.R. Hill, J.H. Futrell and A.M. Harper, 1982, Characterization of Rocky Mountain coals and coal liquids by computerized analytical techniques: Final Report to DOE, Grant No. DE-FG22-80PC30242, 117 p.
- Meuzelaar, H.L.C., R.E. Wood, J.H. Futrell and L.H. Wojcik, 1980, Proc. 28th ASMS Conf. on Mass Spectrom. All. Topics, New York, pp. 460-461.
- Montgomery, D.C., 1976, Design and analysis of experiments, John Wiley and Sons, N.Y., 418 p.
- Neavel, R.C., 1981, Exxon Donor Solvent liquefaction process: Phil. Trans. R. Soc. Lond. A, v. 300, p. 141-156.
- Neavel, R.C., S.E. Smith, E.J. Hippo, and R.N. Miller, 1981, Optimum classification of coals: Proc. Int. Conf. Coal Sci., Dusseldorf, 1-9.
- Romovacek, J. and J. Kubat, 1968, Characterization of coal substance by pyrolysis-gas chromatography: Anal. Chem., v. 40, pp. 1119-1126.
- Rouxhet, Paul G., M. Villey and A. Oberlin, 1979, Infrared study of the pyrolysis products of sporopollenin and lignite: Geochim. Cosmochim. Acta, v. 43, pp. 1705-13.
- Rozett, R.W. and E.M. Petersen, 1975, Methods of factor analysis of mass spectra: Anal. Chem., v. 47, pp. 1301-1308.
- Rummel, R.J., 1970, Applied Factor Analysis, Evanston, Ill. (Northwestern University Press) 617 p.
- Sammon, J.W., 1969, A nonlinear mapping for data structure analysis: IEEE Transact. Computers, v. C-18, pp. 401-9.

- Schlosberg, R.H., T.R. Ashe, R.J. Pancirov and M. Donaldson, 1981, Pyrolysis of benzyl ether under hydrogen starvation conditions: Fuel, v. 60, pp. 155-157.
- Schlosberg, R.H., W.H. Davis and T.R. Ashe, 1980, Pyrolysis studies of organic oxygenates. 2. Benzyl phenyl ether pyrolysis under batch autoclave conditions: Fuel, v. 60, pp. 201-204.
- Schlosberg, R.H., P.F. Szajowski, G.D. Dupre, J.A. Donik, A. Kurs, T.R. Ashe and W.N. Olmstead, 1983, Pyrolysis studies of organic oxygenates 3. High temperature rearrangement of aryl alkyl ethers: Fuel, v. 62, pp. 690-694.
- Searle, S.R., 1982, Matrix Algebra Useful for Statistics: New York, (John Wiley and Sons), 438p.
- Shadle, L.J. and P.H. Given, 1982, Dependence of coal liquefaction behaviour on coal characteristics. 7. Structural differences between coals of different rank and the asphaltenes derived from them: Fuel, v. 61, p.972-979.
- Siskin, M. and T. Aczel, 1983, Pyrolysis studies on the structure of ethers and phenols in coal: Fuel, v. 62, pp. 1321-1326.
- Snee, R.D., 1977, Validation of regression models: methods and examples: Technometrics, v. 19, pp. 415-428.
- Solli, H., S.R. Larter, and A.G. Douglas, 1979, Analysis of kerogens by pyrolysis-gas chromatography mass spectrometry using selective ion monitoring - III. Long chain alkylbenzenes: Adv. Org. Geochem., pp. 591-598.
- Solli, H., S.R. Larter and A.G. Douglas, 1980, The analysis of kerogens by pyrolysis-gas chromatography-mass spectrometry using selective ion monitoring. 2. Alkyl naphthalenes: J. Anal. App. Pyrol., v. 1, pp. 231-241.

- Tissot, B.P. and D.H. Welte, 1978, Petroleum Formation and Occurrence, Berlin, (Springer-Verlag), 538 p.
- van Graas, G., J.W. de Leeuw and P.A. Schenck, 1979, Characterization of coals and sedimentary organic matter by Curie point pyrolysis mass spectrometry, Part II.: in A.G. Douglas and J.R. Maxwell (Eds.), Advances in Organic Geochemistry, Oxford, (Pergamon Press) pp. 485-494.
- van Graas, G., J.W. De Leeuw and P.A. Schenck, 1980, Characterization of coals and sedimentary organic matter by Curie-point pyrolysis-mass spectrometry. Part 1.: J. Anal. App. Pyrol., v. 2, pp. 265-76.
- Villey, M., A. Oberlin and A. Combaz, 1979, Influence of elemental composition on carbonization pyrolysis of sporopollenin and lignite as models of kerogens: Carbon, v. 17, pp. 77-86.
- Voorhees, K.J. and S.L. Durfee, 1983, The analysis of naturally occurring polymers by pyrolysis/mass spectrometry: Colorado Sch. Mines Q., v. 78, pp. 23-29.
- Voorhees, K.J., S.L. Durfee and R.M. Baldwin, 1981, Liquefaction reactivity correlations using pyrolysis/mass spectrometry/pattern recognition procedures: Polymer Preprints, v. 22, pp 280-281.
- Walberg, R.L., 1984, Correlations between parent coal properties and kinetically-defined hydroliquefaction reactivities, M.Sc. Thesis, Colorado School of Mines, Golden, CO.
- Whitehurst, D.D.; Mitchell, T.O.; Farcasieu, M., 1980, In Coal Liquefaction, the Chemistry and Technology of Thermal Processes: New York, (Academic), p. 178.
- Windig, W., P.G. Kistemaker and J. Haverkamp, 1981, Chemical interpretation of differences in pyrolysis-Mass spectra of simulated mixtures of biopolymers by factor analysis with graphical rotation: J. Anal. Appl. Pyrol., v. 3, pp. 199-212.

- Windig, W. and H.L.C. Meuzelaar, 1984, Nonsupervised numerical component extraction from pyrolysis mass spectra of complex mixtures: *Anal. Chem.*, v. 56, pp. 2297-2303.
- Wold, S. and M. Sjostrom, 1977, SIMCA: A method for analyzing chemical data in terms of similarity and analogy: in B.R. Kowalski (Ed.), *Chemometrics: Theory and Application*, ACS Symp. Ser. no. 52, Wash. D.C., Amer. Chem. Soc., pp. 243-282.
- Yarzab, R.F.; Abdel-Baset, Z.; Given, P.H. *Geochim. Cosmochim. Acta*, 1979, 43, 281.
- Yarzab, R.F., P.H. Given, W. Spackman and A. Davis, 1980, Dependence of coal liquefaction behaviour on coal characteristics. 4. Cluster analyses for characteristics of 104 coals: *Fuel*, v. 59, p. 81-92.

APPENDIX I. DOCUMENTATION FOR PROGRAM SPIN

SPIN is a program which provides a variety of tools for the rotation and display of multivariate data. The program was written on VAX 780 and VAX 8600 computers in FORTRAN IV. The graphics in the program were designed for the HP 2600 series terminal, although KL plots may be displayed in text mode on any ANSI-standard terminal by setting the appropriate flag in the flags module described below. The program is entirely menu-driven, but because of the dual text/graphics modes on the HP graphics terminals, the text mode may be turned off and the program then functions as a command driven program.

At least two input files are required in order to run SPIN. The first is a raw data set formatted as an ARTHUR input file. The second is a card punch formatted file produced by ARTHUR which contains the means and standard deviations (if autoscaling was used) and the loadings from principal component or discriminant analysis. Varimax rotated files may also serve as input, as can any ARTHUR-generated card-punch formatted vector file. The program recognizes

internally whether the data were autoscaled, and treats the data accordingly.

Starting the program SPIN is started by typing "R spin" at the \$ prompt of the VAX/VMS operating system.

Data input. Once the names of the ARTHUR-formatted data file and the vector (FOR007.DAT) file have been entered, SPIN reads the vector file and plots the eigenvalues. In principal components analysis, this plot reveals much about the complexity of the data, as described in the INTRODUCTION of this thesis or any text about factor analysis, such as Rummel's (1970).

After the eigenvalue plot has been placed on the screen, the user has the option of reading the entire data matrix or user-defined spectra. If user-defined spectra are selected, the user must specify their order in the file as a range. For instance, "5,5" indicates the fifth data vector in the file; "7,18" indicates the seventh through the eighteenth data vectors, inclusive; and "0,0" indicates that all of the vectors desired have been entered. The ranges specified must not overlap or decrease in value. Once the desired data vectors have been specified, the data matrix is read,

the scores are read and the Givens rotation matrix (see INTRODUCTION) is fixed as an identity matrix. After the data are read, the Main Menu appears.

Main Menu. The Main Menu provides the general options available during the interactive session with SPIN. Any rotation or change in the data which is performed in a subroutine affects the entire data matrix (unless the "reinitialize rotations" option number 3 is selected), so that, for instance, a rotation in the K-L branch will also produce a rotation in the Factor Spectrum or Variance Diagram subroutines. This feature makes it possible to jump from module to module and see the effect of a change in one module on the appearance of another. The options available in the Main Menu will now be described in sequence.

1. Factor spectrum. This subroutine plots the factor spectrum associated with the first of a pair of factors. Two factors are involved so that the effect of rotating a second factor into a first can be seen. The factors do not necessarily have to be in order, so that, at the "What two factors rotated?" prompt, an answer of "1,2" will give a factor spectrum for factor

1, but an answer of "2,1" will give a factor spectrum for factor 2. Following the entry of the two factors, a rotation angle is requested. After the spectrum is plotted, it is possible to obtain a hardcopy of the factor spectrum and the K-L plot which goes with it by answering "y" to the prompt.

2. K-L Plot. This option invokes the most complex of the subroutines in SPIN. Therefore, each of the options on the K-L Plot submenu will be described separately. As with the Factor Spectrum subroutine, the user is asked to select two initial factors at the start of the routine. These factors may be changed while in the routine.

a. Scale up. This option magnifies the plot uniformly, so that the detail near the middle of the plot may be seen in large data sets. Points which would fall outside the plot are displayed as an "x" at the perimeter of the plot. Scale up values less than 1 reduce the size of the plot, and values greater than 1 expand it.

b. Shift axes. This option slides the plot up, down, right, and left. As with the scale up option,

the data remain unchanged - only the display is changed. Using a combination of the previous option and this one, it is possible to highlight any portion of the data desired. The blank character method in the "Change characters" option described below is another method for automatically focusing on the desired information.

c. List scores. This option makes it possible to print the scores (after rotation) in an ordinary file, easily accessible for other programs such as SPSS and LOTUS, which allow variable data formats. It will not work as ARTHUR input because ARTHUR has strict requirements for data entry. To produce an ARTHUR-compatible file, use option 5 in the Main Menu.

d. Change characters. Usually, the samples plotted on a K-L plot can be associated into groups, such as groups of replicates or different species of bacteria belonging to the same genus. In order to rapidly detect which samples go together on K-L plots, all of the samples belonging to the same group may be flagged with the same character. The prompts under this option are self-explanatory. The ranges of sample

names belonging to each character follow the same conventions used for ranges throughout SPIN: "5,5" indicates sample 5, "9,16" indicates samples 9 through 16 inclusive, and "0,0" ends input. The range is not restricted to the number of samples ("1,10000" is a valid entry) and successive ranges do not have to be sequential ("2,3" can follow "500,501"). Some skill and forethought can dramatically cut down on the number of keystrokes needed to flag large data sets with characters. For instance, if you have exactly two classes which are mixed together, label all of them as one class by choosing the range "1,10000", and then add the other class over the first one. Both classes will be correctly assigned.

Only normally printing characters are allowed, but the blank character has special meaning. If a character is flagged with a blank, the sample it is associated with is not used for scaling. If plotted offscreen it will still be flagged with an "x" along the perimeter, but only nonblank characters will be used to expand the plot into the area viewed. This makes it possible, by "blanking out" outlying or undesired data points, to focus automatically on the

samples of interest without explicitly shifting and/or scaling the plot.

e. Plot sample numbers The order of each sample in the file may be plotted using this option. Once this option or option f. is chosen, subsequent options will automatically produce the appropriate plot (sample number or character) upon return. Once selected, option e. becomes the default until option f. is selected or the K-L routine is exited. The hardcopy produced on exit will correspond to the most recent choice of option e. or f. This cuts down on the number of keystrokes required by making all of the other options context sensitive.

f. Plot characters. This selection plots the characters for the most recent pair of factors selected, with the special treatment of blank characters described under option d. Also, on an ANSI-compatible color terminal, each character is displayed in a different color for enhanced contrast. See the previous paragraph for other properties.

g. Rerotate the same factors. This option is used to change the angle of rotation of the two factors

plotted. The results are cumulative, and modify the actual internal data (not the files on disk - these files cannot be altered within SPIN), so that factor spectra, scores and variance diagrams will also reflect these rotations. The geometrically correct results of the rotation may be viewed by displaying the matrix of rotation from the flags subroutine of the Main Menu.

h. Rotate different factors. This selection changes which two factors are plotted. In all other respects it is the same as option g.

i. Exit. This option returns to the Main Menu. Once selected, the program asks whether a hardcopy is desired. The hardcopy is actually a file which may be used as input to the PQ decoder program (also written by Steve Durfee), which decodes symbolic plotter commands to commands which are compatible with CALCOMP plotter commands using the system program PLTMGR (see CSM computing center documentation). The symbolic plotter code was used to make the plot files ASCII in nature, so that they can be easily transferred to a wide variety of machines for plotting. The plot includes both the K-L plot and the corresponding factor

spectrum. The CSM version of SPIN conforms to the 11" by 11" window available to students.

3. Reinitialize Rotations This option rereads the original data (without requesting the filenames or ranges of sample numbers if only part of the data matrix has been used) and sets the Givens rotation matrix back to an identity matrix. It is generally used for one of three reasons: 1) The rotations have gotten completely out of hand, and you want to start over. 2) You have just saved a rotation for one property and want to try rotations for a different set of properties using the same character labelings. This often saves time if the variance has been "spread out" through all of the factors because of a large number of rotations. It also would cut down on cumulative round-off errors from repeated rotations, (although these have never been seen because of the 32 bit word size of the VAX, even after hundreds of rotations). 3) A set of characters has been read from a rotation file (option 9), but the rotations are not desired.

4. Targeted Least Squares. This option was implemented in an earlier version of SPIN. It appears to work, but

the results are not guaranteed because of the many other changes which have been introduced to SPIN since it was last debugged. However, the old code for targeted least squares is still in place, and could be implemented with minor debugging of this section. At this time, the preferred way of performing targeted least squares is to enter the constants into the Givens matrix using the appropriate sub-option in option 9, described below. The least squares factor spectra for this thesis were produced using the program NEW3. NEW3 has not been implemented on a VAX, but should run with appropriate changes to the OPEN and CLOSE statements. NEW3 is a very straightforward program which accomplishes essentially one purpose - the plotting of factor spectra using targeted least squares - so it will not be described here.

5. Output rotated score matrix to ARTHUR. This option creates an ARTHUR-formatted file of rotated scores of the user's choosing. This file may be used as input for ARTHUR pattern recognition routines.

6. Exit gracefully. The exit is considered graceful because it closes any open files and insures that an HP

2600 series terminal is returned to text mode. The user is placed immediately at VMS command level.

7. Choose different spectra. It is unusual to want to change the selected spectra when the rotations are reinitialized under option 3. This routine provides the capability to change the spectra. Also, it can be used for true supervised learning by cross-validation. A subset of the data are used for development of any model imaginable, with no restrictions at all, and then the rest of the data may be read and tested as an evaluation set using this routine.

8. Change flags. This subroutine alters certain default options used in all of the other subroutines. It enables the user to customize the entire program to his own preferences, prejudices, or hardware. The options are listed below.

a. Autoscaled data. You can't change this one. It reminds you which type of data you're working with.

b. True horizontal & vertical K-L plots. This option chooses whether the K-L plots are scaled the same horizontally as vertically with the center point

of the plot at (0,0) (answer yes) or the K-L plots are scaled arbitrarily to fill the screen (answer no). True scaling reflects the amount of variance in each plot, so that the range is visibly smaller for the higher factors. It's quite interesting to toggle this switch after "meaningful" information has been forced on the data by rotation.

c. Multiply by Std. Dev. for Factor Spectra.

Alice Harper (1982 argues that it can be misleading to multiply by the standard deviation, as advocated by Windig (1981). With this option, you can have it either way, although multiplication is the default.

d. HP 2600 series terminal. The default is the HP series graphics terminal; however any ANSI-standard text terminal will do. The text terminal loses overprinting of characters and microcentering of characters since the resolution is a poor 80 horizontal by 24 vertical. A color terminal will display each character in a different color for heightened visibility.

e. Display rotations for each pass. The display of the Givens matrix is normally suppressed. It

contains the true n-dimensional effect of rotations, and can be stored and reused by selecting option 9 of the Main Menu. When this option is selected, the Givens matrix is printed above the Main Menu, showing the position of the current (rotated) factors compared to the original factor axes in degrees. An entry in the matrix shows the angle between the original factor (row index) and the current, rotated factor (column index).

f. Std. Dev. first (for VAX). Unaccountably, current versions of ARTHUR differ from the original version by placing the standard deviations before the means in the card punch file. If SPIN is used to read the old (non-VAX) ARTHUR data files, this switch must be toggled and rotations reinitialized in the Main Menu before any other data manipulation is attempted in order to reflect autoscaling correctly.

g. Positive eigenvalues only. This flag was added as an experiment, the significance of which hasn't been fully tested. Windig (1985) in particular argues that negative eigenvalues are artifacts of autoscaling. While this is undoubtedly true, the significance of

negative eigenvalues is not clear. This flag makes it possible to calculate pseudo-scores based only on those eigenvalues greater than zero. The default mode, which uses both positive and negative eigenvalues, is recommended for most work.

9. Input, output or modify matrix of rotations. SPIN keeps track of all of the rotations which have been made by calculating a Givens matrix, which is orthonormal and symmetric (see INTRODUCTION). In matrix notation, $D = R T T^t C$, where T is the Givens matrix. This subroutine permits the storage and retrieval of the Givens matrix, so that the user may return at a later time to a session and resume where he was previously. It also stores the character assignments, which may take a great deal of time to enter the first time for large data sets, so this option may be selected even without rotation so time-consuming character assignments only need to be performed once. To do this, input the matrix of rotations using this option, say yes to the "Characters?" request, and then Reinitialize Rotations in the Main Menu. The character assignments from a completely different data set may also be used by

inputting the previous rotation matrix, which reads the characters, and then Reinitializing Rotations.

There are a large number of creative uses for this subroutine, many of which were not used in this thesis work. Among the most fruitful possibilities are the following: 1) Data from other programs, such as discriminant analysis in SPSS, may be made into a square matrix by augmenting the rectangular matrix with zeroes in the off diagonal and ones on the main diagonal. These matrices may be used to plot oblique rotations such as discriminant functions (although rotating them is not appropriate). 2) A completely different data set may be made conformable with the matrix used to derive the rotation matrix. Then plots of the new data using the derived rotation matrix are projections of the new data modeled on the space of the previous data. This method makes it possible to collect data at a later date and then determine how it relates to data modeled previously. 3) Targeted factor rotation for several dependent variables may be performed simultaneously by augmenting the matrix as in 1) above.

The input and output options are self-explanatory. The loading vector output option is similar to the score matrix output option in the K-L subroutine. It was added in order to provide a convenient method for preparing loading vectors for input to most other programs, such as SPSS and LOTUS. It produces both a score and a loading vector as output.

The modify rotations option provides the capability to alter the Givens matrix manually. Once manually altered, it is generally no longer a true Givens matrix, since it is no longer a product of a set of orthogonal rotations. The "Display matrix of rotations" sub-option in the Flags option will produce nonsense if the Givens matrix is modified.

This option is used when a set of multipliers have been found because of some other procedure. To reproduce the data of this thesis, the coefficients from multiple linear regression would be entered. Coefficients for each dependent variable (model) would be placed in a different vector. The routine requests one vector at a time. When a vector is selected, the vector is first zeroed and then the values of zero are replaced by the values entered by the user.

10. Variance diagram. The variance diagramming procedure described by Windig (1985) has been implemented in this subroutine. The algorithm used in SPIN is identical mathematically, but is 36 times faster for a 10 degree window than described in that paper (and is proportionately faster for a narrower window). As in the Factor Spectrum and K-L subroutines, this subroutine modifies the internal data in order to make all of the subroutines consistent.

a. Windig's laundering. This option implements the negative eigenvalue correction in strict accordance with Windig (1985). As outlined in the Flags subroutine, it compensates for negative eigenvalues, which are due to autoscaling. Windig suggests that negative eigenvalues should be subtracted from positive ones in order to compensate for this scaling phenomenon.

b. Durfee's laundering. I agree that negative eigenvalues are the result of autoscaling. I do not believe that they are a problem. However, I also believe that if you think they must be compensated for, the compensation should involve the addition, rather

than the subtraction of the absolute value of the complementary eigenvalues. This option adds the absolute value of the opposite eigenvalues to the eigenvalue plotted.

c. Smooth it. This routine smooths the variance diagrams in order to make them appear more meaningful (Windig, 1985).

APPENDIX II. PROCEDURE FOR MOVING DATA BETWEEN
PROGRAMS

What follows is a step-by-step procedure for producing the factor rotations and factor spectra described in this thesis. The starting point is a normalized, ARTHUR-formatted input file, which may be produced in a very straightforward manner using the programs ANALOG and ARTFIL written by the author for the PDP11-series of computers under the RT-11 operating system.

The precise sequence of steps and the programs used to produce the results described in this thesis cannot be recreated. All of the data analysis was performed before the Colorado School of Mines upgraded from a DEC-10 system running TOPS-10 to a VAX system running the VMS operating system. The current version of ARTHUR is different, the original programs written by the author have not been implemented on the VAX, and the version of SPSSX is so different that even the most trivial SPSS control file written under the old version will not work under the new one without modification. Instead of describing these old programs and procedures, which would have only academic

significance, I shall describe a step-by-step procedure for producing the same results using modern programs also written by me. The sequence of steps will be described as a series of small goals which lead to the final conclusion.

It must be assumed that the reader is familiar with ARTHUR commands, a text editor (TECO commands are described), the SPSSX command structure, a programming language capable of floating point arithmetic (FORTRAN here), the program SPIN (described in Appendix I), and the standard utilities of the VMS operating system. The ARTHUR-formatted file and the set of dependent variables are small, convenient data sets which do not have any particular chemical significance. To make the presentation as clear as possible, input and output to the computer will be left justified. Comments will be preceded by double colons and indented. The escape character (ASCII code 27 decimal), used heavily in TECO, will be indicated by <esc>.

Goal 1. Perform component analysis using ARTHUR
\$ REN K59.ART FOR001.DAT

```
    :: Copy K59.ART to FOR001.DAT as standard ARTHUR
    :: input (see listing of FOR001.DAT)
$ R ARTHUR
    :: Run ARTHUR
$ REN FOR001.DAT K59.ART
    :: Restore ARTHUR file to original name.
$ COPY APC.DAT FOR001.DAT
    :: Copy ARTHUR autoscaling and principal
    :: components instructions to standard ARTHUR
    :: input. (see listing of APC.DAT)
$ R ARTHUR
    :: Run ARTHUR
$ PURGE
    :: Clean up directory
$ COPY UTILIT.DAT FOR001.DAT
    :: Copy score calculation and output commands to
    :: standard ARTHUR output (see listing of
    :: UTILIT.DAT)
$ R ARTHUR
    :: Run ARTHUR

    Goal 2. Prepare ARTHUR output for SPSSX input.
$ COPY FOR007.DAT INP.ART
```

```
:: Copy ARTHUR-formatted score matrix produced by
:: ARTHUR to input file for REDRIT

$ R REDRIT

:: Run program which converts from FORTRAN-
:: formatted to normal data file.

$ COPY OUP.ART K59FA.ART

:: Copy back to a more informative filename

$ TEC K59FA.ART

:: Prepare file for SPSS by eliminating ARTHUR .
:: stuff.

*S999<esc><esc>

:: Search for line of nines.

*0,.K<esc><esc>

:: Kill all lines up to this point.

*T<esc><esc>

:: Type current line

*KT<esc><esc>

:: Kill line if it's all 9's. Repeat this
:: line until the line is no longer a line of 9's.

*<40DLLL><esc><esc>

:: Delete all the filenames and indices. Ignore
:: the error message when TECO gets to EOF.

*J<SE<esc>CCCI <esc>><esc><esc>
```

:: Insert blanks between the numbers so they don't
:: run together. Ignore error message at EOF.
:: Note the real right angle bracket (>) between
:: two of the <esc> characters.

*J<L44CI

<esc>L44CI

<esc>LI

<esc>><esc><esc>

:: Break the lines up and insert a blank line
:: between samples. This is partly for
:: readability and partly because SPSSX can't read
:: a line longer than 80 characters without
:: special formatting. Note: use carriage returns
:: between the separate lines, and note the real
:: right angle bracket (>) between two of the
:: <esc> characters.

*ZJ-20LS9999<esc><esc>

:: Search for trailing 999's

*0L.,ZK<esc><esc>

:: Delete 'em

*EX<esc><esc>

:: Exit TECO

\$ PURGE

:: Delete old files.

Goal 3. Run SPSSX stepwise linear regression

\$ SPSSX/OUT=K59IN.PRN K59IN.CTL

:: Read file into SPSSX and store in SPSS format
:: (see listing of K59IN.CTL for control file).
:: The listing for the run is stored in K59IN.PRN,
:: analogous to ARTHUR's FOR020.DAT. Print or
:: type it to taste.

\$ SPSSX/OUT=K59DVIN.PRN K59DVIN.CTL

:: Read file of dependent variables into SPSSX and
:: store in SPSS format (see listing of
:: K59DVIN.CTL for control file). The listing is
:: stored in K59DVIN.PRN. Print or type it.

\$ SPSSX/OUT=K59FDR.PRN K59FDR.CTL

:: Merge the files K59FA.SPS and K59DV.SPS in
:: SPSSX and perform stepwise multiple linear
:: regression with factors F1 through F20 as the
:: independent variables and dependent variables
:: DV1 through DV6 (see control file K59DV.SPS).
:: Print or type K59FDR.PRN.

Goal 4. Use output from SPSS in SPIN

\$ R SPIN

:: Run SPIN

What's the ARTHUR file name? K59.ART

What's the vector file? FOR007.DAT

Will you be using the full data matrix? Y

:: Standard SPIN beginning. SPIN responds with

:: the Main Menu (see Appendix I)

9

:: Select "Input, output or modify matrix of

:: rotations".

5

:: Alter matrix of rotations

Which vector (0 to quit)? 1

:: Select first vector

Which element (0 to quit)? 9

:: This example will show data entry for equation

:: 2 with six coefficients produced in SPSSX.

-88220

Which element (0 to quit)? 8

80067

Which element (0 to quit)? 7

62610

Which element (0 to quit)? 18

164119

Which element (0 to quit)? 10

59808

Which element (0 to quit)? 5

-29230

Which element (0 to quit)? 0

:: 0 ends data entry.

Which vector (0 to quit)? 0

:: More vectors could be entered if desired.

Will you be using the full data matrix? Y

:: Specific spectra could have been chosen. After

:: the full matrix is read in, the main menu

:: appears.

1

:: Choose factor spectrum. Any of the options

:: will work, but rotation should not be

:: performed.

Which two factors rotated? 1,2

:: Select factors 1 and 2 in order to plot number

:: 1.

How many degrees rotation? 0

:: No rotation. After the factor spectrum has

:: been plotted, the Main Menu appears.

6

:: Exit gracefully.