ORNL/TM-2002/16

PCR+ in Diesel Fuels and Emissions Research

MARCH 2002

Prepared by

H. T. McAdams
AccaMath Services
Carrollton, Illinois

R. W. Crawford
RWCrawford Energy Systems
Tucson, Arizona

G. R. Hadder
Oak Ridge National Laboratory
Oak Ridge, Tennessee

ORNL/TM-2002/16

PCR+ IN DIESEL FUELS AND EMISSIONS RESEARCH

H. T. McAdams
AccaMath Services
Carrollton, Illinois

R. W. Crawford
RWCrawford Energy Systems
Tucson, Arizona

G. R. Hadder
Oak Ridge National Laboratory
Oak Ridge, Tennessee

March 2002

Prepared by
OAK RIDGE NATIONAL LABORATORY
P.O. Box 2008
Oak Ridge, Tennessee 37831-6285
managed by
UT-Battelle, LLC
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

AAM	Alliance of Automobile Manufacturers
CFD	Commercial Fuels Database
CO	Carbon monoxide
DOE	U.S. Department of Energy
EGR	Exhaust gas recirculation
EPA	U.S. Environmental Protection Agency
FBP	Final boiling point
FM	Figure of Merit
HC	Hydrocarbons
HDD	Heavy Duty Diesel
HDEWG	Heavy Duty Engine Working Group
IBP	Initial boiling point
NO_x	Nitrogen oxides
OLS	Ordinary Least Squares
ORNL	Oak Ridge National Laboratory
PCA	Principal Components Analysis
PCR	Principal Components Regression
PCR+	Principal Components Regression Plus
PM	Particulate Matter
RFG	Reformulated Gasoline
SS	Sum of Squares
SWRI	Southwest Research Institute
T10	Ten percent evaporation temperature
T50	Fifty percent evaporation temperature
T90	Ninety percent evaporation temperature

ACKNOWLEDGMENTS

This research, performed under contract with the Engineering Science and Technology Division of Oak Ridge National Laboratory (ORNL), was sponsored by the U.S. Department of Energy Offices of Energy Efficiency and Renewable Energy, Fossil Energy, and Policy and International Affairs. ORNL is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The authors wish to acknowledge the guidance of Barry McNutt of the U.S. Department of Energy Office of Policy and International Affairs.

EXECUTIVE SUMMARY

In past work for the U.S. Department of Energy (DOE) and Oak Ridge National Laboratory (ORNL), PCR+ was developed as an alternative methodology for building statistical models. PCR+ is an extension of Principal Components Regression (PCR), in which the eigenvectors resulting from Principal Components Analysis (PCA) are used as predictor variables in regression analysis. The work was motivated by the observation that most heavy-duty diesel (HDD) engine research was conducted with test fuels that had been "concocted" in the laboratory to vary selected fuel properties in isolation from each other. This approach departs markedly from the real world, where the reformulation of diesel fuels for almost any purpose leads to changes in a number of interrelated properties.

In this work, we present new information regarding the problems encountered in the conventional approach to model-building and how the PCR+ method can be used to improve research on the relationship between fuel characteristics and engine emissions. We also discuss how PCR+ can be applied to a variety of other research problems related to diesel fuels.

ISSUES WITH STEPWISE REGRESSION

Stepwise regression is among the most widely used research methodologies for expressing the dependence of a response variable on several predictor variables. However, we take issue with the technique when it is used in an environment where the predictor variables are interrelated to an appreciable extent. In such cases, the interrelatedness leads to aliasing among variables that can confuse and confound efforts to identify the variables having the greatest effect on the response. The dispute does not involve the method of estimating coefficient values – whether Ordinary Least Squares (OLS) or another method – or whether the method is highly automated or not, but rather the emphasis on individual fuel properties as predictor variables and the use of stepwise procedures for selecting among them to build statistical models.

Diesel fuels are strongly affected by naturally-occurring relationships among the individual fuel properties, as are all diesel fuel and emissions data in which the relationships have not been artificially eliminated. In this realm, the influential factors are better described by *vector variables* representing combinations of the fuel properties. The PCR+ approach was developed as an alternative to stepwise regression for these reasons:

1. We believe that emissions are better described as a response to vector variables, characteristic features of fuels in which the properties act in concert, rather than as a response to single fuel properties acting in isolation.

2. The variables selected by the stepwise process for inclusion in a model can be arbitrary, inasmuch as there are multiple solutions that are essentially equivalent in explanatory power when gauged by common statistical measures such as R^2.

3. Stepwise regression does not redefine variables and therefore *cannot* correct the difficulties caused by interrelated predictors. It merely consolidates aliased effects of variables excluded from the model under the names of the variables that remain, thereby confusing the causal relationships between predictors and response.

4. Aliasing casts doubt on whether the final model selected by a stepwise process emphasizes the "most important" or the "right" variables. If it does *not*, then the model

will be misleading as a basis for fuel improvement, because it will misidentify the proper set of variables and will be incapable of making correct predictions when applied outside the data used in estimating the model.

We believe that analysis should be done in the space of eigenvectors, where the vector variables are defined to be mathematically independent and where model-building is subject to little or no ambiguity. Orthogonality of predictors eliminates the problems inherent in stepwise regression and provides a unique means for assessing the relative importance of fuel properties. Orthogonality also provides maximum discrimination between variables and tests of significance with maximum power.

For example, the Unified Model (U.S. EPA 2001) includes aromatics content, additized cetane, specific gravity, and the mid-point temperature on the distillation curve (T50) as predictor variables for HDD NO_x emissions. Aliasing among variables and the confounding effect it has on identifying the proper set of predictor variables are, we believe, the reasons that the Unified Model omits natural cetane. It seems counterintuitive to suggest that the natural cetane rating of a fuel has no effect on NO_x formation. It seems more plausible to suggest that the effect of natural cetane has been incorporated in the model coefficients for total aromatics and specific gravity.

That aliasing is present among interrelated predictors is well understood, but conventional statistical practice has focused primarily on problems where variables are collinear to the extent that a solution of the general linear model may not exist. Our concern, and the main target for the PCR+ approach, is *not* the set of computationally pathological problems, but rather those in which aliasing is present to the extent that a stepwise analysis risks misidentifying the proper set of predictor variables.

Further, natural correlations exist among diesel fuel properties as an expression of the characteristics of blendstocks and the effects of refining processes. It is the natural structure of correlations underlying diesel fuels that PCR+ attempts to identify and harness. In this environment, where fuel properties do not vary independently, it is more reasonable to believe that the eigenvector variables exert independent effects on emissions and that efforts to attribute these effects to the individual fuel properties will necessarily be frustrated.

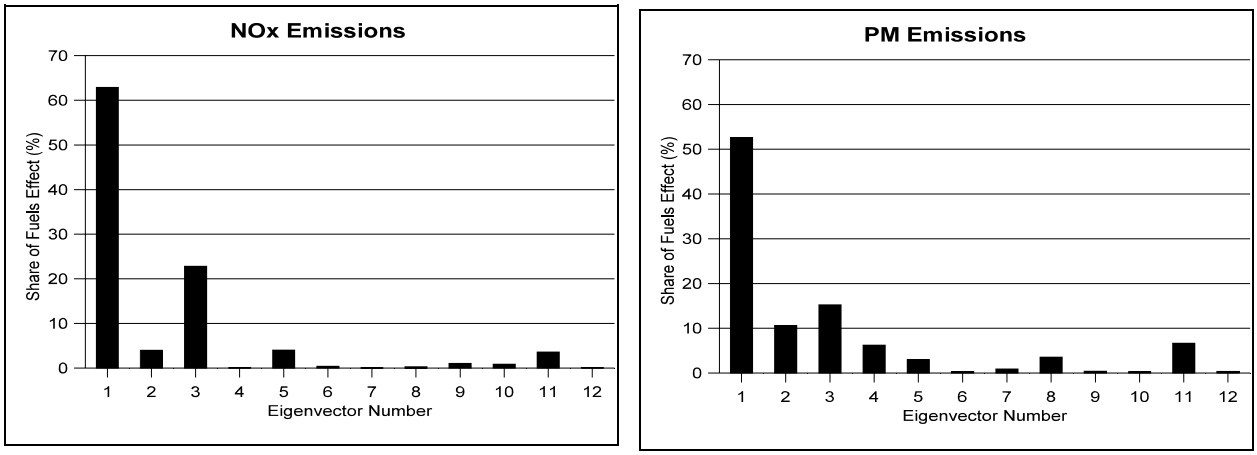
APPLICATION OF PCR+ TO DIESEL EMISSIONS

PCR+ is a distinct approach to fuels and emissions research aimed at a more natural interpretation of the relevant fuel factors influencing emissions. Because they act like mathematical blendstocks in describing fuels, the vector variables have been termed "eigenfuels." When applied to a large data set of emissions testing on HDD engines compiled by the U.S. Environmental Protection Agency (EPA), we find that:

1. Only six vector features are needed to explain nearly 95 percent of the differences among experimental fuels used in past research. The first five features – representing vector variations related to aromatics content, natural cetane, additized cetane, oxygen content, and sulfur content – are ones logically found in a data set developed to test for fuel effects on emissions. The sixth feature appears related to controlling the flash point of test fuels to commercial specifications.

2. When the vector variables are used as orthogonal predictors for emissions in regression analysis, we find that one vector stands out overwhelmingly as the most important influence on NO_x and PM emissions. This vector, also termed the "light cycle oil" vector, involves a *simultaneous* variation in aromatics content, natural cetane, and specific gravity with emissions effects that *cannot* be ascribed to any one

Figure ES.1. Eigenfuel Impact on HDD Engine Emissions



fuel property in isolation from the others. Figure ES.1 shows the impact of the vectors on HDD engine emissions.

3. Hypothetical scenarios for diesel fuel reformulation were evaluated to compare the predictive abilities of PCR+ and stepwise regression models. Three ways of varying the blendstock composition were identified from the characteristics of commercial diesel fuels; cetane additives and oxygenates also were considered. The comparisons presented in Table ES.1 make clear that the conclusions one draws on the variables affecting emissions, and the magnitude of their effects, will depend to a significant degree on the choice of analysis methodology.

Table ES.1. Summary of NO_x Predictions (percent reduction from average commercial fuel)

	PCR+ Model	EPA Unified Model
Blendstock Composition Change (Aromatics 33 ➜ 10 percent)		
Light Cycle Oil Vector	-9.1	-9.5
Hydroprocessed Heavy Distillate Vector	-8.5	-12.5
Straight-Run Light Distillate Vector	-9.5	-10.7
Additized Cetane + 10 numbers	-3.5	-2.7
Oxygen Content + 4 percent	2	none

Changing long-held views is not an easy task, and a simple example may be more powerful than lengthy argument. A recurring observation in the EPA analysis and the technical literature is the association of cetane number with aromatics and specific gravity and the attendant practical difficulty of separating the effects of these variables on emissions. Why not think of these three as one composite variable, as the PCR+ approach does? As long as one tries to partition this vector into three separate effects, there is a high probability of confusion and misattribution.

ADDITIONAL APPLICATIONS FOR PCR+ IN DIESEL RESEARCH

PCR+ offers benefits in areas of diesel research beyond the relationship of fuels to emissions. For example, it may be possible to improve the prediction of natural cetane rating for refinery blending operations using as predictors the vector features characteristic of fuels. A database of commercial diesel fuels currently in the marketplace was examined in this work to identify the vector features that describe real-world fuels.

When compared to commercial fuels, the experimental fuels used in past emissions research are found to be intentional recombinations of selected commercial fuel features for the purpose of creating *experimental* features that fit pre-selected experiment designs. That the experimental fuels are largely re-expressions of the commercial fuel characteristics lends credence, we believe, to our conviction that the eigenvectors of commercial fuels offer a more natural and fundamental set of variables for diesel fuels analysis.

Since the experimental features can be traced back to underlying characteristics of commercial fuels, it is possible, and we believe more productive, to base experiments on the commercial characteristics directly. Such experiment designs would be based on the variation in vector characteristics of fuels, rather than the variation in individual properties. Principles and examples of the role for eigenvectors in the design of diesel fuel experiments (without regard to the response variable of interest) are given in the report.

Although the motivation for PCR+ was as a technique for improving analysis of fuels and engine emissions data, eigenfuels can also be useful in solving a variety of problems related to fuel characteristics. The report demonstrates a methodology for conducting Monte Carlo simulations to generate synthetic, but realistic, fuels data. The utility of this capability ranges from estimating sample sizes for new testing to assessing the likely performance of statistical models when applied to new data.

Using eigenfuels as the building blocks of fuels, hypothetical fuels can be created that are indistinguishable from real fuels in terms of average values for fuel properties, the standard deviations, and the correlations among the properties. New and distinctive fuels data can be created under the researcher's control by modifying the parameters of the simulation or by selecting simulated subsets of particular interest. A hypothetical study is outlined in which a realistic slate of fuel characteristics is estimated for fuels that might be produced to meet a hypothetical standard of not less than 50 natural cetane rating.

Eigenfuels can also provide a framework for modifying fuels in the real world to achieve a predetermined objective, whether that be emissions reduction, cetane number control, or something else. The outline of such a framework is:

- Candidate fuel modifications are translated into their equivalent expressions in eigenvector terms
- A PCR+ model, formulated in terms of the eigenvectors, "scores" the candidate modifications in terms of the objective of interest
- Fuel modifications are selected and combined until a predefined change in the objective is achieved, while taking account of cost at each step in the process to reach an optimal result.

Because eigenfuel-based models are expected to be more accurate predictors than models based on correlated variables, a reformulation process based on eigenfuels should be subject to fewer limitations, inaccuracies, and inefficiencies. Whether the objective is regulatory or commercial, an eigenfuel-based process is more likely to achieve its expected results in the real-world, at lower total cost, than a system based on the individual fuel properties.

1. INTRODUCTION

1.1 BACKGROUND

Multiple regression is one of the most widely used methodologies for expressing the dependence of a response variable on several predictor variables. In typical usage, data consisting of a response variable and a number of potential predictor variables are compiled – often from a variety of sources – either from prior research or new data collection efforts. Customarily, a large number of multiple regression models are estimated, the most complex being the model that includes all available predictor variables. The end product is usually a simpler model resulting from the exclusion of some of the predictors that are deemed statistically non-significant.

Conventionally, the t-test of statistical significance, as applied to the regression coefficients, is the basis for excluding predictor variables from the model. When the predictor variables are interdependent (as is usually the case), the variable selection process becomes complex, however, and may require considerable iteration before a final version of the model emerges.

Stepwise regression is a procedure commonly used for this purpose, as incorporated in such statistical software packages as SAS, SPSS or SYSTAT. The procedure can be fully automated or may be exercised in an interactive manner in which the data analyst can examine the results of each step and interject judgment in determining subsequent steps. Ordinary Least Squares (OLS) is the optimizing methodology most commonly used to estimate the regression coefficients, but other estimation methods, such as Maximum Likelihood, are also used.

This conventional model building process is quite familiar to analysts involved in fuels and emissions research. Although it is widely used, we contend that stepwise regression is flawed in principle and is not appropriate for formulating a model for predicting heavy duty diesel (HDD) emissions in terms of the physical and chemical properties of fuels. These properties are inherently interrelated through the characteristics of diesel blendstocks and the effects that refining processes have on those characteristics.

For research purposes, efforts are often made to “break" the correlations among fuel properties through specialized blending of test fuels. Such efforts usually meet with only limited success and, in doing so, may create test fuels that are unrepresentative of fuels produced in the real world. On the other hand, efforts to understand the effects of individual fuel properties on emissions can be frustrated when test fuels are allowed to retain the naturally occurring relationships among properties (Lee *et al.* 1998). The frustration results from the inherent inability of stepwise regression to identify and separate – clearly and unambiguously – how individual fuel properties influence emissions in circumstances where an individual fuel property is related to many other fuel properties, sometimes strongly.

In an attempt to remedy these concerns, we have developed an alternative model-building approach that represents fuels in terms of combinations of fuel properties, these combinations being referred to as "eigenfuels." The eigenfuels are unique and mathematically independent characteristics of diesel fuels; they offer concise and natural descriptions of fuels and circumvent the confusion that accompanies fuel characterization in terms of interrelated, separate fuel properties.

The eigenfuel approach is based on PCR, in which the predictor variables derive from the eigenvectors of the correlation matrix of the original fuel properties. Our approach corrects past shortcomings of PCR as perceived in the statistical literature (see, for example, Hadi and Ling 1998). Accordingly, we identify our

approach as PCR+, the distinction being essential to prevent the perceived limitations of PCR from being an impediment to implementation of our approach. The PCR+ methodology is documented in an SAE paper (McAdams *et al*. 2000a) and a technical report (McAdams *et al*. 2000b) published by the ORNL and was described in a presentation (Crawford and McAdams 2001) at an EPA Workshop on diesel fuel effects on emissions.

In this report, we present new information regarding the problems encountered in using stepwise regression, and we discuss applications of PCR+ to a variety of research problems that include, but are not limited to, model building. Our dispute with stepwise regression is not concerned with the method of estimation of parameters (whether by OLS or Maximum Likelihood), nor with the degree of automation attending the stepwise process. Rather, we reject the notion that stepwise regression can find a meaningful and defensible way to identify emissions effects with separate fuel properties when the fuel properties are inextricably related.

Our primary concern is that the result of stepwise regression leaves the impression that the variables driving the emissions response have been identified and that those effects have been quantified. In reality, the named effects may be considerably more complex than their labels imply and could just as logically be assigned to alternative predictor variables. In short, the names may not necessarily reflect the facts of the matter.

1.2 BASIC TERMINOLOGY

We use the term "stepwise regression" as short-hand terminology to identify the conventional model-building approach that we believe to be flawed – i.e., the use of conventional regression techniques to assess diesel engine emissions when the predictor variables are physical and chemical properties of fuels that evidence strong relationships to each other. As has been stated, our issue is not with the method of estimating parameters, or with the degree of process automation. Nor would we object to the use of conventional regression techniques with the separate fuel property variables, *if the fuel properties occurred in nature without strong relationships to each other.* It is because the properties of diesel fuels exhibit strong relationships to each other that we argue for a different approach to regression analysis, and it is the *stepwise process of selecting among interrelated variables* that is at the heart of our dispute.

Other terminology comes into play in this report. The term "P-Space" refers to the variable space defined by the original, fuel property variables. When the variables in P-Space are interrelated, the conduct of regression analysis (by any means of estimation) typically implies a stepwise process of selecting variables (whether automated or not) for inclusion in the model. We do not dispute other uses of regression analysis in P-Space, including:

- The display of "all possible" regressions involving interrelated variables, when there is no attempt to select among the variables by this method
- The estimation of model parameters when variable selection has already been made by some other means.

The term "E-Space" refers to the space in which the predictor variables are orthogonal. In instances such as diesel fuels, the orthogonal variables are derived from the eigenvectors of the correlation matrix of the original fuel property variables and used to define E-Space.[1] Because the vector variables in E-space are

[1] When the original variables are independent, as may result from a planned experiment design, the correlation matrix is diagonal and the eigenvectors of the matrix are unit vectors. In this instance, P-Space and E-Space are synonymous and the orthogonal predictors are defined by the experiment design.

inherently independent, there is no ambiguity in regression analysis and no need to resort to stepwise procedures for variable selection. PCR+ is the application of regression analysis in E-Space, as advocated in this report, using OLS as the method for estimating model parameters. However, there should be no barrier to extending PCR+ to other formulations using the orthogonal predictors, such as Mixed Effects models that would be estimated with Maximum Likelihood methods.

1.3 ORGANIZATION OF REPORT

This report documents research performed during 2001. Portions of the work have been previously distributed in the form of working papers and presentations in forums that reached government and industry audiences. Section 2 describes the data on diesel fuel characteristics and HDD engine emissions that are used here for the development and demonstration of the PCR+ approach. The section also summarizes the EPA Unified Model resulting from that agency's work on diesel engine emissions. Section 3 presents our assessment of the issues with the conventional analysis approach based on stepwise regression when applied to the problem of diesel fuels and emissions. Section 4 presents the PCR+ approach as an alternative approach to the analysis of the problem. It summarizes the PCR+ methodology, presents empirical results on diesel fuel effects on emissions, and responds to issues and objections raised in regard to the PCR+ methodology.

Subsequent sections broaden the perspective on eigenfuels beyond its statistical foundations to include a range of applications. Section 5 examines the eigenfuel characteristics of current commercial diesel fuels using a fuels survey made available to this study. It also shows how experimental fuels used in past research compare to the actual characteristics of commercial diesel fuels. Section 6 then considers how an accepted set of eigenfuels, such as those of the commercial diesel fuels, can form a practical basis for the design of experiments. Section 7 demonstrates how simulation exercises based on eigenfuels can be used to answer a range of questions related to the potential characteristics of diesel fuels in a manner that takes full account of the interrelationships among fuel properties. Finally, Section 8 reviews how eigenfuels can be used to guide and measure the reformulation of diesel fuels.

Six appendices are included to amplify the information presented in the report. Appendix A compiles basic numerical results from analyses presented in the report. The five other appendices address in greater detail selected topics that are discussed in the report.

2. METHODOLOGY AND DATA

This section describes the primary data sets that are used to develop and demonstrate the PCR+ methodology in this report. It also describes briefly the EPA Unified Models for diesel engine emissions, which are referenced in this work as a point of comparison for emissions models developed using the PCR+ approach.

2.1 DIESEL FUEL AND EMISSIONS DATABASES

The PCR+ approach was originally developed during 1999 and 2000 using a database of some 280 individual emissions tests of HDD engines that had been compiled specifically for the purpose of supporting the methodological development.[2] The original database described fuels in terms of the twelve physical and chemical properties shown in Table 2.1. The original database differs from that used more recently by EPA in that the original database contained the breakdown of aromatics content into mono- and poly-aromatics species, cetane additives were found in only a small number of fuels, and no fuels containing oxygenates were present. The original data, and emissions models based on them, were used in early portions of the work presented in this report, most prominently in portions of the appendices. The use of the original data and models can be identified by reference to the original work and by the number of emissions tests (280).

Table 2.1. Fuel Properties in the Original HDD Emissions Database

Fuel Property	Units	Description
Natural Cetane	number	natural cetane number
Cetane Difference	number	cetane number increase due to additives
Specific Gravity	gm/cm^3	
Viscosity	mm^2/sec	at/near 40 degrees C
Sulfur Content	ppm	
Mono-Aromatics Content	vol percent	
Poly-Aromatics Content	vol percent	
IBP	Celsius	Initial boiling point
T10	Celsius	10 percent evaporation temperature
T50	Celsius	50 percent evaporation temperature
T90	Celsius	90 percent evaporation temperature
FBP	Celsius	Final boiling point

During the first half of 2001, EPA undertook a research project to assess the relationship between diesel fuel characteristics and engine emissions using all relevant data that could be identified (U.S. EPA 2001, SWRI 2001). One result of their work was the creation of a greatly enlarged database of engine emissions tests and

[2] See McAdams *et al.* 2000b, pp. 9-11 for documentation of the original database, and pp. 20-32 for the emissions models based on this data.

diesel fuel properties,[3] containing more detailed information on a wider range of diesel engines and fuels than had been previously available in any one place.

As is easily understood, the information available on engines and fuels varies throughout the EPA database depending on what was published in the original sources. EPA's analysis used varying subsets of the database depending on the particular group(s) of engine technologies being considered and the fuel descriptors that were required. A decision was made by EPA to use total aromatics content as a fuel descriptor, rather than the mono- and poly-aromatics speciation, in order to maximize the number of engine tests that could be used. We have adopted that decision in this work for the purposes both of increasing the available data and of maintaining a basis of comparison to the EPA work.

A consistent subset of the EPA database, consisting of approximately 70 percent of the data, was selected for use in this work. The subset spans ten different engine technology groups, including the dominant technology types on the road,[4] that were found in EPA work to share a common emissions response to fuels. The subset contains 906 emissions tests, out of the 1315 tests in the entire database, for which NO_x and PM emissions and a minimum of nine different fuel properties had been measured, as shown in Table 2.2. The fuel properties required are: natural cetane number, cetane difference, specific gravity, sulfur content, total aromatics content, T10, T50, T90, and oxygen content. (Additional fuel properties, including viscosity, IBP, and FBP, are available for many of the observations in the selected subset.) As a matter of shorthand, this consistent subset is often referred to as the EPA database in this report.

Table 2.2. Required Fuel Properties in the EPA Database

Fuel Property	Units
Natural Cetane	number
Cetane Difference	number
Specific Gravity	gm/cm^3
Sulfur Content	ppm
Total Aromatics Content	volume percent
T10	Fahrenheit
T50	Fahrenheit
T90	Fahrenheit
Oxygen Content	weight percent

Two different ways of describing fuels are used in this report. The first is adopted from the EPA work in order to facilitate comparison to their results; the second is a modification of the twelve fuel properties used in the original work on PCR+. EPA selected the nine major fuel properties shown in the table and added to them the squared terms for natural cetane, cetane difference, and aromatics content to produce a total of twelve predictor variables.

[3] See U.S. EPA 2001, pp 6-22 for a full description of the database.

[4] The dominant technology groups are Groups T and F, consisting of low-speed turbocharged engines below 500 horsepower with electronic fuel injection (Group T) and mid-speed turbocharged engines of any horsepower rating with mechanical fuel injection (Group F). Engines in these groups are not equipped with EGR systems, oxidations catalysts, or particulate traps. Some analysis specific to Group T is presented in this report.

The selection of fuel properties was narrowed to nine by EPA based on its assessment of the literature on the properties thought to influence emissions. The presence of the square terms in the variable set allows for the possibility that the emissions response could exhibit nonlinear behavior, such as saturation (diminishing returns) as the fuel property approaches high or low values.

A different list of twelve fuel properties, without squared terms, has been used in other portions of this report for continuity with the original work on PCR+:

- Natural cetane
- Cetane difference
- Total aromatics
- Viscosity
- Specific gravity
- Sulfur content
- Five points (IBP, T10, T50, T90, and FBP) on the distillation curve
- Oxygen content

This choice of variables adopts total aromatics content, in place of the earlier mono- and poly-aromatics breakdown, and adds oxygen content to the original list of variables. These two sets of variables are merely alternate ways of describing fuels that may be convenient and appropriate in different contexts. There is no intent to argue that one set or the other is to be generally preferred for emissions analyses or other purposes.

2.2 EPA UNIFIED MODEL

One product of EPA's work on engine emissions was the estimation of models for HC, CO, NO_x, and PM as a function of engine technology and diesel fuel characteristics, termed the Unified Model. EPA summarizes the Unified Model as follows:

> *In this approach, forward stepwise regressions were carried out on the database as a whole, but technology group-specific effects were also permitted to enter the model if significant. We also made efforts to more properly account for engine variability and the impact that such variability should have on the statistical significance of fuel property coefficients. The resulting Unified Model is the model that we are proposing in this staff discussion document as a means for predicting the impact of changes in diesel fuel properties on emissions.* (U.S. EPA 2001, p. 27)

The EPA Unified Model was presented at an EPA Workshop[5] on diesel fuel effects and critiqued by government and industry representatives. In this report, we use the EPA Unified Model as an example of the application of stepwise regression techniques in fuels and emissions research – both in the discussion of statistical issues and in the presentation of empirical results. However, we want to note that the issues we take with the EPA work are criticisms of the genre of stepwise regression analysis in this area and not a criticism specific to the EPA analysis, which was thoroughly conducted in its entirety.

---

[5] *Diesel Fuel Effects on Emissions Workshop*. National Vehicle and Fuel Emissions Laboratory, Ann Arbor, MI. August 28, 2001.

Portions of this report assume that readers have a basic familiarity with the work conducted by EPA to develop the Unified Model. The Unified Model is compared to models based on PCR+ in a number of ways, including the selection of fuel property variables used to characterize emissions and the numerical predictions of emissions effects based on fuel property changes. Because technology effects are not directly germane to the purpose of this report, the discussion and comparisons focus on the so-called "common terms" of the Unified Model, which represent the emissions model developed for the 10 engine technology groups that were found to share a common response to fuels, and which represent the large majority of diesel engines on the road today. Table 2.3 lists the variables and the NO_x and PM coefficients of the EPA Unified Model for these groups. The Unified Model is formulated in terms of the logarithm of emissions in gm/bhp-hr. Fuel variables that were considered, but not included in the models, are indicated in the table with hyphens.

Table 2.3. Emissions Coefficients for EPA Unified Model [a/]

Fuel Property	Units	log(NO_x)	log(PM)
Natural Cetane	number	–	-0.004521
Natural Cetane^2	number^2	–	–
Cetane Difference	number	-0.002779	-0.04825
Cetane Difference^2	number^2	–	–
Aromatics Content	vol percent	0.002922	0.002157
Aromatics Content^2	vol percent^2	–	–
Sulfur Content	ppm	–	0.00008386
Specific Gravity	gm/cm^3	1.3966	2.3708
T10	Fahrenheit	–	–
T50	Fahrenheit	-0.0004023	–
T90	Fahrenheit	–	–
Oxygen Content	wt percent	–	-0.07193
Natural Cetane x Cetane Difference	number^2	–	0.001009

[a/] Common terms applicable to the large majority of diesel engines.

3. ISSUES WITH STEPWISE REGRESSION

The EPA assessment of diesel fuel effects on emissions is a recent example of the use of stepwise regression in fuels and emissions research. The process employed by EPA in the effort to construct a statistical model of diesel fuel effects was both thorough and complex in the number of fuel properties and interactive terms considered in the models and in how the model parameters were selected and estimated. Reduced to its basics, the process employs stepwise regression techniques to identify potentially influential variables and to test their statistical significance in building emissions models. Emission response coefficients were estimated using the SAS Mixed Effects procedure, rather than the more commonly used OLS method.

While thoroughly conducted in its entirety, we take issue with some of the study's conclusions regarding the identity of fuel properties that influence emissions and the magnitude of their effect, because the analysis was conducted in an environment of substantial correlation among the individual predictor variables. Problems with stepwise regression in this environment include the following:

- The choice of variables in a final model can be arbitrary, inasmuch as there are multiple sets of variables that are essentially equivalent as predictors according to common statistical measures
- Aliasing present among inter-related predictors casts doubt on whether the final stepwise equation emphasizes the "most important" or the "right" variables
- If a stepwise model fails to select the correct set of variables, then the equation may be misleading as a basis for fuel reformulation, since it will be incapable of making accurate predictions for postulated "improved" fuels.

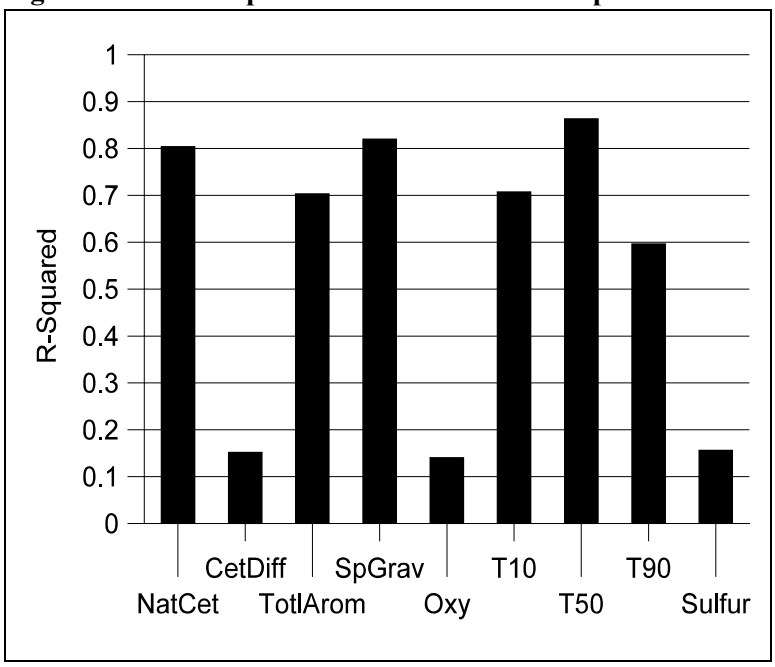
3.1 RELATIONSHIPS AMONG FUEL PROPERTIES

The heart of our dispute with stepwise techniques is that the analysis of fuel effects on emissions will be complicated and confused when any appreciable degree of correlation exists among the explanatory variables. While the association of physical and chemical properties of automotive fuels is well known, the degree of interdependence may be surprising to some. In fact, the EPA diesel emissions data set is strongly affected by relationships among the individual fuel properties, as are all other diesel fuel and emissions data in which the naturally-occurring relationships among properties have not been artificially eliminated.

Figure 3.1 demonstrates the degree of inter-relatedness that exists in the data. The measure of inter-relatedness is the R^2 statistic that is obtained when each fuel property is treated as a response variable and regressed against all other properties. One sees that only cetane difference, oxygen content and (surprisingly) sulfur content are relatively independent of the other properties. Natural cetane – a measure of a fuel's ignition characteristics, rather than a physical or chemical property – is strongly related to the other fuel properties . Overall, six of the properties (natural cetane, total aromatics, specific gravity, T10, T50 and T90) can be largely explained or predicted as a function of other properties.

It should not be a real surprise that analysis based on redundant variables such as these will have difficulty determining which variables have an effect on emissions and what subset of variables is best used for

Figure 3.1. Interdependence of Diesel Fuel Properties



predictions. This difficulty will be inherent whenever the correlations are more than modest. The specific problems encountered in using correlated predictors are:

- No unique definition can be given to statistical significance, since this can be judged only conditionally based on the other variables included at that stage in model development. Statistical significance of any given term will change whenever other variables enter or exit the model.
- Parameter estimates made in the presence of appreciable correlations are subject to variance inflation, where amount of inflation in the standard errors depends on the degree of correlation that is present. Variance inflation reduces the ability of the analysis to properly identify effects that are actually present in the data, in the same manner as if less data were available, thereby increasing the risk of incomplete models (i.e., missing one or more influential variables).

These two problems are well-known, but their consequences may be underestimated and overlooked because stepwise techniques have become so familiar and widely used. It may appear that there is no alternative to the conditional test of significance – perhaps even that statistical significance has no practical meaning except in relation to the list of other variables considered – and that the loss of power for detecting influential effects is something best addressed by increasing sample size. We dispute the contentions that the consequences are so benign and that no practical alternative exists, and we offer PCR+ as an alternative.

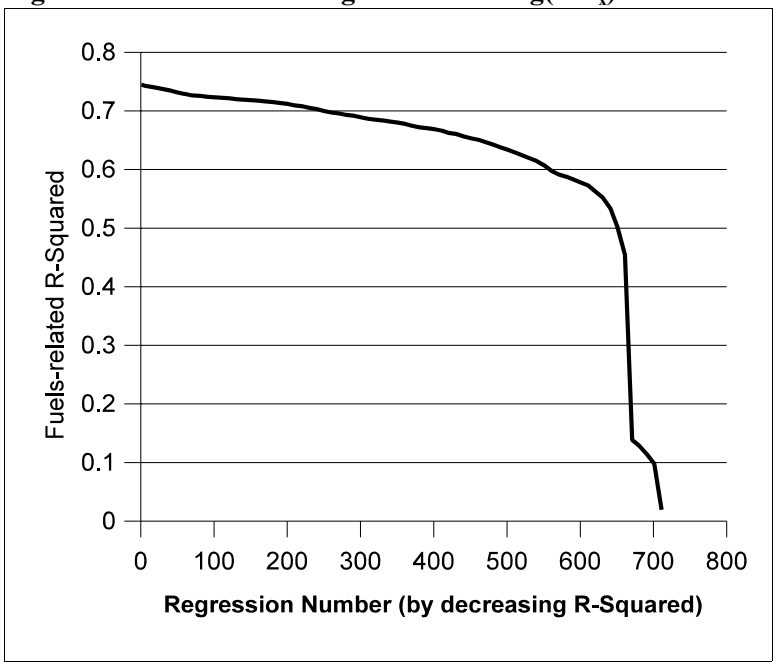
3.2 THE CONSEQUENCE OF CORRELATIONS

One consequence of having substantial correlations among predictor variables is that there will, in general, be many difference combinations of variables that provide good predictive power. This occurs because the exclusion of one predictor from the model can be compensated, to varying degrees, by other variables to

which the predictor is correlated. This is termed "aliasing" among variables in that a selection of two or more variables can account for the effect of another variable.

There are *4095* different models that can be formed from twelve fuel variables, in regressions that also contain controls for individual engine effects. Here, at least, it is possible to estimate each of the 4095 models – the "all possible" regressions approach – to illustrate the ambiguity inherent in models based on interrelated predictors. Of the 4095 models in which the dependent variable is log(NO_x), 711 are ones in which all included fuel terms are statistically significant at the 0.05 level. Figure 3.2 plots the explanatory power of these significant models using the "fuels-related" R^2 statistic, which is the R^2 of the fuel variables once the explanatory power of individual engine terms has been removed. The sharp rise at model 664 is the first entry of natural cetane as a variable. By itself as the only fuel variable, natural cetane produces a fuels-related R^2 of 0.43. From there, the curve of increasing R^2 bends over at about model 650 and begins a slow increase to its maximum of 0.745.

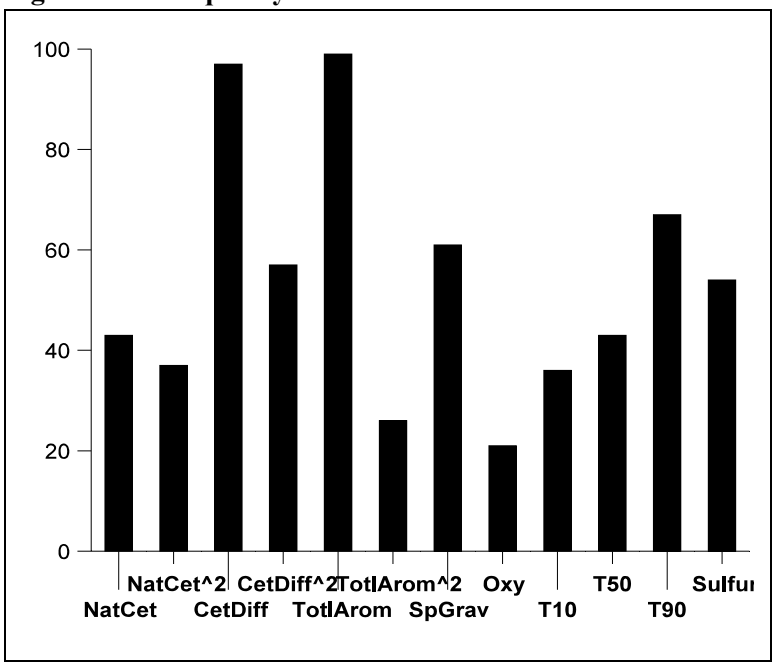
Figure 3.2. All Possible Regressions for log(NO_x)



If one were concerned with only the best 100 models, there is little difference in their fit to the data, at least as measured by R^2. The best R^2 is 0.745 and the 100th best R^2 is 0.723; it might seem that any of these models could predict well.

While R^2 may be about the same, the 100 models differ greatly in the number of terms included, ranging from four fuel variables to nine, and in which variables are selected. Figure 3.3 shows the frequency with which the twelve fuel property variables are included in the 100 best models. There is little doubt that cetane difference and total aromatics are important variables, because they are included in all but a few of the models. The case is a little less clear for other variables. For example, cetane difference^2, specific gravity, T90, and sulfur content are contained in more than half of the models, while four other variables (natural cetane, natural cetane^2, T10, and T50) are included in a large minority of the models. Only two variables (total aromatics^2 and oxygen content) are included in fewer than one-third of the models.

Figure 3.3. Frequency of Terms in 100 Best Models



The problem with stepwise-style variable selection using interrelated predictors is the large number of good models that can differ substantially in the identity and number of fuel terms. The absence of a variable from a model can be due to the fact that its effect is accounted for by other variables present in the model, and not because the variable is without effect. Relatively small differences in the data set can lead the analyst to select a different model with a different set of predictors.

3.3 THE EFFECT OF ALIASING ON REGRESSION COEFFICIENTS

The existence of so many near-equal choices means that the one model chosen as the result of a stepwise process does not provide a unique answer to the question of which variables influence the response. The presence of aliasing often means that not all of the influential variables can be included in the model with acceptable levels of statistical significance. Therefore, the stepwise model-building process becomes an effort to find a subset of variables that passes the conditional test of statistical significance and provides good predictive power, as judged by the performance of the model in predicting the *observed* responses.

Consider a case where correlated variables A, B, and C all have an effect on the response variable but cannot all be included in a predictive model with acceptable levels of statistical significance. Perhaps only A can pass the t-test for significance, while B and C fail, when all three variables are included in the model. But, if C is excluded, then B will appear significant by virtue of "picking up" contributions from C. When one or more influential variables are omitted from a model, the response coefficients for the remaining variables will "pick up" contributions from excluded variables with which they are correlated. This means, in the example above, that the coefficients estimated for variables A and B in the model will include contributions from the correlated variable C that was excluded.

As a result, we do not know either the true identity or the influence of a variable when building models with correlated predictors, since the estimated coefficients typically will include aliased contributions from excluded variables. Also, we do not necessarily know that the variables chosen by the stepwise technique

are the only ones driving the response. Some may argue that this is irrelevant because the excluded variables were judged to be non-significant. Nonetheless, their contributions are still present computationally and may be strong enough to "tip the scales" so that the exclusion of a non-significant variable enables another non-significant variable to become significant.

Let us look at aliasing using a regression for the dominant HDD engine technology group T that was estimated for demonstration purposes. As shown in Table 3.1, a model consisting of all twelve EPA fuel variables was initially estimated for log(NO_x) in a format where individual engine effects have been removed from the data. With all fuel terms included, seven of the twelve variables are found to be statistically significant at the 0.05 level, as indicated with asterisks. The five variables failing the t-test are then dropped, and the model is re-estimated. After selection, the coefficient estimates and t statistics change for each of the retained variables in the subset model, but the difference is most marked for three – aromatics content, specific gravity, and T50, as shown in boldface.

Table 3.1. Effect of Variable Selection on Regression Coefficients

	Before Selection		After Selection	
Fuel Property	Coefficient	t value	Coefficient	t value
Natural Cetane	-0.0077	0.58		
Natural Cetane ^{2}	-0.0042	0.31		
Cetane Difference	-0.0289	6.15*	-0.0273	5.76*
Cetane Difference ^{2}	-0.0127	2.78*	0.0122	2.62*
Total Aromatics	**0.0324**	**5.24***	**0.0248**	**10.6***
Total Aromatics ^{2}	-0.0098	1.81		
Specific Gravity	**0.0106**	**3.20***	**0.0203**	**8.61***
Oxygen Content	0.0053	3.68*	0.0055	3.74*
T10	0.0104	4.52*	0.0103	4.82*
T50	**-0.0104**	**3.05***	**-0.0173**	**7.57***
T90	0.0021	0.95		
Sulfur Content	-0.0021	1.43		

The coefficients estimated for a regression model can be decomposed into two parts: (1) the part that originates with each variable; and (2) the parts that are "picked up" from other variables that have been excluded and with which the retained variables are aliased. The method involves computation of the "alias" or "bias" matrix as shown in Appendix C. When applied to the coefficient estimated for aromatics content in the seven-term model, we find that the coefficient includes contributions *from all of the five other variables that were excluded*, as shown in Table 3.2. The largest aliased contribution is from the quadratic term in aromatics content, which has a high correlation to the linear term. But the other variables, including natural cetane, T90 and sulfur content also contribute to the re-estimated aromatics term.

Table 3.2. Aliasing of Total Aromatics to Other Variables

Total Aromatics Coefficient (full model)	0.0324
+ Contributions from	
Natural Cetane	0.0016
Natural Cetane^2	0.0007
Aromatics^2	-0.0103
T90	0.0005
Sulfur Content	-0.0001
	0.0248
Aromatics Coefficient (subset model)	0.0248

Table 3.3 presents the aliasing of the specific gravity term in the model. The specific gravity coefficient nearly doubles between the full and subset regressions because it picks up strong contributions from the natural cetane variables (linear and quadratic). The other excluded variables make smaller contributions to the specific gravity term. What one originally thought was the effect of specific gravity is now seen to include the effects of natural cetane and its square.

Table 3.3. Aliasing of Specific Gravity to Other Variables

Specific Gravity Coefficient (full model)	0.0106
+ Contributions from	
Natural Cetane	0.0058
Natural Cetane^2	0.0034
Aromatics^2	0.0013
T90	-0.0004
Sulfur Content	-0.0004
	0.0203
Specific Gravity Coefficient (subset model)	0.0203

The aliasing shown here comes about because the predictor variables are related to each other. The subset model (after selecting terms) actually includes computational contributions from *all of the variables*, whether those contributions are separately identified or not, and the variables included in the model stand for more than just themselves. As a result, the coefficients in the subset model are "biased" relative to the full model. The bias will be small when the terms excluded have computationally negligible effect on the response variable. The bias will be large whenever the model excludes, for any reason, aliased terms that carry an appreciable effect on the response.

The aliasing will be somewhat different in the EPA Unified Model, compared to this simple demonstration, but aliasing is present nevertheless. If cetane difference, aromatics content, specific gravity, and T50 – the four primary variables of the EPA NO_x model – are not the only ones driving the response, then there is not a solid basis for predicting the effects of varying fuel formulations simply by noting the effects of varying these four variables. It seems counterintuitive to suggest that the natural cetane rating of a fuel has no effect on NO_x formation in engine emissions. It seems more plausible to suggest that the natural cetane effect has been incorporated in the coefficients for total aromatics and specific gravity

A further problem with models based on correlated predictors is that variance inflation increases the risk of failing to detect an influential variable, compared to the risk faced with independent variables. When one or more influential variables are omitted, the resulting model can be misleading in regard to which variables are influential, since exclusion from the final model does not necessarily imply the absence of effect, and coefficient estimates can represent something different from the individual effect of the variable to which they are assigned. The resulting model is "tuned" for the aliasing present in the data set used for estimation. It may predict well on this "training" data, but perform poorly when applied to data having a substantially different correlation structure.

In contrast, a model based on eigenvectors, defined to be independent predictors, will omit an influential vector variable only when the sample size is inadequate to detect an effect of its size. The remaining vector variables are unaffected by exclusion of the one term, and their coefficient estimates are unbiased with respect to the true population effects for those terms. The resulting model will be inadequate with respect to an influential term that could not be statistically detected, but it is otherwise unbiased by exclusion of the term.

3.4 THE EFFECT OF ALIASING ON PREDICTIVE POWER

The limitations of models based on correlated predictors are of more than mere theoretical concern and can, in fact, result in appreciable loss of predictive ability when such models are exercised in real-world applications. Ironically, what appears to be a quite satisfactory model may fail in the very purpose for which it was intended – namely, to interpolate the response at points in the treatment space where no direct observations are available.

It is instructive to think of the observations in a data set as a function whose domain of definition is a discrete set of points in N-dimensional space. In the case of emissions as a function of fuel properties, the domain points are simply the test fuels as defined by N fuel properties such as cetane number, cetane difference, aromatics, specific gravity and so on. The purpose of a regression model is to provide a means by which emissions may be estimated for fuels not represented in the data set. Unfortunately, the conventional approach to evaluating the worth of such a model is to determine how well the model predicts emissions for the known fuels, not for fuels that have not yet been tested. The following will demonstrate that good prediction for the known fuels does not necessarily assure good prediction for postulated fuels.

Mathematically, the problem can be viewed as one of extending the domain of the emissions function from a discrete set of points in N-dimensional space to a N-dimensional continuum. It is well known, however, that a finitely defined function can have a multiplicity of extensions, and it is that fact that casts doubt on the applicability of a regression model outside the data set on which it is based. The data set on which a model is based will be referred to as the "training set," while the corresponding continuum will be referred to as the "extension set." To demonstrate the dilemma that may arise, we consider two models exhibiting essentially equivalent performance on the training set and proceed to show that the models may perform quite differently on the extension set. Models based on eigenvectors are less subject to such inconsistencies than are models based on interrelated variables.

Two of the 4095 possible regression models detailed in Appendix B are used for demonstration purposes. The models are based on engine-corrected log(NO_x) emissions for 480 engine tests from technology Group T of the EPA database. The regression coefficients and their t-values, together with the model SS and R^2 statistics are tabulated in Table 3.4. Note that the regression coefficients of both models are significant at better than the 0.05 level, and the models are very similar with regard to model SS and R^2. It remains to be seen only how well the two models agree on a fuel-by-fuel basis.

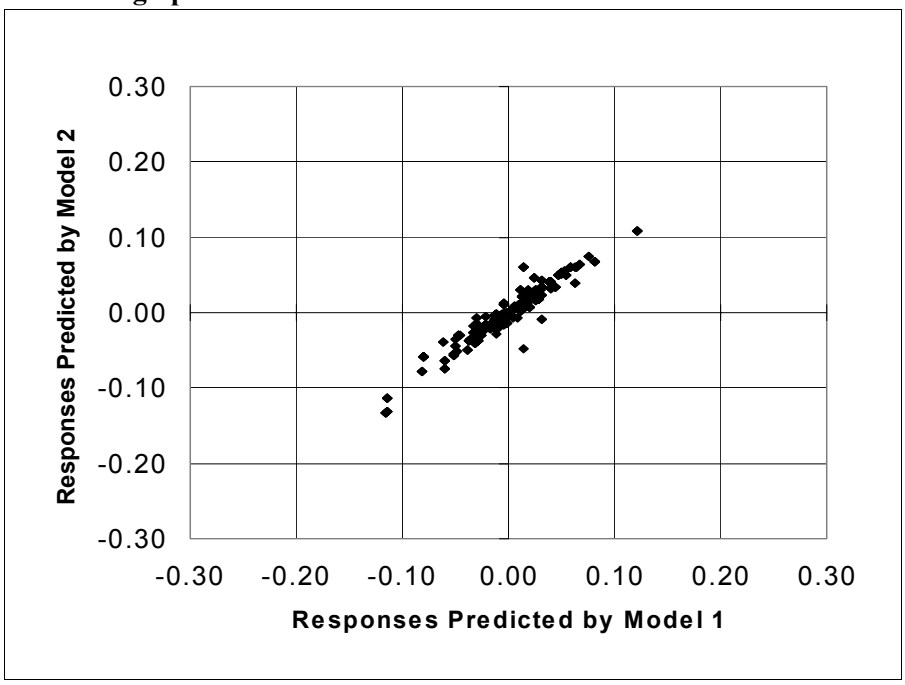
Table 3.4. Comparison of Two log(NO_x) Models Based on Fuel Property Variables

	Model 1		**Model 2**	
Model SS	0.6758		0.6763	
R^2	0.6116		0.6120	
Fuel Property	Regression Coefficient	t value	Regression Coefficient	t value
Natural Cetane	n/a	n/a	-0.0177	8.29
Cetane Difference	-0.0160	10.80	-0.0161	11.06
Aromatics	0.0243	10.73	0.0189	8.28
Specific Gravity	0.0208	8.74	0.0067	2.46
Oxygen Content	0.0050	3.42	n/a	n/a
T10	0.0103	4.75	0.0047	2.90
T50	-0.0169	7.33	n/a	n/a
Sulfur Content	-0.0033	2.24	n/a	n/a

Accordingly, each of the models was exercised to compute the predicted response for each of the 480 tests in the data set. Then, the predictions from Model 2 were plotted against the predictions from Model 1, as shown in Figure 3.4. For perfect agreement, the points should fall ideally on a straight line with 45° slope. It is seen that the points deviate minimally from this ideal. Clearly, there is little difference in the predictions made by the two models, so long as those predictions are constrained to the training set.

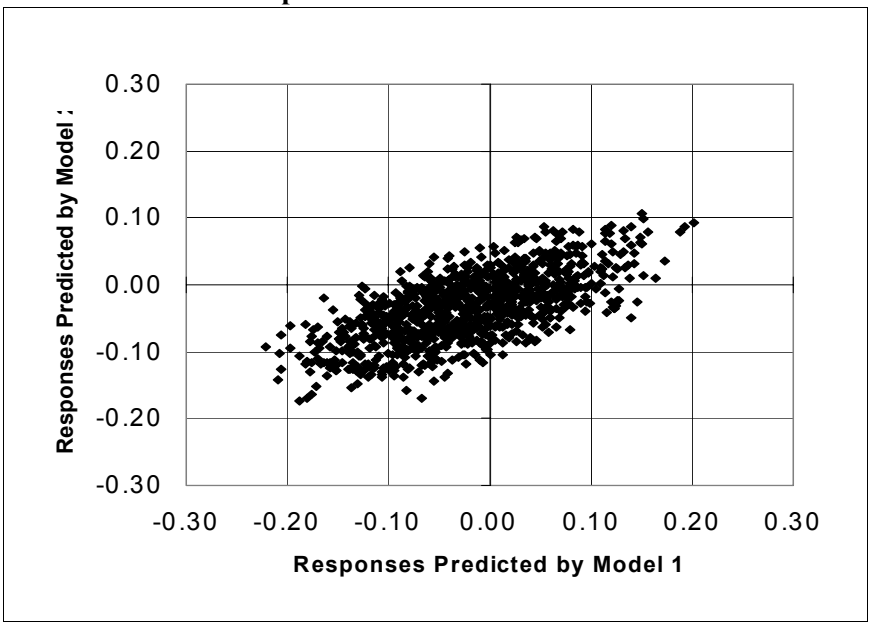
We now employ a random-balance approach (McAdams 1995) to construct an extension set. First, we determine the minimum and maximum values for each of the fuel variables. Then, we generate 1000 values for each of the fuel variables by random sampling from a uniform distribution scaled to have the maximum and minimum values that were observed in the training set. The extended set of points in the fuel-property space has the property that all predictor variables are essentially uncorrelated. Whereas the training set exhibits only specific combinations of fuel property values between the observed minima and maxima, the extension set exhibits, for each fuel variable, a distribution that is virtually continuous between the two extremes.

Figure 3.4. Comparison of Predictions by Two Fuel-Property Models in Training Space



As in the case of the training set, the two models were used to compute the predicted emissions for each of the 1000 randomly generated "fuels." Then, the predicted values from Model 2 were plotted against the predicted values for Model 1, as shown in Figure 3.5. Though the points roughly follow a line of unit slope, they constitute a broad band rather than a we1l-defined pattern. There is considerable disagreement between predictions by the two models in extension space, in spite of their agreement in training space.

Figure 3.5. Comparison of Predictions by Two Fuel-Property Models in Extension Space



A corresponding comparison was made of two models based on eigenvectors rather than on the interrelated fuel-property variables. As before, the two models were selected from among those detailed in Appendix B. The models were selected to have essentially equivalent performance on the training set; their regression coefficients, t-values, model SS and R^2 values are tabulated in Table 3.5. Note that all regression coefficients are significant at better than the 0.05 level and that R^2 and the model SS are strictly comparable for the two models.

Table 3.5. Comparison of Two log(NO_x) Models Based on Principal Components

	Model 1		**Model 2**	
Model SS	0.6755		0.6764	
R^2	0.6114		0.6121	
Principal Component	Regression Coefficient	t value	Regression Coefficient	t value
1	-0.0150	22.94	-0.0150	22.06
3	-0.0152	14.55	-0.0152	14.56
5	-0.0072	5.04	-0.0072	5.04
6	n/a	n/a	0.0037	2.18
9	0.0138	3.09	n/a	n/a
10	0.0237	3.59	n/a	n/a
11	n/a	n/a	-0.0352	4.32

Predicted responses for each of the training-set fuels were computed using each of the two eigenvector models. Then, the predictions by Model 2 were plotted against the predictions by Model 1, as shown in Figure 3.6. As is evident, the two sets of predictions track each other with only minimal deviation from a unit-slope straight line. To complete the comparison, a set of 1000 randomly selected weights were drawn for each principal component from a uniform distribution scaled so as to duplicate the maxima and minima of the eigenvector weights in the training set. Predicted emissions for each of the 1000 randomly generated "fuels" were computed by the two models, and predictions by Model 2 were plotted against predictions by Model 1, as shown in Figure 3.7. Note that the points tend to follow a 45° straight line with only slightly more deviation than in the case of the training set.

Figure 3.6. Comparison of Predictions by Two Principal Components Models in Training Space

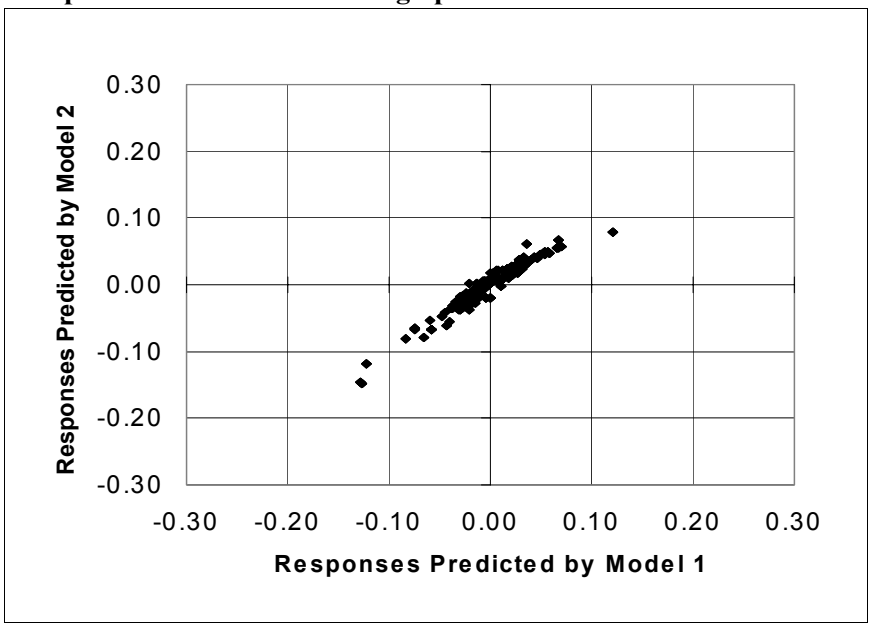
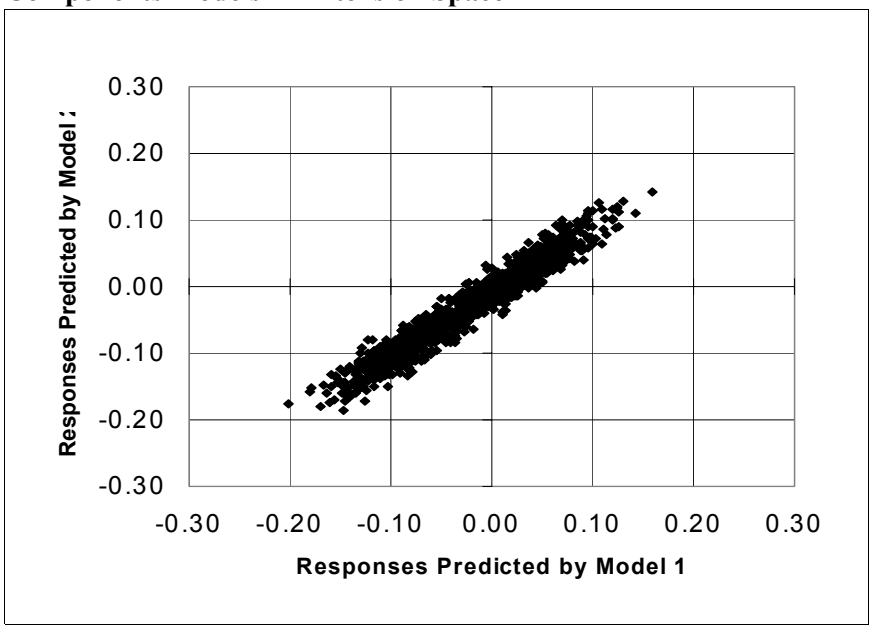


Figure 3.7. Comparison of Predictions by Two Principal Components Models in Extension Space



Another assessment of the comparable performance of models based on fuel-property variables and principal components is afforded by computing the difference between predictions of models 1 and 2 in each case and compiling statistics on the minimum, maximum, and standard deviation of those differences, as shown in Table 3.6. For the models based on fuel-property variables, the magnitude of the extreme differences (minimums and maximums) doubles in moving from training space to extension space, while the width of the distribution of differences increases by more than a factor of five. For the models based on principal components, however, the extreme differences increase by less, and the width of the distribution of differences only doubles.

Table 3.6. Summary of Model Performance in Training and Extension Space
(Comparison of Predictions by Two Models of Each Kind)

	Training Space	Extension Space	Percent Difference
Fuel Property Models			
Minimum Difference	-0.0470	-0.1142	+143
Maximum Difference	0.0627	0.1889	+201
Standard Deviation	0.0099	0.0565	+470
Principal Components Models			
Minimum Difference	-0.0243	-0.0416	+71
Maximum Difference	0.0409	0.0537	+31
Standard Deviation	0.0093	0.0175	+88

It seems evident that models based on principal components perform more consistently than do models based on interrelated predictor variables. The importance of this statement can be better appreciated if one realizes that the first model of each type can be postulated as the "true" response of NO_x to fuel variables and the second model as the empirical approximation derived from regression analysis of the training set. It is evident that the eigenvector model is a better estimator of the true response than is the model based on interrelated variables.

The relatively poor performance of a regression model based on fuel-property variables is a consequence of the difference in the correlation structure for variables in the training set and in the extension set. A model based on correlated predictors will “tune” the coefficients to match the aliasing present in the training data set. The “tuning” falters whenever the aliasing differs in the extension data set. In the case of the eigenvector models, such a difference does not occur, because both the training set and the extension set are essentially orthogonal.

3.5 COMMENTS ON THE EPA UNIFIED EMISSION MODELS

The thrust of our concerns regarding stepwise regression should make clear that we disagree with the methodological approach used by EPA in building the Unified Model. We are concerned that the predictive variables included in the models are fewer in number than the full list of fuel properties that influence emissions and that, by omitting one or more influential variables, such as natural cetane, the models may be misleading and inaccurate as predictive tools.

We hold these concerns most strongly in regard to the potential use of the Unified Model in estimating emissions reductions over a wide range of diesel fuel reformulations, in a manner similar to how the Complex Model for Reformulated Gasoline (RFG) is used (DOE, 1994). If one or more influential variables are omitted from the model, but are accounted for indirectly through aliased variables, the model may seem to be right in training space, but refiners and fuel blenders will receive incomplete and potentially misleading guidance on the appropriate strategies for reformulation in extension space.

Such use would advance an inefficient system for seeking emissions reductions, as refiners and blenders pursue fuel modifications with potentially different aliasing among fuel properties than those on which the predictive model was built. For example, we have seen that the specific gravity term in the Unified Model for NO_x contains aliased contributions from natural cetane that double the regression coefficient. If a refiner were to modify specific gravity in a way that had a little effect on natural cetane, then a smaller emissions change would take place than the Unified Model would predict. Conversely, if a refiner could modify natural cetane with little or no effect on specific gravity, then an emissions change might take place under circumstances where the Unified Model would predict none. While the EPA Unified Model is an advance in the understanding of diesel fuel effects on emissions, we believe that further work is needed to resolve methodological issues regarding the model-building process and their impact on the model's predictive capabilities.

3.6 SUMMARY

The highlighted difficulties in working with correlated predictors are well known, but perhaps under-estimated by many analysts in applying stepwise regression techniques. Often, the conventional wisdom is that stepwise regression techniques can be used successfully if the degree of collinearity among variables does not exceed some threshold. Condition Number[6] is often cited as an appropriate guard against such circumstances. We disagree, however, viewing the Condition Number (or Index) as primarily a measure of the computational difficulty faced in the solution of linear equations and the resulting loss of precision.

Our concern, and the main target for the eigenfuel approach, actually is NOT the set of problems that are computationally pathological – i.e., situations where the computations will lose precision. It is our view that the confusion among variables that results from aliasing and the susceptibility to misleading analytical results set in much sooner. The question is really, "How low must the Condition Number be for one to conclude that aliasing is of no appreciable effect." We are unaware that a suitable threshold of safety has been developed in answer to this question.

It may also help to understand that conventional stepwise regression techniques were developed and are successfully applied in circumstances where the aliasing among variables results from uncontrolled methods of sampling and not from naturally-occurring relationships among the variables. In such a case, the physical phenomenon being studied would permit the variables to be sampled independently in a controlled experiment, and it is reasonable to claim that each variable exerts an independent effect on the response variable. The correlations present in the data derive from the methods of sampling and could change, or even disappear, in a new data set. In this environment, the correlations between variables contain no information pertaining to the physical phenomenon, and the presence of correlations is of concern primarily to the extent that the standard errors of estimate are inflated (i.e., the data set is less efficient for estimation).

In the circumstance of fuel effects on emissions, natural correlations exist among diesel fuel properties as an expression of the characteristics of blendstocks and the effects of refining processes. The natural correlations

[6] Not to be confused with Condition Index, which is the square root of Condition Number.

will tend to recur across repeated samples of fuels manufactured using similar blendstocks and processes, although each data set may also have unique or "volitional" correlations that result from the specific sampling scheme. It is the recurring, natural structure underlying diesel fuels that the PCR+ approach attempts to identify and harness. In this environment, where fuel properties do not vary independently, it is more reasonable to believe that the eigenvector variables exert independent effects on the response variable and that attempts to attribute these effects to the individual fuel properties will necessarily be frustrated.

4. APPLICATION OF PCR+ TO DIESEL EMISSIONS

4.1 STATISTICAL BACKGROUND

Although the discussion of stepwise regression has emphasized the difficulties caused by the inter-relatedness of predictors, the motivation for PCR+ is much more than an effort to eliminate collinearity between variables. It is a distinct approach to fuels and emissions research in its own right, aimed at a more natural interpretation of the relevant factors that influence emissions, in addition to exploiting orthogonalization to eliminate the aliasing or confounding of variables. The objective of PCR+ is to assess emissions in terms of the joint forces that work together in determining emissions levels for a fuel. It is *not* a round-about-way to develop a regression equation for individual fuel properties.

Although PCR has been in existence for a long time, there remain opportunities for development of the technique as indicated by the following quotation from J.E. Jackson's (1991) *A User's Guide to Principal Components*:

> *If, in your experience with PCA, you discover something that surprises you, do not assume that it is a property of PCA known to everyone but you. If no one else has written about it, sit down with pencil and paper ... and share it with the rest of us. ...[T]here are still plenty of development opportunities to make PCA an even more useful technique.*

Our use differs materially from the manner in which PCR has been used historically in the literature. To distinguish our approach, we label it PCR+, the "plus" denoting the innovations we have introduced.

The fact that every data set has an orthogonal basis is at the root of the PCR+ methodology. We use the eigenvectors of PCA, rather than the individual fuel properties, to describe fuels and as the independent variables in regression analysis against engine emissions. Initially, we incorporate all of the vectors in the regression model and then rely on tests of statistical significance and substantiality (meaningfulness of contributions) to select a smaller number of terms to be retained. This method avoids certain problems related to the practice of arbitrarily dropping eigenvectors with small eigenvalues and is remarkably consonant with the recommendations in Jackson's (1991) *User's Guide* as follows:

> *The answer ... would appear to be to obtain all of the pc's and regress the response on all of them. Since they are independent of each other, the amounts of the response variability accounted for by each of the pc's are also independent and hence the pc's could be ranked in order by that criterion and a cutoff employed when the desired residual has been obtained.*

In effect, we treat the "eigenized" X-matrix like any other matrix of predictor variables, perform the same tests of significance, and claim all the properties, such as unbiasedness, of the method of estimation that is used. In reality, the columns of numbers in a design matrix of principal components are just that: columns of numbers. If they were labeled X_1, X_2, ... X_n and given as a textbook exercise for the student, no one would ever suspect their shady past. Given a response vector Y, the student would faithfully do what is expected: derive a multiple regression equation, perform some tests of significance and publish the results. Moreover, it is not subject to bias (as stepwise regression is) when terms are removed from the model, because the alias matrix is null. Therefore, coefficients are invariant when terms are added to or removed from the model.

The method of estimation is OLS in our case, but there is nothing to preclude the use of eigenvectors in a Mixed Effects model. The "eigenized" X-matrix is the result of transforming to a new set of variables, unique to the data set at hand, that is completely orthogonal. From orthogonality flow a number of desirable statistical properties, including:

- Elimination of all aliasing among predictor variables
- Elimination of variance inflation resulting from the inter-relatedness of the predictors
- Independence and, therefore, additivity of the individual vector contributions to the response sums of squares and R^2
- Creation of a variable space in which the associations of individual terms to the response variable can be identified without ambiguity.

We will now look at how the PCR+ approach is applied to the problem of diesel fuel effects on HDD engine emissions. A tutorial presentation of the approach is given in McAdams *et al*., 2000b.

4.2 EIGENVECTOR REPRESENTATION OF FUELS

As defined by PCA, each eigenvector is a linear combination of fuel properties, consisting of coefficients or weights for each property variable, that expresses a unique feature found empirically in the fuels. The eigenvectors are defined in such as way as to partition the variation among fuels into its independent components. The first vector represents the feature by which fuels differ most, while the last vector represents the feature that varies least.

In this section, we will work with eigenvectors defined using the EPA list of twelve variables, which consists of:

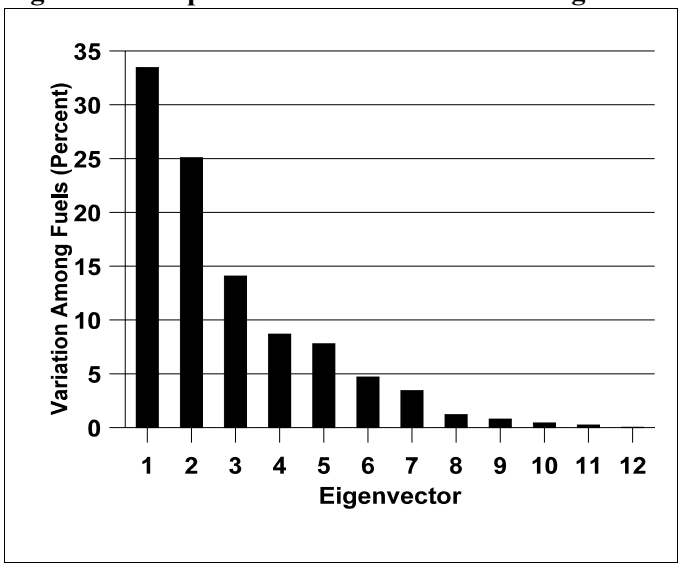
- Seven native fuel properties, meaning physical and chemical properties of fuels that are clear of additives
- Two property variables representing the use of cetane additives and oxygenates in the fuels
- Three quadratic variables to represent possible nonlinear effects for natural cetane, cetane difference, and total aromatics.

Cetane difference, defined as the increase in total cetane number over the natural cetane number, is used to measure the effect of cetane additives in the fuels, while the percent of oxygen by weight is used to measure the use of oxygenates.

When PCA is applied to the experimental fuels in the EPA database using this variable list, we find that only six features are needed to explain nearly 95 percent of the differences among fuels.[7] Figure 4.1 shows how the first (major) eigenvectors are most important for explaining the differences among fuels. The first vector alone accounts for one-third of the differences among fuels, and the second vector accounts for one-quarter. The vector contributions decline rapidly after that to reach less than 5 percent for the sixth vector.

[7] Appendix Table A.1 summarizes the PCA analysis of the EPA Experimental Fuels.

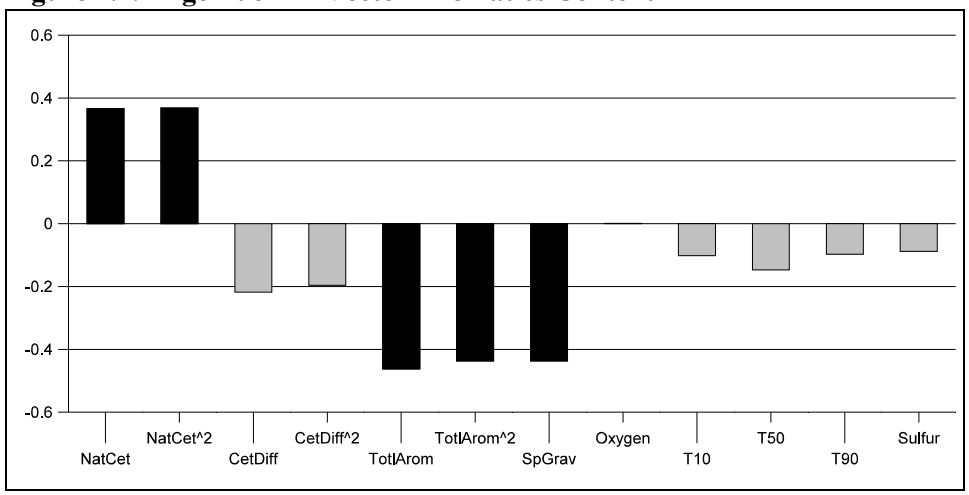
Figure 4.1. Explanation of Differences Among Fuels



We term these vectors "eigenfuels" because, in both experimental and commercial fuels data sets, the eigenvectors appear to serve as building blocks of the fuels and can be given interpretations in terms of refining and blending processes. Although eigenfuels may be unfamiliar at first, they soon become as natural a way to describe fuels as the individual fuel properties are now.

Each eigenvector is composed of a weighted combination of the original fuel property variables that expresses a particular relationship among the properties existing in the data. Figure 4.2 shows, for example, that Vector 1 of the EPA data set is an expression of the fuel's total aromatics content in conjunction with natural cetane and specific gravity. The vertical axis gives the weight for each of the twelve fuel property variables that make up the vector. The dark bars show the terms that make the largest contributions to the vector. We can "read" the vector as saying that a decrease in total aromatics content (both linear and quadratic terms) is associated with an increase in natural cetane (linear and quadratic) and a decrease in specific gravity, with smaller effects on other properties. The property changes are ones that occur simultaneously whenever the amount of this eigenfuel varies. This is the feature that varies most in the experimental fuels found in the EPA data set. A refinery-based interpretation would call this the "light cycle oil" vector.

Figure 4.2. Eigenfuel 1 – Vector Aromatics Content



The number of vectors defined for a data set is determined by the number of fuel property variables used to describe the data. More resolution of underlying fuel features will be afforded when a longer list of variables is considered, while less resolution is permitted with a shorter list of variables. Using the EPA slate of variables, the fuel features summarized in Table 4.1 are found. The first five vectors, representing vector variations in fuel characteristics related to aromatics content, natural cetane, additized cetane, oxygen content, and sulfur content, are ones logically found in a data set developed to test for fuel effects on emissions. The sixth feature, the slope of the distillation curve, may be related to controlling flash and pour points to commercial specifications. Eigenfuels 10, 11, and 12 represent the nonlinear effects for additized cetane, total aromatics, and natural cetane.

Table 4.1. Features of Experimental Fuels in EPA Database

Eigenvector	Description
1	Aromatics variation in association with natural cetane and specific gravity
2	Natural cetane variation in association with other properties, but largely independent of aromatics content
3	Additized cetane (and associated properties)
4	Oxygen content (and associated properties)
5	Sulfur content (and associated properties)
6	Slope of distillation curve
7 - 9	Minor vectors not given interpretations
10 - 12	Nonlinear terms for additized cetane, total aromatics, and natural cetane, respectively.

Eigenfuels offer a concise and insightful method for describing how fuels have been formed that we find to be more natural than the individual fuel properties. A real fuel is described mathematically as a weighted combination of the eigenvectors, in which the weights vary from one fuel to the next, just as a blend can be described in terms of the relative amounts of the various blendstocks that make up the blend – hence, the term "eigenfuel."

4.3 A PCR+ MODEL OF EMISSIONS

Having expressed real fuels in terms of eigenfuels, the coefficients of that expression can then be used as predictors for emissions in regression analysis, just like any other variable. The PCR+ approach is to conduct a two-stage analysis.[8] In the first stage, the emissions effects associated with individual engines are removed from the response data. A model of the form:

$$\log(\text{emissions}) = A_1 + A_2*\text{Eng}_2 + A_3*\text{Eng}_3 + \ldots + A_n*\text{Eng}_n \qquad (1)$$

[8] See SWRI, 2001, pp. 35-39 for further description.

is estimated for each pollutant studied, where the variables Eng_n are dummy variables for N-1 engines and the coefficients A_i represent mean (log) emissions levels for the engines i = 1 ... N related to characteristics of the individual engines and a variety of effects associated with engines, including laboratory and test cycle.

The residuals of Equation (1) are the deviations of actual emissions from the engine-specific mean (log) emissions levels. The residuals include the emissions effects of the variation in fuel characteristics, sources of systematic variation not associated with fuels or engines (if any), and test-to-test variability. The residuals become the response variable for a second stage regression in which the predictor variables are the coefficients in the eigenvector representation of tests. The model is of the form:

$$\log(\text{residuals}) = A + B_1*EF_1 + B_2* EF_2 + ... + B_n*EF_n \qquad (2)$$

where the variables EF_i are the weights of vector i in the eigenfuel expansion of the fuels, the intercept term A is identically zero because engine effects have been removed, and the coefficients B_i are the regression coefficients for emissions.

The PCR+ approach is to enter all vector coefficients into the regression and then exclude vectors from the model based on the evidence for their statistical significance and substantiality. The substantiality criterion recognizes that an effect could be found statistically significant in a large data set even though it contributed very little to explaining emissions. The threshold at which one would exclude a statistically significant term based on substantiality will depend on the nature and use the resulting model. We suggest a threshold on the order of 1 percent of the response sums of squares or, equivalently, 0.01 contribution to the R^2 statistic.

The PCR+ approach was applied to the EPA database to develop eigenvector-based emissions models for NO_x and PM.[9] Figures 4.3(a) and 4.3(b) show emissions models in terms of each eigenvector's contribution to the total effect of fuels on emissions, plotted on the vertical axis as a percent of the fuels-related model SS. We see that eigenvector 1 – the aromatics, natural cetane, and specific gravity vector – has the single largest effect on both $\log(NO_x)$ and log(PM) emissions, accounting for more than 60 percent of the fuels-related variation in NO_x and more than 50 percent of the fuels-related variation in PM. From the perspective of PCR+, the effects are related to the *joint variation* in aromatics, natural cetane, and specific gravity and *cannot* be ascribed to any one fuel property in isolation from the other. When we hear discussion of the emissions effects of the "Big Three" diesel fuel variables – natural cetane, aromatics, and specific gravity – we conclude that the discussion is about this one vector and its individual emissions effect.

There are other effects on emissions. For NO_x, the second largest effect is due to Vector 3 (additized cetane), which accounts for more than 20 percent of the fuels-related variation in NO_x. Vector 2 (natural cetane variation independently of aromatics), Vector 5 (oxygen content), and Vector 11 (nonlinear aromatics content) make smaller contributions. Eight vectors (numbers 1, 2, 3, 5, 6, 9, 10, and 11) are statistically significant at the 0.05 level and five of the vectors (numbers 1, 2, 3, 5, 9, and 11) are substantial, in that they contribute at least 1 percent to the model SS. It should be remembered that the SS contribution depends on both the strength (magnitude) of the vector effect and on how much the vector was varied in the test set. A small share here does not necessarily mean that an eigenfuel is unimportant in all cases, since it might vary much more in another data set.

[9] Appendix Tables A.2 and A.3 summarize the PCR+ regression analysis for NO_x and PM, respectively.

Figure 4.3(a). Eigenfuel Impact on NO_x Emissions

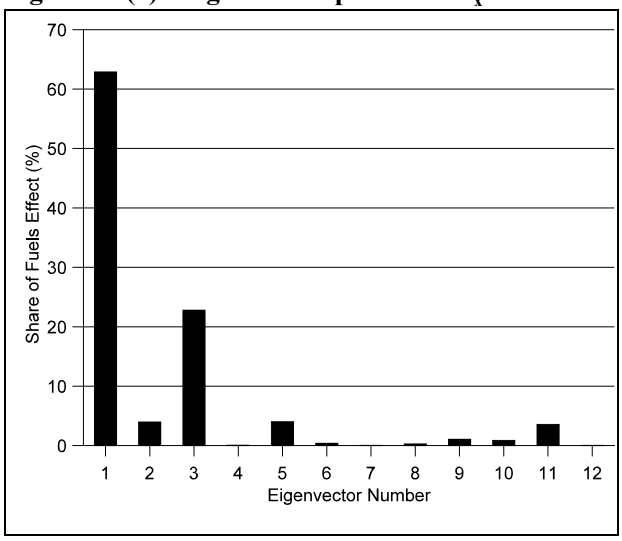
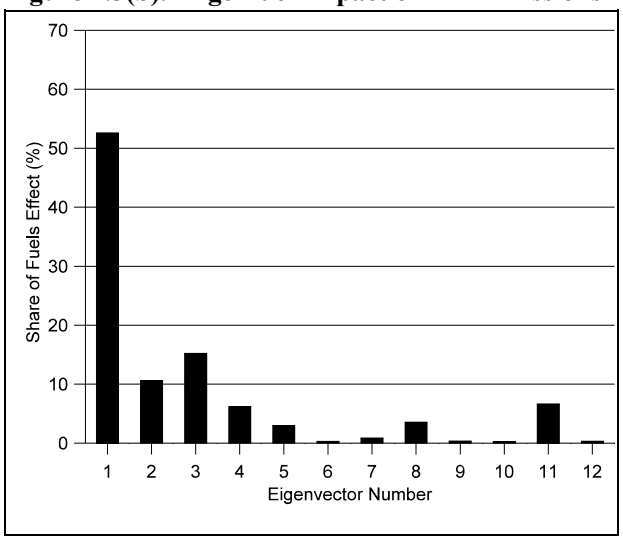


Figure 4.3(b). Eigenfuel Impact on PM Emissions



For PM, Vector 3 (additized cetane) is also the second largest effect, although several eigenfuels make comparably sized contributions including:

- Vector 2 (natural cetane variation independent of aromatics)
- Vector 4 (oxygen content)
- Vector 5 (sulfur content)
- Vector 8 (un-interpreted)
- Vector 11 (quadratic aromatics).

Eight vectors (numbers 1, 2, 3, 4, 5, 7, 8, and 11) are statistically significant at the 0.05 level, and seven vectors (numbers 1, 2, 3, 4, 5, 8, and 11) are substantial at the 1 percent level.

Turning to the interpretation of the eigenvectors given in Table 4.1, we see that the "light cycle oil" Vector 1, which accounts for one-third of the variation among fuels, is the primary fuel factor associated with NO_x and PM emissions. Vectors 2 through 5 have significant and substantial effects on either NO_x or PM emissions, while nonlinear effects on NO_x and PM are found for aromatics content only. Of the six major fuel characteristics, only Vector 6 (distillation curve slope) is unrelated to emissions, while uninterpreted Vectors 8 and 9 make small contributions. As suggested here an discussed in greater detail in Section 5, the PCR+ approach to emissions analysis produces a concise, unambiguous, and easily understood basis for assessing the characteristics of fuels that affect emissions.

4.4 DIESEL FUEL EFFECTS ON EMISSIONS

The models presented above have been described in terms of the fuel characteristics found to influence emissions, but without information on the direction or magnitude of the effects. Let us now look more closely at the diesel fuel effect by examining the predicted emissions impact of hypothetical changes in fuel characteristics. We will define the hypothetical changes based on an analysis of features that are characteristic of commercial diesel fuels. In a sense, this is the acid test for an emissions model, because the objective in its development is the prediction of emissions reductions achieved by modifying commercial fuels. We will also estimate the emissions reduction achieved by a typical diesel fuel sold in the Los Angeles

area under California's diesel fuel formulation requirements, in comparison to emissions from an average U.S. diesel fuel.

The predictions of the PCR+ emissions models are compared with predictions of EPA's Unified Model. We cannot say from the comparisons that one model is right and the other is wrong, but we should be concerned if we find them to give substantially different predictions, inasmuch as they were based on nearly identical data and differ primarily in terms of the regression methodology. Although comparisons will be shown for both NO_x and PM, we will give weight only to the NO_x comparisons, where the eigenfuel and EPA models considered the same twelve fuel property variables and the only appreciable difference is the model-building methodology. The differences in prediction are actually greater for PM, but the EPA model considers an interactive term for natural cetane and cetane difference that was not considered by the eigenfuel model.

It should be noted that the PCR+ emissions models were originally developed as a demonstration of the PCR+ technique and for comparison to the Unified Model developed by EPA and, as a result, the PCR+ models adopt the fuel variables considered by EPA. The PCR+ models do not represent a final determination by DOE or ORNL on the identity of the fuel characteristics affecting diesel emissions or on the magnitude of that effect. Emission models also are subject to change with the addition of new emissions test data, innovative fuels formulations, and updated information on refining technologies.

4.4.1 Hypothetical Changes in Diesel Fuel Characteristics

In Section 5.1 of this report, we examine the eigenfuel characteristics of commercial diesel fuels using a proprietary database. As shown there, commercial diesel fuels have an eigenvector structure based on five primary features that can be related to common diesel blendstocks, including light cycle oil, hydroprocessed heavy distillate, and straight-run heavy and light distillates. The first three eigenvectors express features in which aromatics content varies in association with other properties, the fourth eigenvector is closely related to sulfur content, and the fifth to the IBP. The first three features involve blendstock variations that have distinctive effects on fuel aromatics content and other properties, including natural cetane and specific gravity. In comparison, there is only a single feature related to aromatics variation in the experimental fuels, which corresponds most closely with Vector 1 in the commercial data.

The method for comparing emissions predictions is based on varying the eigenfuel content of commercial diesel fuel in hypothetical ways. Starting with the characteristics of the average commercial fuel, we change the content of a selected commercial eigenvector in steps, moving in the direction of reduced emissions. At each step, the modified fuel is resolved back into its fuel property values and into its expression based on the experimental-fuel eigenvectors of the EPA data set. The PCR+ and the EPA Unified Model are then evaluated to predict emissions for the hypothetical fuel. Thus, the emissions response is mapped out for a vector-based modification to the average commercial fuel. We will consider five different vector changes: three associated with the aromatics-related vectors found in commercial fuels, one for the use of cetane improvers, and one for the use of oxygenates.

Figures 4.4 through 4.6 show the predicted impact on NO_x and PM emissions of changes in the aromatics-related eigenvectors of commercial fuels. The vertical axes give the emissions impact measured relative to a value of zero for the average commercial fuel. The horizontal axes give the fuel's content of the eigenvector, where the content is zero for the average commercial fuel and is measured in units of the standard deviations of the eigenvector content found in commercial fuels. The emissions predictions start at zero, by definition, with the average commercial fuel (with 33 percent aromatics content), and continue until a sufficient vector change has been made to reduce total aromatics content to the level of 10 percent. The fuel modifications shown here involve changes in *all* of the fuel properties associated with each vector,

Figure 4.4. Predicted Emissions Effects of Commercial Eigenvector 1

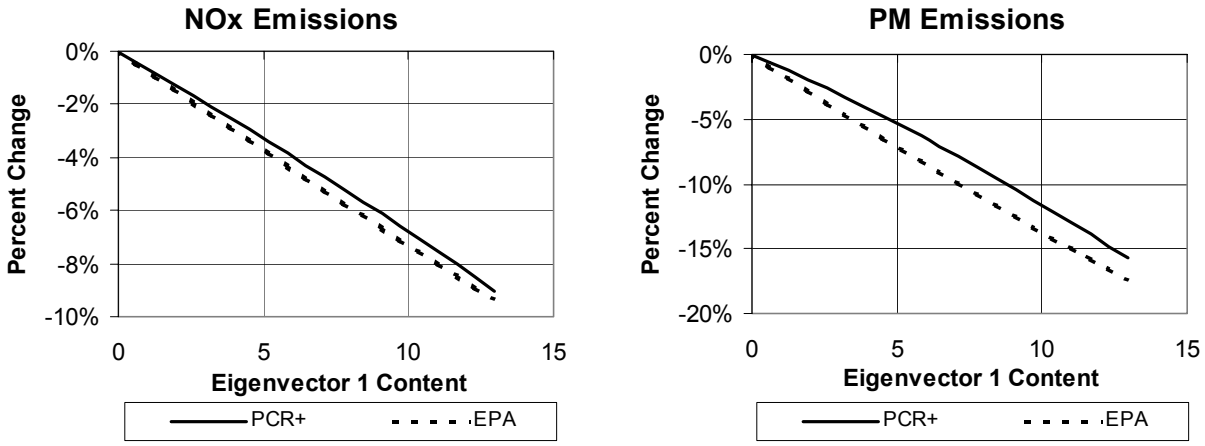


Figure 4.5. Predicted Emissions Effects of Commercial Eigenvector 2

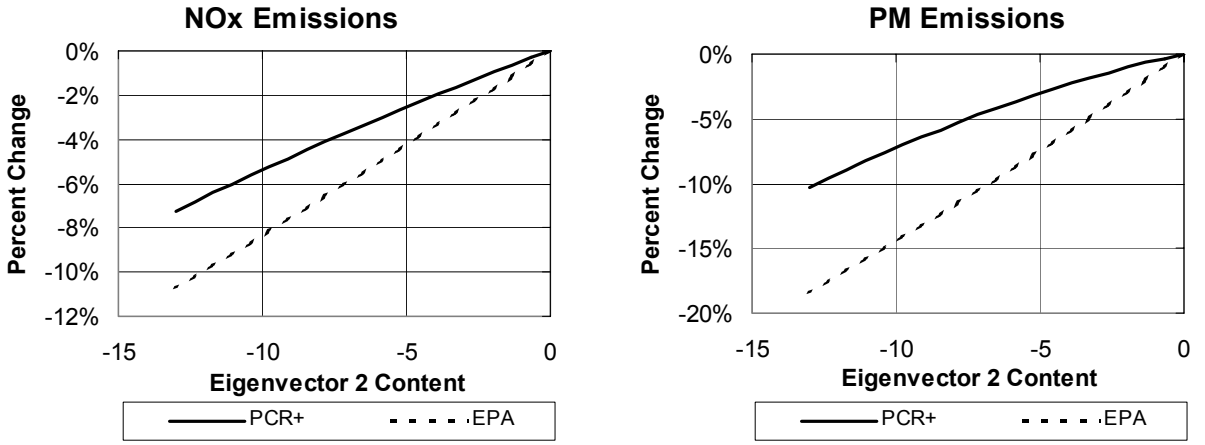
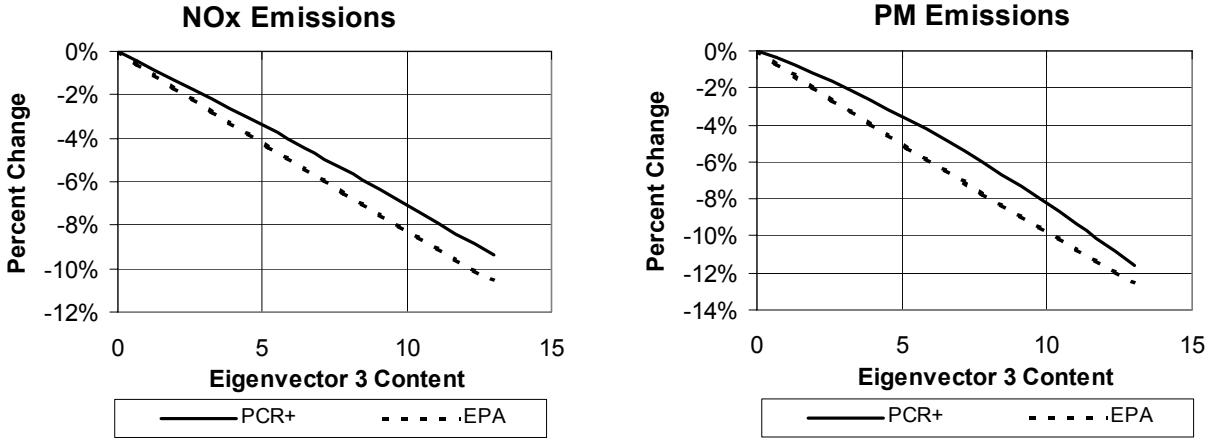


Figure 4.6. Predicted Emissions Effects of Commercial Eigenvector 3



not just aromatics content, and we will label this approach in which all properties participate as a "vector aromatics" change.

Figure 4.4 shows the emissions effect of varying Commercial Vector 1 (Light Cycle Oil) across the range needed to reduce the vector aromatics content to 10 percent, from the commercial average of 33 percent. The PCR+ model predicts that this vector change will reduce NO_x emissions in a nearly linear fashion by a total of 9.1 percent. This is closely matched by the EPA Unified Model, which predicts a 9.5 percent NO_x reduction for the same total change in fuel properties. The two models predict greater, and similar, effects on PM, reaching the range of a 15 to 17 percent reduction when 10 percent vector aromatics is reached. While the predictions exhibit slight curvature in the graphs, they are straight lines in the space of log(emissions).

Commercial Vector 2 (Hydroprocessed Heavy Distillate Feature) involves three times the increase in natural cetane for each unit of aromatics reduction compared to Commercial Vector 1. As shown in Figure 4.5, the PCR+ model predicts a smaller change in NO_x emissions when Vector 2 is reduced, amounting to an 8.5 percent NO_x reduction when the endpoint of 10 percent vector aromatics is reached, compared to the 12.5 percent reduction that is predicted by the EPA model. The predicted PM emissions impacts are much larger, ranging from a 10 percent to nearly a 19 percent reduction, and there is substantially more difference between the models.

Commercial Vector 3 (Straight-Run Heavy Distillate) involves an intermediate effect on natural cetane compared to Commercial Vectors 1 and 2. As shown in Figure 4.6 the models predict similar reductions in NO_x emissions: a 9.5 percent reduction for the PCR+ model when the 10 percent vector aromatics endpoint is reached, compared with a 10.7 percent reduction for the EPA model. The emissions reductions predicted for PM are also similar, reaching approximately 12 percent reduction at the endpoint of the range.

Figures 4.7 and 4.8 show the predicted effects of cetane improvers and oxygenates, respectively, on NO_x and PM emissions. Here, the horizontal axes are specified in units of cetane number increase and the percent oxygen content by weight. These hypothetical fuel changes have been developed from the effect of additives used in the experimental fuels and do not depend on the commercial fuels, which were clear of the additives.

Figure 4.7 shows that the PCR+ model predicts a larger effect of cetane improvers on NO_x than does the EPA model. At a maximum of +10 cetane numbers, the PCR+ model predicts a NO_x reduction of 3.5 percent, while the EPA model predicts a 2.7 percent reduction. PM predictions differ enormously, but the predictions may not be comparable because the PCR+ model does not include an interactive term (natural cetane x cetane difference) that EPA found to be significant for PM.

As shown in Figure 4.8, the PCR+ model predicts a small, but statistically significant increase in NO_x emissions as the result of oxygenates in fuels, amounting to a 2.0 percent increase in NO_x at 4 percent oxygen content. The EPA Unified Model excludes oxygen content as a predictor variable for NO_x and, therefore, predicts zero effect. The models predict a substantial effect on PM emissions, ranging from a 12 to a 25 percent reduction at 4 percent oxygen contents.

Table 4.2 summarizes the models' NO_x predictions. The results from the PCR+ and EPA models agree closely on the NO_x impact of reducing aromatics in accord with Commercial Vector 1, but disagree significantly regarding the impact of Vector 2 and, to a lesser extent, Vector 3. For Vector 2, the EPA estimate of the NO_x effect is nearly half again as large. Overall, the PCR+ estimates a 9 percent NO_x reduction in going from 33 to 10 percent vector aromatics, whether this is accomplished via Vectors 1, 2, or 3. The EPA Unified Model predicts different impacts depending on which vector is varied and, consequently, on how the other fuel properties are affected. The EPA model would appear to estimate a somewhat

Figure 4.7. Predicted Emissions Effects of Cetane Improvers

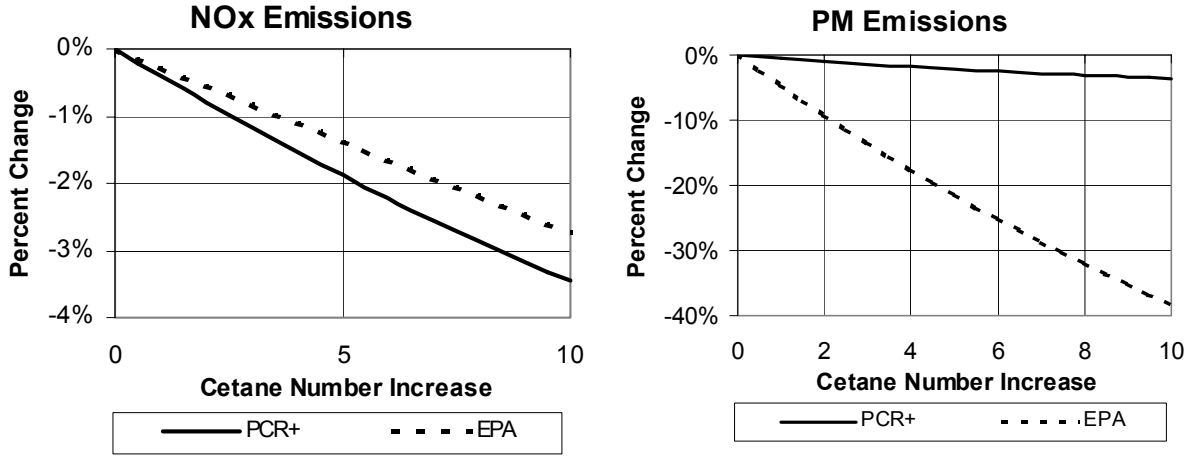
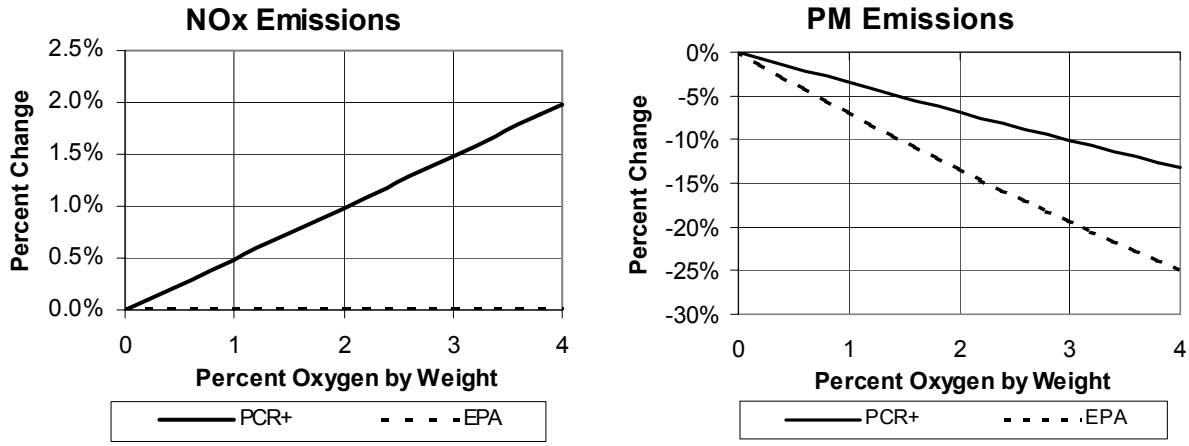


Figure 4.8. Predicted Emissions Effects of Oxygenates



greater overall effect of aromatics reduction. The PCR+ and EPA models disagree on the magnitude of the additized cetane effect and disagree on whether oxygen content adversely affects NO_x.

The foregoing comparisons between model predictions are not intended to argue that one model is correct and the other wrong, and cannot do so in fact. The purpose of the comparisons is to make clear that the conclusions one draws on which variables affect emissions, and by how much, depend to a significant degree on the choice of analysis methodology – PCR+ or stepwise regression – in the current circumstance where the data employed is otherwise the same.

Table 4.2. Summary of NO_x Predictions (Changes Relative to U.S. Average Diesel Fuel)

	PCR+ Model	EPA Unified Model
Blendstock Composition Change (Aromatics 33 ➜10 percent)		
Light Cycle Oil Vector	-9.1%	-9.5%
Hydroprocessed Heavy Distillate Vector	-8.5%	-12.5%
Straight-Run Heavy Distillate Vector	-9.5%	-10.7%
Additized Cetane + 10 numbers	-3.5%	-2.7%
Oxygen Content + 4 percent	0.02	none

4.4.2 Predicted Emission Changes for Los Angeles Fuel

Notwithstanding different views regarding the methods of analysis, it is possible that some consensus has been reached on the overall magnitude of the diesel fuel emissions effect. At EPA's request, estimates were made of the emissions benefit associated with moving from an average U.S. diesel fuel to California diesel specifications. The information was used by EPA to look for such a consensus, even if there is not a consensus regarding the modeling approach or the predictions for individual fuel properties.

Two fuels are considered. The first is a baseline fuel identified by EPA as representative of the U.S. average diesel fuel.[10] The second is a California average diesel fuel[11] as determined by the Alliance of Automobile Manufacturers (AAM) survey in Los Angeles. The California fuel represents the result of diesel fuel reformulation under the Air Resources Board program currently in place. Emission reduction estimates are made for the change from the U.S. average fuel to the California average fuel. As shown in Table 4.3, the PCR+ model predicts somewhat smaller emissions effects: a 5.0 percent reduction for NO_x versus the 6.2 percent reduction predicted by the Unified Model, and a 6.9% reduction for PM versus the 8.5% reduction predicted by the Unified Model.

The PCR+ model presented here will give emissions estimates that differ from those of the eigenfuel model originally published (McAdams *et al*. 2001b). The original model was based on a small database of 280 tests on 11 different engines and was developed for the primary purpose of demonstrating methodology. Its database was less representative of the overall diesel engine fleet than the greatly enlarged database compiled by EPA, and predictions of this early model are less comparable to the EPA Unified Model. Further, the PCR+ model presented here is not a final determination by DOE or ORNL on the identity of the fuel characteristics affecting diesel emissions or on the magnitude of those effects.

[10] See U.S. EPA 2001, Table III.D-1 Baseline Fuel Properties.
[11] See U.S. EPA 2001, Table III.F-1 Average California Fuel Properties.

Table 4.3. Summary of Predicted Emission Changes for California Diesel Fuel (Changes Relative to U.S. Average Diesel Fuel)

	NO_x	PM
EPA Unified Model		
Average California Fuel	-6.2%	-8.5%
PCR+ Model		
Average California Fuel	-5.0%	-6.9%

4.5 PARTITIONING OF THE MODEL SUMS OF SQUARES

An important objective in most studies is the identification of the variables that exert an effect on the response of interest and the elimination of other variables judged to have no effect. In stepwise regression, the variable selection procedure is based on a conditional t-test of statistical significance. This test is based on the regression coefficient and its standard error, in circumstances where the standard error is increased by the covariant relationship of the term to the other variables present in the model. An incremental Sum of Squares (SS) can be formulated for the variable by determining the reduction in model SS when the variable is dropped from the model. The t-test and the SS must be formulated on a marginal basis because, in the presence of aliasing among variables in a nonorthogonal set, there can be no unique way to partition the SS.

The circumstance is quite different in the orthogonal environment defined by PCA. Each eigenvector is independent of the others and the model SS can be uniquely partitioned among them when used as explanatory variables in a regression model. Once the unique SS partitioning is recognized by eigenvector, it becomes possible to partition the SS among the underlying fuel variables by virtue of the fact that each eigenvector can be expressed as a linear combination of the original variables. A method for the SS partitioning by fuel property was developed in prior work,[12] which proposed a variable selection method that could assist in reducing the dimensionality of the problem. Some controversy attended this method, so that additional consideration has been given to the issue of SS partitioning and the methods proposed for the purpose.

As developed in Appendix D, the model SS for a general linear model can be partitioned into terms formed from the variances and covariances of the predictor variables. If the variables are orthogonal, the covariance terms are identically zero and, in this case only, the SS can be partitioned into terms uniquely associated with the predictor variables. This is the circumstance encountered with eigenvector variables. More commonly, when the variables are nonorthogonal, no partitioning of effects is possible without taking into account the covariances among the predictor variables. Any attempt at partitioning the SS among the predictor variables must then propose a method for attributing the covariance terms to the individual variables involved. Many such partitionings are possible and none can be held out uniquely as the "true" partitioning of the model SS.

The absence of a unique solution in the nonorthogonal case does not preclude the possibility that SS partitioning may offer insight to the problem of variable selection. Appendix D advances two candidate approaches to variable selection. The first is the method proposed in prior work for distributing the SS associated with the eigenvectors to the individual fuel property variables of which they are composed. The second is based on what we choose to call an "even-handed" attribution of the response SS associated with the covariance of two predictors. The even-handed principle divides that SS equally between the two

[12] See McAdams *et al.*, 2001b, Section 2.4 and Appendix D, Addendum D-1: A Program for Computing SS Contributions.

covariant predictors involved. Computer algorithms Simplify.m and SortSS.m, which implement the eigenvector-based and even-handed methods, respectively, may be found in the appendix, as well as numerical demonstrations of the procedures.

The eigenvector-based partitioning done by Simplify.m takes into consideration:

- The regression coefficients for the eigenvectors, giving their relationship to the response variable of interest
- The eigenvalues, giving the variance associated with the eigenvectors in the data set
- The eigenvector components, giving the weight of each fuel property variable in forming the vectors.

While other partitionings of the SS are possible, the partitioning among fuel property variables by this method is implemented by the computational expansion of the SS associated with each eigenvector and involves no element of judgement in attributing the SS to the fuel property variables. This is true regardless of what subset of eigenvectors is retained in the model. Further, when the SS partitioning is performed using all eigenvectors, it is found that the attributed SS exactly total the model SS as computed by regression either on all fuel property variables or on all eigenvectors.

The even-handed method provides a complementary approach to the partitioning of effects in a regression model. It incorporates covariant effects directly and yields a breakdown capable of providing insight into aliasing effects that influence variable selection. The partitioning done in SortSS.m takes into consideration:

- The regression coefficients for the fuel property variables
- The variances of the fuel property variables in the data set
- The covariance terms between each pair of variables, which are split equally between them for purposes of attributing the SS to the individual variables.

The method does not depend on the eigenvector representation of the data in any manner. In fact, it is the result of a general treatment of the model SS within the framework of error propagation in linear models. It considers predictor variables, whether orthogonal or not, and would reproduce the partitioning of SS by eigenvectors when applied to data expressed in those terms. When applied to nonorthogonal data having non-zero covariance terms, it distributes the terms in the one even-handed manner that is possible. As seen in the appendix, this method admits the possibility that SS terms can be negative as a result of the negative covariance that can occur between two predictor variables.

We recommend the Simplify.m algorithm as the primary discriminant for variable selection because it is derived from the SS partitioning for eigenvectors. The algorithm SortSS.m is offered as a complementary approach that may provide further insight. Researchers working within the conventional paradigm based on fuel properties may see variable selection procedures as a way to determine which fuel properties should be retained in a final emissions model. We see this from a different perspective, however, since we believe that the emissions effects are better represented by the eigenvectors (i.e., *combinations* of fuel properties).

From our viewpoint, variable selection procedures are most useful for determining the number of fuel property variables needed to characterize the eigenvectors that are found to be relevant to emissions. We would first prune a PCR+ model of eigenvector terms that fail tests of significance or substantiality, as required. Then, we would strike from the variable space the fuel property variables (if any) that are shown to be superfluous based on the model SS attributed to them by the Simply.m algorithm. If the dimensionality of the variable space is reduced by this process, the eigenvectors and the PCR+ emissions model would be

re-estimated. The expected outcome of this process would be that the dimensionality of the variable space N will typically be greater than the number of eigenvectors M that are retained in the final PCR+ model.

Table 4.4 shows the SS partitioning based on applying the recommended Simplify.m algorithm to the PCR+ models described above, in which all eigenvector terms that are statistically significant at $p>0.05$ are retained. We see that natural cetane and its squared term are associated with more than 20 percent of the model SS for NO_x and would be identified as influential variables, along with cetane difference (and square), total aromatics (and square), and specific gravity, by any criteria. Oxygen content, sulfur content, and the distillation curve temperatures are seen as relatively less influential. Except for T10, the last five terms might be candidates for exclusion. For PM, we see that all of the fuel property variables except one (cetane difference^2) contribute at least 1 percent to the model SS.

Table 4.4. Sum of Squares Partitioning for PCR+ Emission Models (Simplify.m Algorithm)

	SS for NO_x (percent)		SS for PM (percent)	
Natural Cetane	9.7	*	7.4	*
Natural Cetane^2	11.1	*	9.1	*
Cetane Difference	6.8	*	2.9	*
Cetane Difference^2	2.6	*	0.4	
Total Aromatics	31.8	*	32.0	*
Total Aromatics^2	9.7	*	4.4	*
Specific Gravity	21.9	*	6.2	*
Oxygen	0.01		8.2	*
T10	4.6	*	9.2	*
T50	0.5		2.1	*
T90	1.2	*	5.0	*
Sulfur Content	0.07		13.1	*

* Signifies variable that contributes more than 1 percent to the model SS and would be retained in the variable space.

From these results we would choose to retain *all* of the fuel properties in the variable space in this instance, because there are clear merits to maintaining a common variable space for both pollutants and the computational cost of additional variables is small in today's world. In more general circumstances, it would be reasonable to have one set of predictors for one purpose and a different set for another purpose.

Whether Simplify.m is used, with the results shown above, or the complementary procedure SortSS.m, we note that both methods lead to a different result than the EPA Unified Model in terms of the fuel properties found to be relevant to NO_x: *natural cetane should not be discarded as a predictive variable*. The clear shortcoming of the Unified Model is its implicit decision to reject natural cetane *because its contributions could be accounted for by aromatics content and specific gravity and not because natural cetane was without effect.* Variable selection should be informed by methods other than the conditional test of significance possible in stepwise regression.

4.6 ISSUES OF STATISTICAL BIAS

EPA perceives PCR+ as a methodology that produces biased estimates for the fuel property coefficients. It is certainly true that, if one or more eigenvectors are removed from the model and the model then transformed to its equivalent in terms of fuel properties, the resulting fuel property coefficients will not be the same as those resulting from a least-squares fit to all of the original variables. This difference also will tend to increase as more eigenvector terms are removed. We take issue, however, with the EPA assessment that PCR+ is inherently biased.

The concern with bias derives from the development of PCR in the statistical literature as an alternative method for estimating regression coefficients in circumstances where near-exact collinearities exist among variables to the extent that the matrix OLS solution cannot be computed. In this circumstance, the predictor variables can be converted to eigenvector space, a regression analysis conducted, and then the regression coefficients converted back to their equivalents in the original space. Although the resulting coefficients are biased with respect to the population values *in the original variable space*, the presence of an often-small bias was seen by the proponents of PCR as preferable to being unable to conduct the analysis.

Though PCR+ is biased when used in the foregoing context, it is incorrect to classify PCR+ as an inherently biased methodology. Bias is a property of the estimation procedure; OLS will give unbiased estimates of the coefficients associated with the X-space variables, regardless of how the X-space is defined. When the X-space consists of the eigenfuel description of fuels, the coefficients are unbiased estimates of the response associated with the eigenvectors. In addition, the coefficients have the desirable property of being unaffected by aliasing, so that the coefficient estimates are invariant with respect to changes in the number of terms retained. Further, the eigenvector model retains its stature as an unbiased estimate of the coefficients for the vectors involved, even when transformed into fuel property terms. Transformation to its equivalent in fuel property terms is simply a method to display the results in another form and is not an alternate method for estimating the coefficients pertaining to fuel properties.

The concern with bias originates from the view that the effect of fuels on emissions is best measured in terms of the individual fuel properties. We dispute this view because the properties of real diesel fuels are subject to naturally occurring correlations that stem from the blendstocks and refining processes used to produce them. PCR+ was not intended as an "end round" means for estimating coefficients for the fuel property terms, but as a direct means for estimating coefficients for the eigenvectors terms that, we believe, are the preferred choice of variable. Our reasons include, on the one hand, the conclusion that the vectors are a more natural and insightful means of describing the composition of fuels and, on the other, a recognition of the desirable statistical properties that flow from their use, including orthogonality and parsimony (reduction in the dimensionality of the problem).

A simple example may be more powerful than lengthy argument in clarifying this view. A recurring observation in the EPA analysis and the technical literature is the association of cetane number with aromatics and specific gravity and the attendant practical difficulty, if not inability, to separate the effects of these variables on emissions. Our eigenvector analysis finds the association among cetane number, aromatics, and specific gravity to be the most important vector feature describing differences among fuels. Why not think of these three as one composite variable, as the PCR+ approach does? As long as we try to partition this vector into three separate effects, there is a high probability of misattribution. In the eigenvector approach, one would refrain from attributing the effects of the combined variable to any of its components.

5. DIESEL FUEL EIGENVECTORS

A fundamental conviction of our approach to emissions analysis is that eigenvectors are a better way to describe diesel fuels than the individual properties, because the vectors reflect the naturally-occurring correlations among the physical and chemical properties of the fuels. This conviction is based on the ability to interpret the eigenvectors of diesel fuel data sets in refining terms and on the repeatability of eigenfuel characteristics from one data set of fuels to another. If we are correct, then the diesel eigenvectors are a more fundamental set of variables for describing fuels than the individual physical and chemical properties.

5.1 EIGENFUEL STRUCTURE OF COMMERCIAL DIESEL FUELS

The assessment of diesel fuels should be informed by a thorough understanding of the characteristics of commercial fuels, yet survey information on commercial fuels is difficult to obtain and often provides very limited information on fuel properties. The *Commercial Fuels Database* (CFD), a proprietary survey of diesel fuels in the United States, was made available to this study by an anonymous member of the automotive fuels industry. The database contains approximately 100 fuels, surveyed during the mid-1990's, with both seasonal and geographic diversity represented in the sampling. Importantly, a wide range of physical and chemical properties are reported for each fuel. The fuels do not contain cetane additives or oxygenates, but other additives (e.g., viscosity improvers) may be present depending on commercial practice. We take these fuels to be representative of commercial diesel fuels in the marketplace.

A PCA analysis was conducted to identify the eigenvector structure of commercial diesel fuels. A total of 10 features can be identified from the 10 fuel properties considered here: natural cetane, specific gravity, viscosity, sulfur content, aromatics content, and five points on the distillation curve. The eigenvector structure tells us how real fuels in the current market vary, but we must remember that this may not fully characterize future fuels. For example, the eigenstructure of current fuels may not fully reflect, or could even omit, modes of variation that could be exploited in reformulating fuels to reduce sulfur content, reduce emissions, or achieve other objectives. The results of the PCA analysis of current commercial fuels may be found in Appendix A, Table A.4.

Table 5.1 summarizes the primary eigenfuel characteristics of commercial fuels. The table describes each vector qualitatively in terms of the primary properties of which it is comprised, the directionality of the relationship between properties, and the strengths of the relationships. The vectors are also given interpretations in terms of the diesel fuel blendstocks that may be associated with the property changes. While we believe these interpretations are plausible, they are subject to change with additional research on the relationships between diesel fuel eigenvectors and refinery blending and processing mechanisms.

Of the total of ten eigenvectors, only five are required to represent 95 percent of the variation among fuels. Commercial Fuel Vector 1, the light cycle oil vector, accounts for nearly half of the differences among commercial fuels. This vector reflects an association of properties in which decreased aromatics coincides with increased natural cetane, decreased specific gravity and viscosity, and lower temperatures throughout the distillation curve. These property changes would be expected with removal, by distillation, of light cycle oil from the diesel stream. Directionally opposite changes would be expected for blending increased percentages of light cycle oil.

Table 5.1. Eigenfuel Structure of Commercial Diesel Fuels [a,b] (representing 95 percent of fuel variation)

Eigenvector	Variance (percent)	Description
1	48	**Light Cycle Oil (or "Back End") Feature**: A decrease in aromatics content is associated with increased natural cetane, decreased in specific gravity and viscosity, and lower temperatures throughout the distillation curve. These property changes would be expected with removal, by distillation, of *light cycle oil*. Directionally opposite property changes would be expected for blending increased percentages of light cycle oil.
2	17	**Hydroprocessed Heavy Distillate Feature**: Decreases in aromatics and sulfur content are associated with increased natural cetane, increased viscosity, and higher temperatures at the low end of the distillation curve. These property changes might result from blending increased percentages of *hydroprocessed (hydrotreated or hydrocracked) heavy distillate*.
3	13	**Straight-Run Heavy Distillate Feature**: A decrease in aromatics content is associated with increased natural cetane, an increased slope to the distillation curve, and increased sulfur content. These property changes might result from increased blending percentages of (unhydrotreated) *straight-run heavy distillate*.
4	9	**Straight-Run Light Distillate Feature:** An increase in sulfur content is associated with decreased back end temperatures, but is largely independent of other property changes. These property changes might result from increased blending percentages of (unhydrotreated) *straight-run light distillate*.
5	7	**Initial Boiling Point Feature**: A vector representing variation in the IBP, largely in isolation from other properties except sulfur content, and apparently representing blending to control flash point. Directionally opposite property changes would be expected with reduced blending percentages of straight-run heavy distillate or increased percentages of straight-run light distillate.

[a] All fuels are clear of cetane additives and oxygenates.
[b] Interpretations presented in this table are subject to change with additional research on the relationship between eigenvectors and refinery blending/processing mechanisms.

Similarly, Commercial Fuel Vectors 2 through 4 reflect distinct associations among the physical and chemical properties that appear to be related to the content of other common diesel blendstocks, including hydroprocessed heavy distillate, straight-run heavy distillate, and straight-run light distillate. The last feature, Commercial Fuel Vector 5, involves variation of the IBP, largely in isolation from other properties, and probably represents blending to control flash point.

As described in Section 6 of this report, it would be possible to base an experimental design for emissions research (or another purpose) on variations in the extent to which diesel fuels express these independent

features. Such experimental fuels would be blended from refinery stocks commonly used in diesel fuels, and the individual physical and chemical properties would be allowed to vary without constraint as the natural outcome of the eigenvector variations imposed by the design. As shown below, a much different and more conventional approach is taken in current fuels research.

5.2 EIGENFUEL STRUCTURE OF EXPERIMENTAL FUELS

Research on the relationship between diesel fuels and HDD emissions has typically been conducted using experimental fuels that are purposefully blended to vary one or more of the individual fuel properties independently, while otherwise attempting to assure that the resulting test fuel is "representative" of real diesel fuels. For example, one study (U.S. EPA 1999) selected a series of blendstocks found in refinery streams that differed in natural cetane, specific gravity, and mono- and poly-aromatics content. Experimental fuels were blended to vary those four properties according to a balanced experimental design. With a sufficient number of blendstocks, the researchers could closely approximate independence among the properties. In another instance (Tanaka *et al.* 1996), researchers used a series of simpler chemical compounds to concoct test fuels that varied desired properties independently. More emphasis was placed on achieving property independence and less emphasis on real-world "representativeness" in the selection of blendstocks.

The construction of fuels in these two studies illustrates an inescapable aspect of experimental design involving diesel fuels: *the correlations among properties that occur in test fuels represent a combination of the naturally-occurring correlations that are characteristic of the blendstocks and volitional correlations introduced by the experimenters as a result of their blending strategy.* We can reasonably expect the natural correlations *to recur* from one set of experimental fuels to another whenever the same or similar sets of diesel blendstocks are used. We can also expect the volitional correlations *to change* from one fuels data set to another as the experimental designs and blending strategies of the studies differ. Therefore, we should not be surprised to see that experimental fuels, to a large extent, express features similar to those in commercial fuels, but may combine them in new ways or mix them with features not found in commercial fuels.

To test this understanding, let us examine the eigenfuel structure of the diesel fuels used in the emissions testing in the EPA database. The test fuels reflect the experimental designs, blendstocks, and blending strategies of a large number of different research studies from which the fuels and emissions data were drawn. The test fuels also introduce two features – cetane additives and oxygenates – that are not found in the database of commercial fuels, and we have therefore excluded such fuels from the following comparison. Having limited the comparison to clear fuels (without cetane additives or oxygenated), a PCA analysis was conducted using the 10 selected property variables: natural cetane, specific gravity, viscosity, sulfur content, total aromatics content, and five points on the distillation curve. The PCA analysis results may be found in Appendix Table A.5.

Table 5.2 summarizes the primary eigenfuel characteristics of the clear test fuels. Five vector features again account for 95 percent of the variation among fuels; the steep decline, from most important to least, in the variance accounted for by each vector is remarkably similar to that seen in the commercial fuels. One can also see similarities in the eigenvectors of the test and commercial fuels by comparing the table with the earlier descriptions of the commercial fuel eigenvectors. For example, the first eigenvector in each set involves a directionally similar relationship among natural cetane, total aromatics content, specific gravity, viscosity and temperatures on the distillation curve. In each case, however, specific details of the vectors differ, and we have therefore given the vectors different labels.

Table 5.2. Eigenfuel Structure of Diesel Test Fuels [a] (representing 95 percent of fuel variation)

Eigenvector	Fuels Variance (percent)	Eigenvector Description
1	49	**Test Fuel Vector 1**: A decrease in specific gravity and viscosity is associated with lower temperatures throughout the distillation curve, modestly lower total aromatics content and modestly higher natural cetane.
2	21	**Test Fuel Vector 2**: A decrease in aromatics content is associated with increased natural cetane, decreased specific gravity, and a modestly increased slope to the distillation curve.
3	12	**Test Fuel Vector 3**: Increased sulfur content is associated with an increased slope to the distillation curve and modestly higher aromatics content and natural cetane.
4	8	**Test Fuel Vector 4**: Decreased sulfur content is associated with an increased slope to the distillation curve, modestly higher aromatics content, and modestly lower natural cetane
5	4	**Test Fuel Vector 5**: A vector representing the variation in IBP that is possibly related to controlling the fuel flash point to commercial standards.

[a] Excludes test fuels containing cetane additives or oxygenates

The differences in labeling have the effect of concealing similarities and do nothing to reveal relationships among the vectors of commercial and test fuels. More insight can be gained by relating the vectors to each other quantitatively. The method[13] of relating the vectors is based on the realization that an eigenvector of the test fuels data set has the form of a fuel and can, therefore, be expressed in terms of the eigenvectors of the commercial fuels data set. Once a test fuel has been expanded in terms of the commercial eigenvectors, its coefficients are then squared. By the orthonormal property, the squared coefficients must sum to unity, and the squared coefficients give us in relative terms the extent to which each test fuel expresses the features found in commercial fuels.

The quantitative relationships among the five primary eigenvectors are displayed in Figure 5.1 and tabulated in the related Table 5.3. In the figure, each row represents one of the five primary eigenvectors of the test fuels data set. The columns in the row represent the extent to which the test fuel vector expresses the eigenfuel features found in the commercial fuels, and the column heights measure the relative contribution. As can be seen, each of the primary vectors in the test fuels data set expresses from one to three primary features of the commercial fuels, and nearly all of the primary test fuel vectors can be accounted for as mixtures of the five primary commercial fuel features.[14]

[13] See McAdams *et al.* 2001b, Section 3.3 and Appendix F.

[14] The columns within each row would sum to 1.00 if all ten commercial eigenvectors were accounted for. Considering only the five primary commercial vectors, the rows sum to values in excess of 0.95 for all test vectors except number 3. Test Fuel Vector 3 sums to 0.86 across the first five commercial vectors and to 0.96 if commercial vector 6 is included.

Figure 5.1. Relationship of Test Fuel Eigenvectors to Commercial Fuel Features

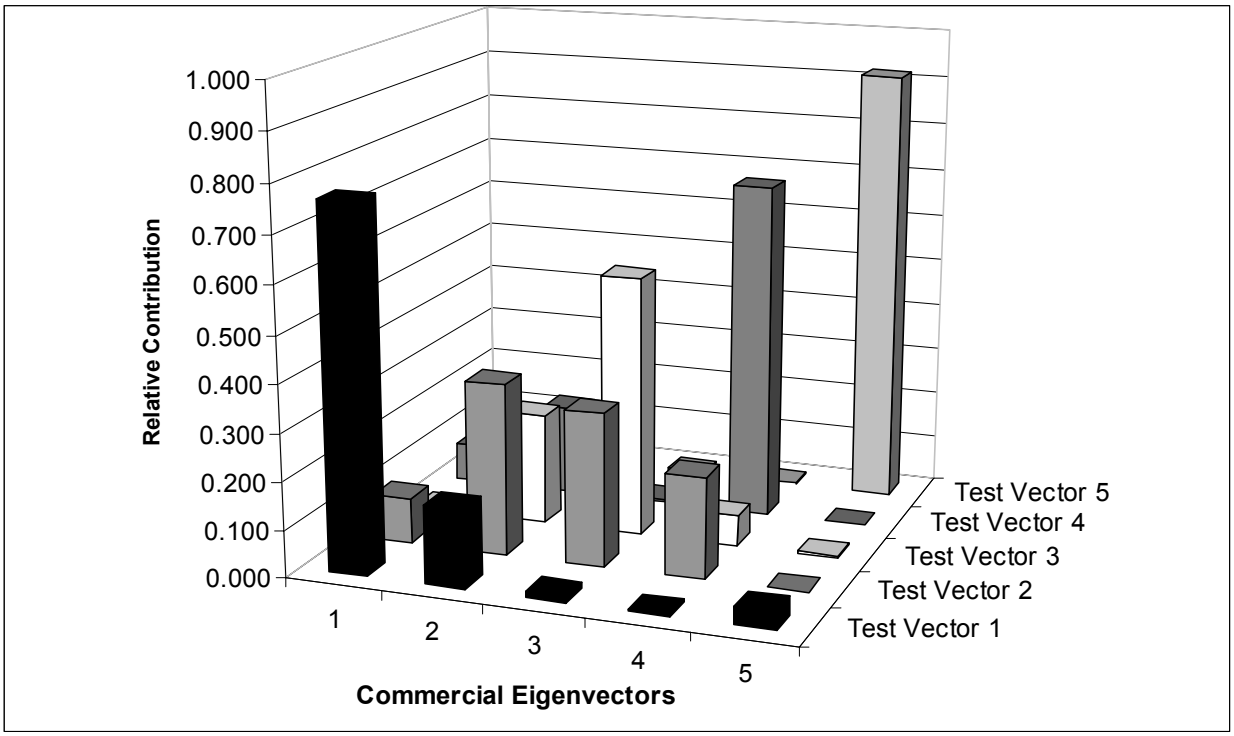


Table 5.3. Representation of Test Fuel Eigenvectors in Terms of Commercial Fuel Features

Commercial Fuel Vector	Squared Commercial Eigenvector Coefficient (Relative Contribution to Test Fuel Vector)				
	Test Fuel Vector 1	Test Fuel Vector 2	Test Fuel Vector 3	Test Fuel Vector 4	Test Fuel Vector 5
1. Light Cycle Oil	**0.763**	0.096	0.006	0.082	0.033
2. Hydroprocessed Heavy Distillate	**0.166**	**0.361**	**0.235**	**0.188**	0.000
3. Straight-Run Heavy Distillate	0.014	**0.323**	**0.549**	0.006	0.005
4. Straight-Run Light Distillate	0.005	**0.214**	0.065	**0.709**	0.003
5. Initial Boiling Point	0.038	0.002	0.004	0.000	**0.918**
6. (uninterpreted)	0.007	0.001	**0.107**	0.005	0.001
All Others	0.007	0.003	0.034	0.010	0.040
Total	1.000	1.000	1.000	1.000	1.000

Overall, the five primary test fuel vectors account for 95 percent of the variation among the test fuels. Test Fuel Vector 5 is essentially the same as Commercial Vector 5, with 92 percent of the test vector being an expression of the corresponding commercial fuel vector. Test Fuel Vectors 1 and 4 are mixtures that are weighted heavily toward one of two commercial fuel vectors, while Test Fuel Vectors 2 and 3 are more

equally-weighted mixtures of 3 commercial fuel vectors. Almost all of the total variation among test fuels is associated with one or more of the five primary commercial fuel vectors, which together account for 92 percent of the variation among the test fuels.

It is clear, therefore, that the commercial fuel eigenvectors tend to be replicated in the data set of experimental fuels, although they often appear in combination with one or more other vectors to form a new "experimental feature" of fuels. A trivial example is that of Test Fuel Vector 5 (Initial Boiling Point Feature), which maps nearly 1:1 into the corresponding Commercial Fuel Vector 5 (Initial Boiling Point Feature). Here, we can interpret the vector as representing the experimental objective that the test fuel flash point must meet commercial standards.

Other test fuel vectors can be interpreted in terms of experiment designs and blending strategies. For example, Test Fuel Vector 1 is formed by mixing two different streams (light cycle oil and hydroprocessed heavy distillate), in which one is increased and the other decreased, to produce fuels that differ in their overall "heaviness." Because both streams have inverse (although different) relationships between aromatics content and natural cetane, the effect on the aromatics content and natural cetane number of the final blend is muted. One stream (light cycle oil) has strong effects on specific gravity, viscosity, and the middle and back end of the distillation curve, while the other (hydroprocessed heavy distillate) has strong effects on specific gravity, viscosity, and the front end of the distillation curve. Combined in varying ratios, we see a test fuels vector that is described as:

> A decrease in specific gravity and viscosity is associated with lower temperatures throughout the distillation curve, modestly lower total aromatics content and modestly higher natural cetane.

The five primary vectors of the test fuels data may represent the following experimental designs and blending strategies:

- **Test Vector 1**. A feature for the study of the effect of fuel "heaviness," particularly specific gravity, on emissions, in which light cycle oil (Commercial Vector 1) and hydroprocessed heavy distillate (Commercial Vector 2) are blended to base fuels to yield finished fuels of varying specific gravity, viscosity, and distillation temperatures. Both stocks have higher specific gravities, but changes in aromatics contents and cetane numbers are offsetting, by design.
- **Test Vector 2**. A feature for the study of the effect of total aromatics content on emissions, in which hydroprocessed and straight-run heavy distillate (Commercial Vectors 2 and 3) are combined with straight-run light distillate (Commercial Vector 4). The heavy distillate stocks are partially hydroprocessed to remove the effects of sulfur content when combined with straight-run light distillate.
- **Test Vector 3**. A feature related to sulfur content, in which hydroprocessed and straight-run heavy distillate (Commercial Vectors 2 and 3) are combined (with contributions from the uninterpreted Commercial Vector 6) to vary fuel sulfur content. There are also variations in distillation characteristics, with relatively smaller effects on total aromatics content and other properties.
- **Test Vector 4**. A second feature related to sulfur content, in which straight-run light distillate (Commercial Vector 4) and hydroprocessed heavy-distillate (Commercial Vector 2) are substituted for each other to vary sulfur content and distillation characteristics, with muted effects on aromatics content and other properties. Unlike the other test vectors, which reflect property variation experiments, it is possible that test vector 4 arises from intentionally designed blendstock substitution experiments.

- **Test Vector 5**. A feature related to the IBP and control of flash point to commercial standards, that is substantially the same as the corresponding vector in the commercial fuels.

5.3 DISCUSSION AND SUMMARY

The foregoing presentation has shown that the major characteristics of the test fuels in the EPA database can be understood as tailored recombinations of the underlying characteristics found in commercial diesel fuels. The methods of recombination appear to represent decisions and strategies employed by researchers in creating test fuels that attempt to vary individual properties in isolation from each other. The commercial fuel characteristics may be combined in novel ways in this quest, but the characteristics are nevertheless carried forward into the final test fuel blends through the diesel fuels and blending components used as raw materials.

That the experimental fuels are largely reexpressions of the commercial fuel characteristics lends credence, we believe, to our conviction that the eigenvectors of the commercial fuels data set offer a more natural and fundamental set of variations for diesel fuels analysis. We see here that widely varying experimental fuels have been created, often with considerable difficulty, by creative combinations of more basic fuel characteristics. These vectors carry forward the *naturally-occurring* correlations among fuel properties inasmuch as commercial fuels were used as the "raw materials" for blending.

The intentional recombination of the commercial vectors into new experimental features that fit preselected experimental designs leads to the introduction of *volitional* correlations among properties in the test fuels. By nature of their definition as eigenvectors of commercial fuels, the commercial vectors are independent of each other and can be freely combined in forming commercial fuels. When combined in a systematic manner designed to achieve an external objective, the researcher has imposed additional constraints that can be observed in the form of differing overall correlations among the fuel properties. These volitional correlations will vary unpredictably from one data set to another as the external objectives themselves change.

These results lead to the conclusion that research into diesel fuel formulation and the relationship of diesel fuels to emissions would be advanced and clarified by adopting the eigenvector characteristics of commercial diesel fuels, in place of the individual fuel properties, as a more appropriate and fundamental set of variables for describing fuels. The volitional choices made by researchers perturb the naturally-occurring relationships among fuel properties in ways that will change from one data set to another. The additional constraints posed by experimental designs based on the individual fuel properties will serve only to confound and confuse the creation of realistic test fuels.

Further, because the characteristics of the test fuels can in large part be traced back to the underlying characteristics of commercial fuels, it should be recognized that test fuels can be designed directly as combinations of the commercial eigenvectors. The characteristic features of commercial fuels can be sampled in ways that differ from the mix found in current commercial fuels, thereby giving the eigenvectors novel weights. Cetane improvers or new features, such as oxygenates, can be added in the form of new vectors. The next section will explore more closely the use of eigenvectors in experiment design.

6. USE OF EIGENFUELS IN EXPERIMENT DESIGN

So widely acknowledged is the intimate relationship between experiment design and data analysis that they form a formal duality bonded by a common calculus. If we design the experiment on an orthogonal basis at the outset, then our data analysis is blessed with a design matrix that is orthogonal in terms of the scalar variables in the data set. If we do otherwise, our design matrix will be nonorthogonal, but the techniques of PCA and PCR+ can be used to identify an orthogonal set of predictor variables. Our "uncontrolled" experiment was, in a sense, implicitly designed on this *ex post* orthogonal basis.

In this section the duality between experiment design and data analysis is explored for the purpose of developing a role for eigenfuels in the design of diesel fuel and emissions experiments. The fundamental methodological principles of eigenfuel-based experiment design are developed first. Then, we show how eigenfuels can be used in practical ways to guide the blending of test fuels to meet the specific requirements imposed by an experiment design. Appendix E explores these concepts in greater detail.

6.1 METHODOLOGICAL PRINCIPLES OF EXPERIMENT DESIGN

We begin with a fundamental postulate:

> For *any* data set with a design matrix of full rank, there exists an orthogonal basis consisting of combinations of the original predictor variables. Data analysis, such as multiple regression analysis, is preferably performed in this orthogonal space, primarily because the transformation removes all aspects of multicollinearity that may have existed in the original predictor variables.

One might say that, in a certain sense, every experiment is designed orthogonally; it is only necessary to find its orthogonal basis. Once that basis is defined, the experiment can be analyzed *as if* it were designed as an orthogonal array at the outset. This array consists, of course, of the vectors that emerge from PCA, and those vectors are the redefined variables.

An investigator has two options at the outset. He or she may *arbitrarily* choose a set of variables believed to be the pertinent ones and then lay out treatments that explore the predictor-variable space, selecting the treatments in such a way that the effects of each of the variables can be independently estimated. Alternatively, he or she may define the treatments *arbitrarily* and then redefine the variables in such a way that the effects of each of the redefined variables can be independently estimated. In this sense, the two approaches can be said to be *duals* of each other. In the one case, the experimenter attempts to untie the knots that tie the variables together. In the other case, the knots are accepted as they are.

The conventional approach, in which variables are used to determine treatments, grew out of applications in which there was no difficulty in controlling the levels of one variable independently of the levels of another. For example, in agricultural experiments, in which the term "treatment" arose, one could apply various amounts of fertilizer A to plots quite independently of applying the same or different amounts of fertilizer B to those plots. In this case, one can define a set of treatments that allows the experimenter to determine the independent effect of each of the two fertilizers on crop yield, as well as certain other effects known as *interactions* in the statistical literature.

When the experiment design methodology is used in another area of inquiry, however, situations may arise in which it is not possible or feasible to vary the levels of two variables independently, because there are natural forces that cause the variables to be correlated to greater or lesser extent. An obvious case is present in the design of experiments related to *mixtures* of materials, as in the formulation of fuel blends. For example, if a mixture is made of three materials, the experimenter can freely choose the amounts of components A and B in the blend, but *cannot* choose the amount of component C inasmuch as the three components must total 100 percent.

What complicates the matter is the circumstance where two or more variables are partially correlated, particularly as a result of natural forces. In some instances the experimenter may attempt to "break" the association between variables, and in other instances he or she may set levels of one variable but let the other variables "seek their own levels," accepting whatever levels the variables will take in nature. The disposition of treatment levels is, therefore, partly *volitional* as a result of experimental choice and partly *involitional* as a result of natural forces.

When a collection of fuels is subjected to PCA, the definition of the characteristic "eigenfuels" may be at least partly be an artifact of conscious design effort. We saw evidence of this in the Section 5, where the eigenfuel characteristics of experimental fuels could be understood as a purposeful blending of the more general characteristics of commercial fuels. A new experiment involving a similar collection of fuels could very well lead to a somewhat different set of eigenvectors simply as a result of the experimenter's effort or lack of effort to "break" the correlations among properties and attain a classical, balanced design.

In what follows it is assumed that a set of eigenvectors, perhaps those found in commercial diesel fuels, has been agreed upon and that we wish to perform additional experiments to explore further the specific effects of those eigenfuels. We consider the process of experiment design as one in which the eigenvariables are held subject to the same constraints as would be the case for any variable capable of being manipulated independently.

6.2 DEMONSTRATION AND APPLICATION OF PRINCIPLES

Rather than proceed with abstract principles, we elect to develop the essential concepts by demonstrating how the principles come into play in connection with real data on diesel fuels. We use for this purpose the data set of 280 diesel fuels that formed the original basis for the development of PCR+ in McAdams *et al.* (2000b). Twelve eigenvectors characterize this data set in which the following twelve properties of diesel fuel were measured: natural cetane, specific gravity, viscosity, sulfur content, mono-aromatics content, poly-aromatics content, and five points on the distillation curve. There is considerable redundancy in the data set compared to the actual number of fuels, because multiple emissions tests have been performed on many of the fuels. The fuel data set, expressed in terms of the coefficients of the eigenfuels, is a 280 by 12 array, the rows denoting the number of emissions tests and the columns denoting the twelve eigenfuels.

The task of eigenfuel-based experiment design is approached by:

- First, specifying the treatment levels of a balanced design in terms of the accepted slate of eigenfuel characteristics
- Then, selecting real fuels as needed to fulfill the chosen design.

To begin, let us set up a very simple array of numbers (see Table 6.1) that can be placed in one-to-one correspondence with a balanced experiment design, specifically a 2^3 factorial, defined as an experiment in which there are three variables, each set at two levels. This array is clearly columnwise orthogonal – that is,

the inner product of any two columns is zero. Consequently, it serves as a model for a fuel subset having like characteristics if such can be found among existing fuels or can be produced by blending existing fuels. This generalized matrix is related to the eigenfuel characteristics of fuels as follows:

1. We assign each column to represent the level of a single eigenfuel. The first column represents eigenfuel 1, which was interpreted in the prior work as a fuel viscosity/specific gravity characteristic. The second column represents eigenfuel 2, an aromatics content feature, and the third column represents eigenfuel 3, a sulfur content feature. These characteristics were found to account for nearly three-fourths of the differences among fuels, and strongly related both to NO_x and PM emissions. We assume in this example they are adequate to represent the fuel characteristics of relevance to the planned experiment.
2. We map each level (-1 and +1) into a quantitative value for the "amount" of the associated eigenvector that the treatment expresses. We will assume in this example that the level values measure eigenfuel weights in terms of standard deviations from the mean, which is a convenient choice when using centered and standardized variables. Further, it means that we are varying each eigenvector by an amount that can be considered of "equivalent difficulty" once we take into account the amounts by which the characteristics are found to vary in real fuels. A greater or lesser range could be employed simply by performing a linear transformation between the level value and the desired eigenfuel weights.

Table 6.1. Binary Representation of a 2^3 Factorial Experiment

-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

Thus, a balanced experimental design of eight fuels is defined in which the levels explore variations in the "amounts" of the first three eigenvectors above and below the amounts found in an average fuel. The task now becomes one of finding, among the fuels of the overall data set, the eight fuels that most closely approximate the above array.

The method used in finding the "best fit" fuels is that of minimizing the differences between the "ideal" or "target" eigenfuel weights in the above table and the actual eigenfuel weights of fuels in the data set. The Figure of Merit (FM) is the squared deviation of a particular fuel from the target weights for the first three eigenvectors; the weights of the remaining nine eigenvectors are not controlled in this experiment design. The fuels listed in Table 6.2 are those from the original data set with the smallest FM, which for a perfect fit would be zero. The "fit" of the candidate fuels to the target fuels is not very good in many cases.

Table 6.2. Approximation of a 2^3 Factorial Array by Fuels Selected from a Fuels Data Set

Fuel Number	Weight of Eigenfuel			Figure of Merit
	1	2	3	
217	-0.8829	-0.5011	-1.2429	0.5671
40	-1.4034	-0.2630	0.8515	0.8532
148	-1.7259	0.7117	-0.6660	0.8494
259	-0.2935	0.7253	1.4268	0.8699
213	0.5218	-0.2641	-1.1766	0.8952
124	0.7486	-0.9352	-0.0175	1.0501
226	0.8011	1.0553	-0.4712	0.5677
223	0.4434	1.2410	0.3738	0.8718

The fit can be improved by *blending* fuels selected from among the 280 available in the fuel data set, rather than looking for the single closest fuel. One method for this is to select randomly sets of four fuels and solve the system of four equations in four unknowns (the three design level values plus the constraint that the blending proportions sum to 1.00). Fuel sets having nonnegative blending proportions are feasible solutions to the problem. With this approach, the design level values can be achieved with arbitrary precision, subject only to the assumption that eigenfuels blend linearly. Table 6.3 shows one such blend, in which four fuels are combined to exactly match the target weights for the first three eigenvectors in a treatment level, while allowing the weights for the fourth (and later) eigenvectors to "seek its own level." Further examples and the results of this blending procedure are presented in Appendix E.

Table 6.3. Sample Fuel Blend Satisfying Target Weights for a Treatment Level

Fuel Number	Blend Weight (percent)	Coefficient of Eigenvector Number			
		1	2	3	4
43	6.14	-0.1530	-0.2095	0.1513	-0.0172
71	70.95	0.0161	-0.5844	-0.8058	-0.1432
176	16.44	-0.8717	-0.0579	-0.3345	-0.0185
114	6.47	0.0086	-0.1482	-0.0110	-0.0086
Sum	100.00	-1.0000	-1.0000	-1.0000	-0.1875
Target Weights		-1.0000	-1.0000	-1.0000	n/a

It should be realized that there can be multiple blended fuels combinations that satisfy the restrictions placed on the first three eigenvectors, but which vary with regard to the levels of the remaining nine eigenvectors. These multiple solutions could be considered as "replicates" from which one could obtain an estimate of the error resulting from the assumption that only the first three eigenvectors need be considered.

6.3 DISCUSSION AND SUMMARY

It has been demonstrated that it is possible to blend real fuels in such a way as to produce "treatments" that satisfy the requirements of factorial design. Application of these principals can satisfy other experimental designs including fractional factorial and random balance designs. Further, as shown in the appendix, the method is practical in that, from the 76 distinct fuels used in the data set of 280 emissions tests, there are thousands of possible fuel blends that are admissible solutions for a three-variable design. On the other hand, the yield of admissible blends could be sparse if one wants to control more than three eigenvector characteristics, and that fact could necessitate having a sizable range of fuel properties from which to draw. The test fuels produced from this method are known to be producible in a refinery by virtue of their occurrence in a data base of real diesel fuels.

Consider how fuels and emissions research might be conducted differently if the eigenfuel methodology has been adopted at an early stage. Firstly, assume that the eigenvector characteristics of the CFD, seen in the prior section, were generally accepted as useful descriptions of the features of diesel fuel. Secondly, assume that vectors have been added to this set to represent the use of additives such as cetane improvers and oxygenates.

Without prior art to focus attention on individual properties of fuels such as specific gravity, aromatics content, or natural cetane, we would formulate a factorial (or other orthogonal) experiment design in seven variables – the five eigenfuels that account for 95 percent of the differences among current commercial fuels plus the two vectors representing additives. Test fuels that are known to be producible in a refinery could be formulated by blending selected diesel fuel stocks as shown above. The resulting data set would carry all of the advantages of an orthogonal basis and would allow the data analysis to assess *without ambiguity* the relationship of real diesel fuel characteristics to emissions, including interaction terms and (allowing for replicates) the assumption that only the selected characteristics need be considered in regard to emissions. We believe that the state of current knowledge regarding the relationship between diesel fuel characteristics and engine emissions would be considerably advanced compared to its present state.

7. EIGENFUELS IN FUELS RESEARCH

Although the motivation for PCR+ was as a technique for improved data analysis in regard to fuels and engine emissions, eigenfuels can also be a useful technique for solving problems in fuels research related to the characteristics of fuels. In this section we present a methodology for conducting a Monte Carlo simulation to generate synthetic, but realistic, fuels data under explicit control of the experimenter. This methodology is applied to a specific problem of estimating fuel characteristics.

7.1 DESIGN OF MONTE CARLO SIMULATIONS

Eigenfuels are understood to be vector descriptions of characteristic fuel properties that are derived through use of PCA. When fuels are expressed in terms of the eigenvectors, the coefficients or weights for the vectors are uncorrelated over the data set from which the vectors were estimated and are distributed with mean zero and variance given by the eigenvalues. Thus far, we have been primarily interested in the eigenvectors as predictor variables for emissions or another response.

Let us now put eigenvectors to work, in reverse, as the basis for synthesizing fuel data. If the vectors are independent of each other, we can sample from them independently to define hypothetical fuels. If there are N vectors, we generate N random values and use those values as the weights associated with the N vectors in the new fuel. If, at the time of sampling, we set the variance for the n^{th} vector equal to the n^{th} eigenvalue, we will sample the eigenvectors using the same weights with which they occur in the original data set of fuels. The hypothetical fuel will have the same distributional properties as the original fuels, and we can reasonably treat it as a new, hypothetical realization of the same processes that created the original fuels. By modifying the variances, we can create new fuels that are based on the same processes but which weight the processes in a unique manner. The utility of this capability, ranging from estimating sample sizes for emissions testing to assessing the likely performance of statistical models when applied to new data, will be apparent to analysts.

7.1.1 Replicating Fuel Characteristics

As a first example, let us use the Monte Carlo process to create a new data set by sampling from the eigenvector slate developed for the commercial diesel fuels. As described in Section 5, the commercial fuels data set has ten fuel property variables and ten eigenvectors. Using the process outlined above, one hypothetical fuel is created by generating ten, normally-distributed random numbers, of mean zero and variance corresponding to the eigenvalues of the commercial fuels data set, and identifying the vector of ten values as the eigenvector expression of a fuel. The eigenvector expression is then converted back to its equivalent in terms of fuel properties. The process is repeated 999 times to create a set of 1000 synthetic diesel fuels.

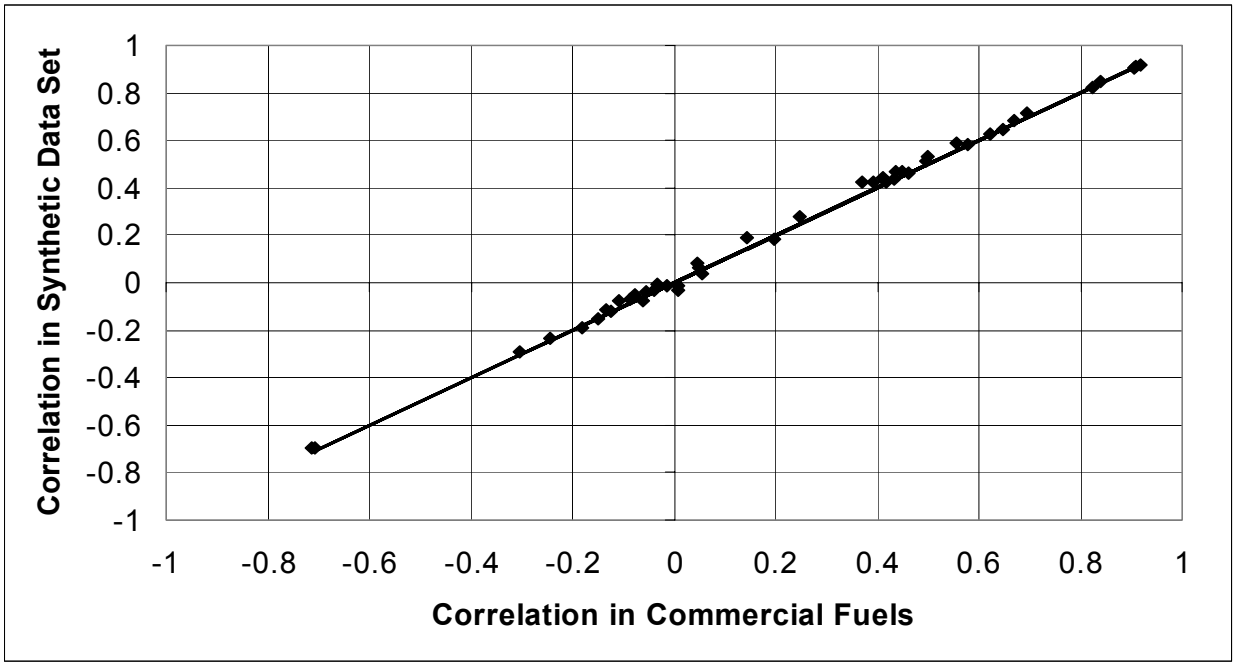
That the synthetic data replicates the characteristics of the original data set can most easily be seen by tabulating the mean values and standard deviations of the individual fuel properties, as in Table 7.1. The mean values and standard deviations of the two data sets are essentially identical, and the differences seen in the table result only from the fluctuations normally encountered in a random process.

Table 7.1. Statistical Summary of Synthetic Data Set No. 1

Fuel Property	Units	Commercial Fuels		Synthetic Data Set No. 1	
		Mean	Std Dev	Mean	Std Dev
Natural Cetane	num	45.4	2.8	45.4	2.8
Specific Gravity	g/cm^3	0.850	0.0086	0.850	0.0086
Viscosity	mm^2/sec	2.67	0.30	2.68	0.31
Sulfur Content	ppm	352	60	353	60
Aromatics Content	vol percent	33.2	5.1	33.3	5.0
IBP	deg F	349	23	348	23
T10	deg F	429	18	429	18
T50	deg F	513	15	514	16
T90	deg F	607	15	608	15
FBP	deg F	653	17	654	17

Further, the sampling process based on eigenvectors replicates the correlations among variables as found in the original data. This occurs because the eigenvectors are defined by the PCA decomposition of the correlation or covariance matrix. We can see this result clearly by calculating the pairwise correlation coefficients among fuel properties in the synthetic sample and then graphing them against the corresponding correlation coefficients in the original data. When this is done, as in Figure 7.1, the points cluster closely around the unity diagonal.

Figure 7.1. Comparison of Fuel Property Correlations in Synthetic Data Set No. 1



Overall, using the ten fuel properties with which the data are described, it is not possible to distinguish the synthetic data from the original (real) data in terms of:

- Average fuel properties
- Standard deviation of the properties
- Correlations among the properties

7.1.2 Creating New Fuel Characteristics

The pairwise correlations existing between fuel properties in the simulated data is controlled by the eigenvalues (variances) associated with the eigenvectors. If we change the variances used in the sampling, new fuels are generated that are possible realizations of the same processes represented by the vectors, but with different weights for the vectors and a different structure of correlations among the properties.

How one would modify the variances will depend, of course, on the problem to be studied. Fuel aromatics content is of great current interest in regard to engine emissions. For purposes of illustration, let us assume that an experimenter might want to know the *widest* range of fuels that are consistent with current methods of varying aromatics content. For commercial diesel fuels, the first three vectors describe features related to aromatics content and account for 48 percent, 17 percent and 13 percent of the differences among fuels, respectively. An illustrative sampling plan would be to give each of the three aromatics vectors an equal opportunity to be given the greatest (or least) emphasis. The following algorithm was used. Given there are six ways in which the variance weights of the first three vectors can be reordered [(48, 17, 13), (48, 13, 17), (17, 48, 13), etc.], one of these ways is chosen at random each time a fuel is generated and used as the weights for vectors 1, 2, and 3. A total of 1000 fuels is generated in the Monte Carlo simulation.

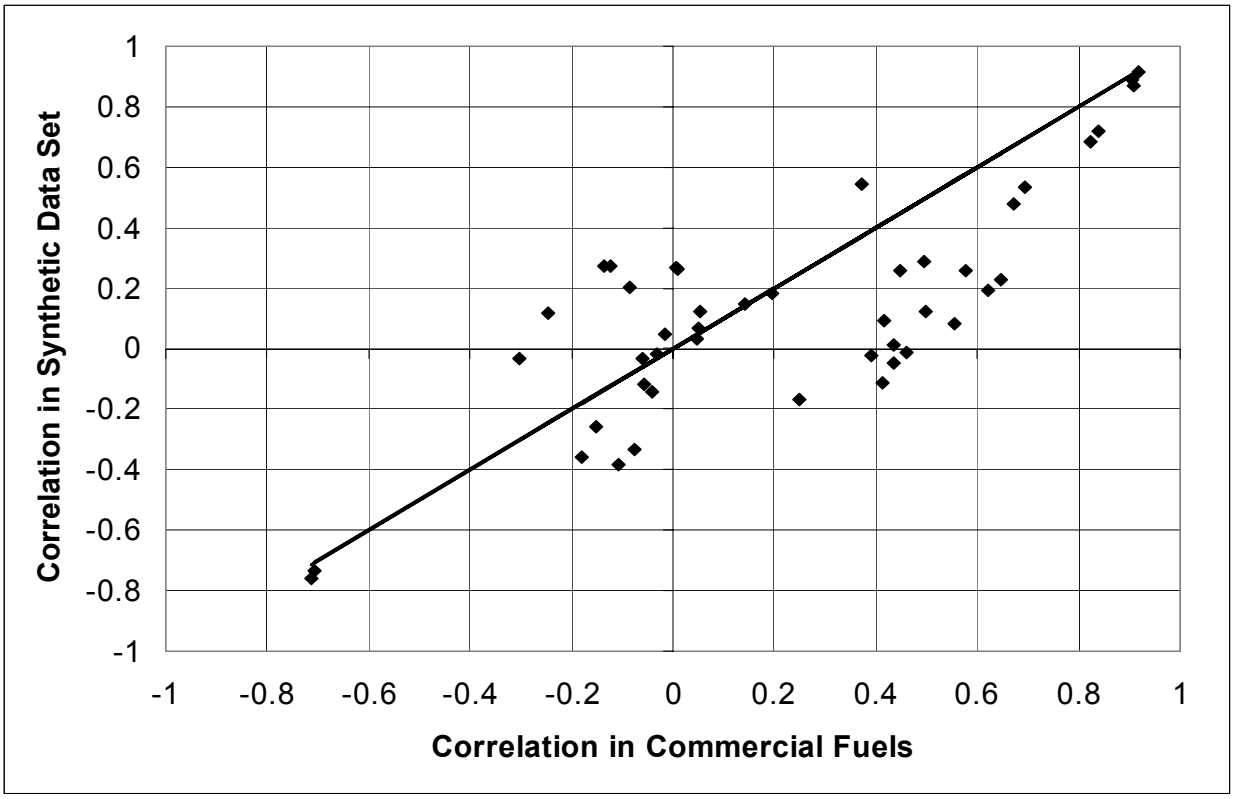
Table 7.2 summarizes the means and standard deviations of the fuel properties for this second set of synthetic data. The mean values are essentially unchanged from those of the original data, which does nothing more than illustrate the property that eigenfuels, as defined using "centered" data, measure variations relative to an average fuel. If all simulated fuels are accepted without *ex post* selection, the mean values of the data set will closely replicate those of the original data. The property standard deviations are also relatively little changed, although they do not replicate the original data as closely as the first synthetic data set.

Table 7.2. Statistical Summary of Synthetic Data Set No. 2

Fuel Property	Units	Original Data		Synthetic Data Set No. 2	
		Mean	Std Dev	Mean	Std Dev
Natural Cetane	num	45.4	2.8	45.2	3.2
Specific Gravity	g/cm3	0.850	0.0086	0.850	0.0075
Viscosity	mm2/sec	2.67	0.30	2.66	0.26
Sulfur Content	ppm	352	60	352	65
Aromatics Content	vol percent	33.2	5.1	33.6	5.1
IBP	deg F	349	23	349	27
T10	deg F	429	18	428	16
T50	deg F	513	15	514	13
T90	deg F	607	15	608	15
FBP	deg F	653	17	654	17

What has been done is to explore a wider range of *ways* in which the aromatics characteristics might be varied in fuels without fundamentally changing the overall range of the original properties. That something is actually different in the second data set can be seen in Figure 7.2. The pairwise correlations among the properties now differ considerably from what was found in the original data, rather than clustering near the line of equality. If we had drawn certain conclusions regarding the implications of varying aromatics content from an analysis of commercial fuels, we might apply them to the new data set in an effort to assess the "robustness" of their predictions.

Figure 7.2. Comparison of Fuel Property Correlations in Synthetic Data Set No. 2



7.2 ESTIMATING THE PROPERTIES OF ALTERNATIVE FUELS

An excellent example of the power of eigenfuel simulations can be found in the question of what diesel fuel properties might look like if one of the properties was subject to regulation. If, for example, the natural cetane rating of the fuel were subject to a minimum requirement of 50, then what might we expect the other properties to look like? The answer will depend on two considerations:

- How refiners would choose to modify the production process of diesel fuel to achieve 50 natural cetane
- How these choices would influence the other fuel properties.

We can give a very useful, if not definitive, answer to this question through a simulation. Let us consider the potential for increasing natural cetane through the set of processes that are represented in current commercial

fuels and the commercial eigenvectors. When synthetic fuels are generated according the variances given by the eigenvalues, we found that the data set was indistinguishable from the original, commercial fuels. Now, let us select only the synthetic fuels for which the simulated natural cetane rating is at least 50. We will assert that these synthetic fuels are likely to be representative of 50+ cetane diesel fuels that could be generated using current production processes and their frequency of use. The fraction of all generated fuels that have 50+ cetane will provide us qualitative guidance on how likely it is that current practices could generate the characteristics of interest.

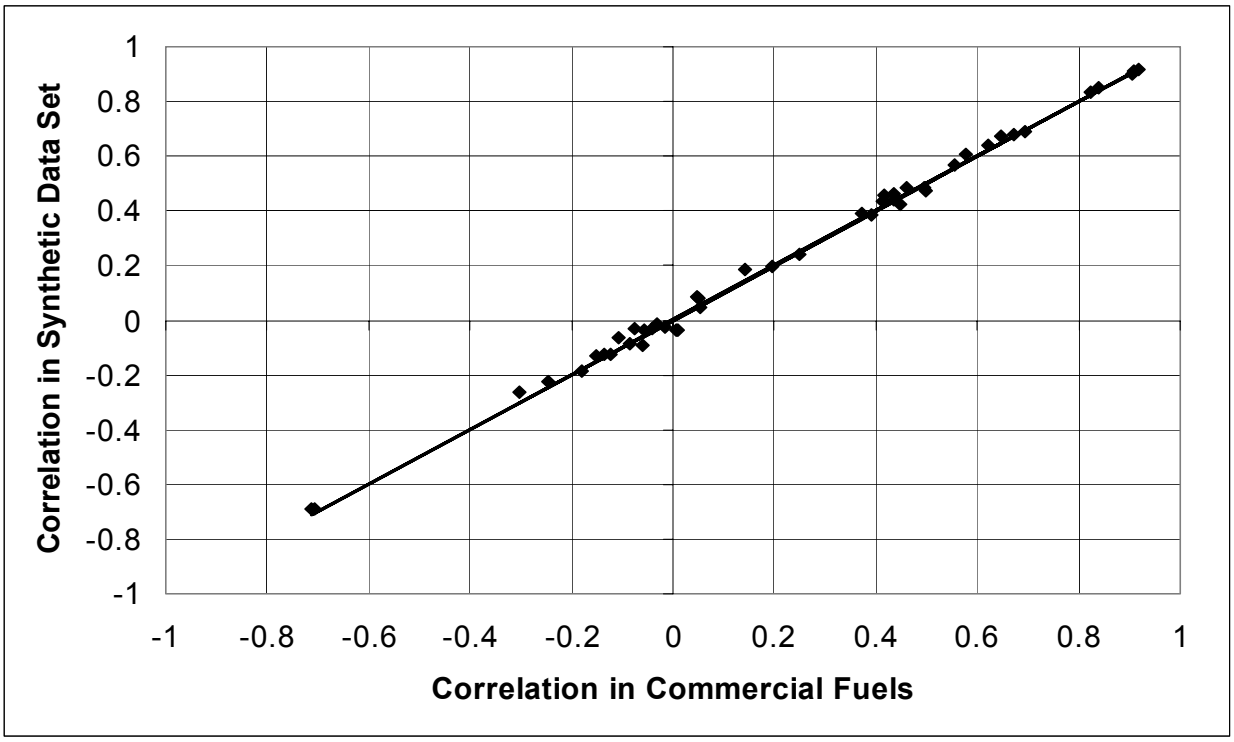
For the exercise as defined above, a simulation of 1000 fuels generated 45 fuels that had a natural cetane rating of 50 or higher. The 4.5 percent yield suggests that current production practices could produce high-cetane fuels, although with modified frequency of use, and it may be useful to examine new or modified practices in a follow-up study. The mean characteristics of the high-cetane fuels are given in Table 7.3. The average natural cetane rating is 51.3. This is accompanied by a reduction in total aromatics content to 25.2 percent from the 33.2 percent in the average commercial fuel, along with reductions in specific gravity, viscosity and the back end of the distillation curve. Other properties are affected relatively little. As shown in Figure 7.3, the high-cetane fuels also exhibit essentially the same correlations among fuel properties as do the commercial fuels overall. This serves to substantiate the assertion that high-cetane fuels generated in this exercise are ones that could, in fact, be realized in the real world.

Table 7.3. Statistical Summary of 50+ Cetane Data Set

Fuel Property	Units	Commercial Fuels		50+ Cetane Data Set	
		Mean	Std Dev	Mean	Std Dev
Natural Cetane	num	45.4	2.8	51.3	1.2
Specific Gravity	g/cm^3	0.850	0.0086	0.836	0.0068
Viscosity	mm^2/sec	2.67	0.30	2.52	0.34
Sulfur Content	ppm	352	60	364	83
Aromatics Content	vol percent	33.2	5.1	25.2	4.1
IBP	deg F	349	23	353	24
T10	deg F	429	18	422	19
T50	deg F	513	15	507	16
T90	deg F	607	15	599	17
FBP	deg F	653	17	642	19

Clearly, this exercise does not examine all ways in which high cetane fuels could be produced. For example, it may underestimate the potential that a "minor" eigenfuel characteristic of commercial fuels could be given much greater weight in achieving the objective. Further, it does not account for the potential that new processes – requiring new eigenvectors – could be brought to bear on the problem. The exercise also may be incomplete, in that it has not explicitly accounted for a cetane "safety margin" that would be needed by refiners to assure that all production batches meet minimum standards. Yet, even these objections could be dealt with through elaborations and extensions of the sampling process. The problem to be addressed here is one in which, fundamentally, a researcher must assess and project the effect of correlations among properties as the ways in which an objective can be reached are enumerated. This is a "vector" problem that eigenfuels can address well.

Figure 7.3. Comparison of Fuel Property Correlations in 50+ Cetane Data Set



8. EIGENFUELS IN FUELS REFORMULATION

While the foregoing sections have explained the potential roles of eigenfuels in emissions analysis, experiment design and fuels research, it may not be completely clear how fuels might be reformulated according to eigenfuel principles. This section begins with the question of what happens to fuels when their eigenfuel content is changed. It then outlines a potential eigenfuel-based system for reformulating fuels to achieve a predetermined objective, such as emissions reduction, cetane number increase, or another desired change in diesel fuel performance.

8.1 WHAT HAPPENS WHEN EIGENFUELS ARE CHANGED?

The formulation of diesel fuels in terms of weighted combinations of fuel properties has been demonstrated in this report and its predecessor. Yet, unless one understands the mechanisms by which fuel properties are transformed into eigenvector weights and vice versa, one might not fully understand the flexibility that makes the eigenvector approach so useful in fuel reformulation. The mechanisms, and the implications of their use, are discussed below. An extended discussion with examples can be found in Appendix F.

Let F be a fuel that is unsatisfactory from some standpoint, and let R be the same fuel after it has been modified (or reformulated) by a change in the value of one or more of the fuel properties. What effect does such a change have on the weights of the eigenvectors? This question can be answered by a general rule that applies to *all* fuel modifications, whether those changes are made by modifying individual fuel properties or the eigenvector characteristics: Simply express F and R in terms of the fuel properties and the eigenvectors of the system and compare.[15] If, for example, the aromatics content of a fuel is to be reduced, leaving all other fuel properties unchanged, one will see that the change ripples through all of the eigenvector weights. The magnitude of the change in weights will vary from one eigenvector to the next, being negligible in some, but substantial in others.

On the other hand, perhaps it is desired to change the eigenvector characteristics of a fuel by modifying the extent to which F expresses a particular vector. Express F in terms of the eigenvectors of the system, modify the weight of the vector in question to yield R, then return the expression of R to the space of fuel properties. If, for example, Vector 1 of the commercial diesel fuels is reduced, leaving all other eigenvector weights unchanged, one will see that the change ripples through to affect all of the individual fuel properties of R.

These two kinds of fuel modifications form a duality that will be referred to as participating and nonparticipating, respectively. The following sections consider the conditions under which these two types of modifications occur and the effect that those modifications have on emissions.

8.1.1 Non-Participating Fuel Modifications

The *nonparticipating* modification is one in which the weight of a specific eigenvector is changed in such a way that it does not induce changes in any of the other eigenvector weights. Let us take the average commercial diesel fuel and reduce its content of Vector 1 by one unit as shown in Table 8.1. Because the average fuel is described in eigenfuel terms as the vector of zeros, the modification can be described as the vector with weight -1 for Vector 1. When the eigenfuel expression of the modified fuel is returned to the

[15] See McAdams *et al.* 2001b, Section 2.3.1 and Appendix C.

space of fuel properties, we see that changes take place in all of the properties. The changes correspond to the fuel properties that make up the vector and are in proportion (in centered and standardized terms) to the weights of the properties making up the vector. In fact, the difference column exactly matches Vector 1 when put into physical units. If the weight of the vector had been changed by 0.5 instead of by 1.0, then the values in the difference column would be just half as great as those shown. Whatever the magnitude of the change made in the content of the vector, the quantities in the difference column will be proportional to the components of the vector.

Table 8.1. A Non-Participating Modification to the Average Commercial Fuel

Eigenvector	Modified Fuel	Fuel Property	Average Commercial Fuel	Modified Fuel	Property Difference
1	-1	Natural Cetane	45.4	44.8	-0.625
2	0	Specific Gravity	0.850	.853	0.003
3	0	Viscosity	2.67	2.79	0.120
4	0	Sulfur Content	352	350	-2.060
5	0	Aromatics Content	33.2	34.9	1.749
6	0	IBP	349	350	0.638
7	0	T10	429	435	6.093
8	0	T50	513	519	6.042
9	0	T90	607	613	5.597
10	0	FBP	653	658	5.365

From this example, it may appear necessary to modify the individual properties of a fuel in strict proportion to the components of an eigenvector in order to implement a reduction in the content of that vector to achieve a goal (such as reducing emissions). For example, if it were decided to control the content of a particular vector in diesel fuels, would we be implicitly requiring that fuel properties be changed in the particular ratios that make up that vector? This would in fact be the requirement *if and only if* it is mandated that there be *no change in the weights* of the other eigenvectors. This is an unnecessary restriction and one that would severely limit blending options. We will see in the next section that fuel modifications may produce finite changes in the weights of many eigenvectors. When, for example, only one (or a few) eigenvectors contribute to emissions, the changes in the weights of the other eigenvectors are irrelevant and can be largely ignored.

8.1.2 Participating Fuel Modifications

The preceding discussion shows that one must adhere to a specific regimen with regard to changes in fuel properties in order to change the weight of a single eigenvector, without also changing the weights of other eigenvectors. When there is no necessity to adhere to that regimen, changes in the weights of other eigenvectors may occur. We will refer to fuel modifications of this latter type as *participating*, in the sense that changes made in the weights of some eigenvectors may also result in changes in the weights of other eigenvectors.

To fully understand the interrelationships among changes in the weights of eigenvectors in a fuel, one must understand the mathematics by which changes in fuel properties are transformed into changes in eigenvectors and vice versa. The mathematics are presented in McAdams *et al.* (2001b, Appendix C) and are described tutorially in Appendix F of this report. A fuel F that has been modified into fuel R by changing one or more properties would be represented in eigenvector terms by inducing changes in the weights of all of the vectors. The more closely the pattern of changes in the properties matches a particular vector, the more the changes in eigenvector weights will be concentrated in that vector. Yet, any desired change in fuel properties can be represented by a corresponding change in which all of the eigenvectors participate.

As an example, let us consider a fuel in which only the natural cetane number is increased to 50, with no changes in its other properties. Although unrealistic, such a fuel could be considered hypothetically as complying with a requirement that the natural cetane rating be at least 50. As shown in Table 8.2, accomplishing this unilateral change involves modifications to the weights of every eigenvector. The largest changes made are to reduce the weights of Vector 2 by -0.890 and of Vector 6 by -1.145, but the weight of every vector is changed to at least some extent. It is in this sense that the fuel modification is said to be "participating."

Table 8.2. A Participating Modification to the Average Commercial Fuel

Fuel Property	Average Commercial Fuel	50 Cetane Fuel	Eigenvector Weights	
			Eigenvector	50 Cetane Fuel
Natural Cetane	45.4	50.0	1	0.369
Specific Gravity	0.850	0.850	2	-0.890
Viscosity	2.67	2.67	3	0.509
Sulfur Content	352	352	4	0.009
Aromatics Content	33.2	33.2	5	0.066
IBP	349	349	6	-1.145
T10	429	429	7	0.338
T50	513	513	8	-0.169
T90	607	607	9	0.235
FBP	653	653	10	-0.127

More realistically, a change desired in one property of diesel fuel will be accompanied by allowed changes in other properties, as was seen in Section 7 where the properties of a "realistic" fuel having natural cetane rating of at least 50 were estimated. If, instead of the isolated change to natural cetane shown above, one were to permit changes in all of the properties as estimated in the simulation, the eigenvector content of the fuel would be modified as shown in Table 8.3. We see that the changes are concentrated in Vectors 1, 2 and 3 – which involve the diesel blendstocks light-cycle oil, hydroprocessed heavy distillate, and straight-run heavy distillate – with only small changes in the other vectors. The average fuel resulting from the simulation process has modified contents of these vectors by amounts that vary from 1.1 to 2.6 standard deviations in the effort to achieve the 50+ cetane rating that was the basis for the simulation. This, also, is a participating modification. Because they allow more flexibility in blending, participating modifications are likely to have refinery economics advantages over non-participating modifications.

Table 8.3. A More Realistic Participating Modification to the Average Commercial Fuel

Fuel Property	Average Commercial Fuel	50 Cetane Fuel	Eigenvector Weights	
			Eigenvector	50 Cetane Fuel
Natural Cetane	45.4	51.3	1	2.559
Specific Gravity	0.850	0.836	2	-1.737
Viscosity	2.67	2.52	3	1.071
Sulfur Content	352	364	4	-0.018
Aromatics Content	33.2	25.2	5	0.108
IBP	349	353	6	-0.344
T10	429	422	7	0.064
T50	513	507	8	0.009
T90	607	599	9	-0.014
FBP	653	642	10	-0.016

8.2 A FRAMEWORK FOR FUEL REFORMULATION

The importance of the distinction between participating and nonparticipating is that most fuel modifications made in the real world will normally be of the participating kind. There is no need to specify a (nonparticipating) modification in which only the weight associated with a single eigenvector is changed. Fuel modifications will more typically occur by modifying the weights of several eigenvectors known to be associated with the objective of interest, as seen in the second example above, while also allowing the weights of other, unrelated vectors to change as needed to make the modification practical.

We can easily envision the outlines of a process for fuel reformation that is based on and guided by an eigenfuel assessment of fuels and their relationship to the objective. If, for example, our objective were to reduce NO_x emissions, we might do the following:

- Translate each candidate modification that might be made in fuel characteristics into its equivalent expression as eigenvectors.
- Use an accepted emissions model formulated in terms of the eigenvectors to "score" the candidate modifications in terms of predicted emissions impacts.
- Select and combine candidate changes until the targeted level of NO_x reduction achieved, while taking account of cost at each step in order to reach an optimal result.

During this process of modifying a fuel, the weights assigned to all of the eigenvectors would likely change. We need not target certain eigenvectors for change by themselves (a nonparticipating modification), but can consider all fuel modifications that are feasible in fuels manufacture.

In fact, the process would be no different in concept to one in which a single fuel property (or even several fuel properties) were used to predict the change in emissions as a fuel was modified in search of lower emissions. Therefore, it is similar to the process in which the Complex Model for RFG is currently used.

The fundamental difference is simply the use of eigenfuels, in place of an equation based on individual fuel properties, to "score" the effect of fuel changes. Eigenfuel-based models are likely to be better predictors than models based on correlated variables. A reformulation process based on eigenfuels should be less subject to the inaccuracies and inefficiencies than would occur in a system driven by less reliable predictions.

Provided that the objective can be stated quantitatively as a function of the eigenfuel description of fuels, eigenfuels can provide the framework for a system of fuel reformulation that is applicable to many areas. The objective could be a regulatory application such as the reduction of engine emissions, or it could be a commercial application such as improving the prediction of cetane rating for blending purposes.

Important characteristics and implications of such a framework are:

- Eigenfuel-based blending offers wide flexibility to refiners to achieve a predefined objective at lowest cost. The degree of flexibility that exists is determined by the number and characteristics of available blendstocks and processing options, and the use of eigenfuels to guide the process in no way limits or constrains use of that flexibility
- Eigenfuels are more likely to give accurate and unbiased guidance on the (vector) characteristics of fuels that must be changed in order to reach the defined objective. Inaccurate or misleading guidance that could result when aliasing obscures the fuel characteristics of interest cannot occur when using eigenfuel-based models
- The improved guidance and ability to predict the effect of fuel changes mean that the reformulation process is more likely to achieve its expected results in real-world use, with less "giveaway" of valuable fuel properties to assure compliance, and at lower total cost than a system in which guidance and predictive ability are poorer.

9. REFERENCES

Crawford, R.W. and McAdams, H.T. 2001. *Issues in Development of Diesel Fuel Emissions Models*. Presented at EPA Diesel Fuel Effects on Emissions Workshop, National Vehicle and Fuel Emissions Laboratory, Ann Arbor, MI.

Hadi, A.S. and Ling, R.F. 1998. *Some Cautionary Notes on the Use of Principal Components Regression*. The American Statistician, Vol. 52, No. 1. February.

Jackson, J.E. 1991. *A User's Guide to Principal Components*. John Wiley & Sons.

Lee, R., J. Pedley and C. Hobbs. 1998. *Fuel Quality Impact on Heavy Duty Diesel Emissions - A Literature Review*. SAE 982649. August 28.

McAdams, H.T. 1995. *A Random Balance Procedure for Simplifying a Complex Model*. American Statistical Association. Proceedings of the Section on Statistics and the Environment, Alexandria, VA.

McAdams, H.T., R.W. Crawford and G.R. Hadder. 2000a. *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*. SAE 2000-01-1961.

McAdams, H.T., R.W. Crawford and G.R. Hadder. 2000b. *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*. ORNL/TM-2000/5.

Southwest Research Institute. 2001. *Diesel Fuel Impact Model Data Analysis Plan Review*. SwRI 08.04075.

Tanaka, S., M. Morinaga, H. Yoshida, H. Takizawa, K. Sanse and H. Ikebe. 1996. *Effects of Fuel Properties on Exhaust Emissions from DI Diesel Engines*. SAE 962114.

U.S. Department of Energy. 1994. *Estimating the Costs and Effects of Reformulated Gasoline*. DOE/PO-0030. December.

U.S. Environmental Protection Agency. 1999. EPA HDEWG Program: Phase II. *Briefing for Meeting of the Mobile Sources Technical Review Subcommittee*, Clean Air Act Advisory Committee, Washington, D.C., January 13.

U.S. Environmental Protection Agency. 2001. *Strategies and Issues in Correlating Diesel Fuel Properties with Emissions: Staff Discussion Document*. EPA420-P-01-001. July.

APPENDIX A

SUPPORTING DATA

Table A.1. PCA Decomposition of EPA Experimental Fuels

PCA Eigen variables

	1	2	3	4	5	6	7	8	9	10	11	12
NatCetane	0.366	-0.327	0.202	-0.032	0.142	-0.085	0.266	-0.264	0.181	-0.069	-0.104	-0.707
NatCet ^2	0.368	-0.318	0.198	-0.034	0.151	-0.102	0.322	-0.221	0.177	-0.135	0.020	0.698
CetDiff	-0.218	0.192	0.618	0.015	0.156	-0.027	0.034	-0.065	0.228	0.656	0.145	0.019
CetDif ^2	-0.196	0.194	0.628	0.053	0.178	-0.025	-0.046	0.140	-0.225	-0.638	-0.116	-0.024
TotlArom	-0.462	-0.070	-0.127	-0.038	0.035	-0.207	0.371	-0.005	0.078	0.090	-0.752	0.054
TArom ^2	-0.437	-0.090	-0.149	-0.021	0.046	-0.196	0.570	0.077	-0.016	-0.147	0.614	-0.088
Density	-0.437	-0.159	-0.024	-0.015	-0.093	0.217	-0.328	-0.649	0.368	-0.237	0.078	0.021
Oxygen	0.001	0.034	0.081	-0.948	-0.138	0.221	0.060	0.127	0.074	-0.037	-0.005	-0.005
T10	-0.102	-0.447	0.128	0.250	-0.158	0.585	0.080	0.506	0.282	-0.012	-0.034	-0.004
T50	-0.147	-0.513	0.152	-0.066	-0.084	0.115	-0.049	-0.213	-0.754	0.225	0.008	0.022
T90	-0.097	-0.442	0.083	-0.118	-0.089	-0.650	-0.449	0.317	0.195	-0.010	0.062	-0.001
Sulfur	-0.088	-0.139	-0.227	-0.119	0.911	0.161	-0.194	0.110	-0.007	0.038	0.017	0.001
Eigenvalues	4.016	3.011	1.691	1.042	0.936	0.565	0.413	0.144	0.094	0.053	0.029	0.004
Pct Variance	33.468	25.093	14.094	8.683	7.802	4.710	3.443	1.202	0.787	0.443	0.238	0.036
Cumulative Pct	33.468	58.561	72.655	81.338	89.140	93.851	97.294	98.496	99.283	99.726	99.964	100.000

PCA Eigen variables in physical units: Deltas

	1	2	3	4	5	6	7	8	9	10	11	12
NatCetane	2.229	-1.992	1.227	-0.194	0.863	-0.520	1.619	-1.606	1.102	-0.419	-0.633	-4.304
NatCet ^2	225.846	-195.266	121.409	-20.861	92.932	-62.672	197.604	-135.473	108.781	-82.729	12.060	428.154
CetDiff	-0.977	0.862	2.773	0.069	0.699	-0.122	0.153	-0.292	1.021	2.943	0.652	0.083
CetDif ^2	-11.021	10.896	35.254	2.962	10.005	-1.407	-2.556	7.844	-12.595	-35.814	-6.504	-1.323
TotlArom	-4.533	-0.690	-1.250	-0.370	0.340	-2.028	3.643	-0.050	0.761	0.880	-7.380	0.530
TArom ^2	-240.792	-49.396	-82.043	-11.470	25.218	-107.734	314.080	42.474	-8.881	-81.031	338.468	-48.502
Density	-0.007	-0.003	-0.000	-0.000	-0.002	0.004	-0.005	-0.011	0.006	-0.004	0.001	0.000
Oxygen	0.000	0.014	0.033	-0.380	-0.055	0.089	0.024	0.051	0.030	-0.015	-0.002	-0.002
T10	-4.138	-18.047	5.177	10.086	-6.367	23.652	3.232	20.454	11.385	-0.493	-1.385	-0.149
T50	-4.951	-17.334	5.140	-2.240	-2.829	3.879	-1.662	-7.202	-25.486	7.609	0.269	0.747
T90	-3.279	-15.006	2.823	-4.012	-3.034	-22.048	-15.236	10.773	6.612	-0.350	2.104	-0.046
Sulfur	-45.855	-72.041	-117.705	-62.005	473.350	83.792	-100.750	57.054	-3.784	19.592	8.596	0.268
Eigenvalues	4.016	3.011	1.691	1.042	0.936	0.565	0.413	0.144	0.094	0.053	0.029	0.004
Pct Variance	33.468	25.093	14.094	8.683	7.802	4.710	3.443	1.202	0.787	0.443	0.238	0.036
Cumulative Pct	33.468	58.561	72.655	81.338	89.140	93.851	97.294	98.496	99.283	99.726	99.964	100.000

Table A.2. PCR+ Regression Model for NO_x Emissions

Stage 1: Regression against engine dummy variables

Mean	2193.5561	1.0000	2193.5561
Model	12.0502	40.0000	0.3013
Error	1.6939	865.0000	0.0020
Total	2207.3001	906.0000	2.4363

F Value: 153.84
R-square: 0.8768

Stage 2(a): Regression against PCA coefficients (eigenfuels)

	SS	DF	Mean Sq
Mean	0.0000	1.0000	0.0000
Model	1.0151	12.0000	0.0846
Error	0.6788	893.0000	0.0008
Total	1.6939	906.0000	0.0019

F Value: 111.28
R-square: 0.5993

Predictive correlation to dependent variable: 0.7741

	Estimate	Std Err	t-value
Intercept	0.00000	0.00092	0.00000
EigFuel 1	-0.01325	0.00046	28.97650
EigFuel 2	-0.00384	0.00053	7.27775
EigFuel 3	-0.01230	0.00070	17.45401
EigFuel 4	-0.00087	0.00090	0.96731
EigFuel 5	-0.00693	0.00095	7.32028
EigFuel 6	0.00281	0.00122	2.30501
EigFuel 7	-0.00096	0.00143	0.67038
EigFuel 8	0.00458	0.00241	1.89964
EigFuel 9	0.01112	0.00298	3.72900
EigFuel 10	-0.01368	0.00397	3.44168
EigFuel 11	-0.03749	0.00543	6.90857
EigFuel 12	-0.01198	0.01389	0.86306

SS contributions by eigenvector

Eigenvector	Reg Coeff	Model SS	Pct SS	F Ratio
1	-0.0133	0.6382	62.88	839.64*
3	-0.0123	0.2316	22.81	304.64*
5	-0.0069	0.0407	4.01	53.59*
2	-0.0038	0.0403	3.97	52.97*
11	-0.0375	0.0363	3.57	47.73*
9	0.0111	0.0106	1.04	13.91*
10	-0.0137	0.0090	0.89	11.85*
6	0.0028	0.0040	0.40	5.31*
8	0.0046	0.0027	0.27	3.61
4	-0.0009	0.0007	0.07	0.94
12	-0.0120	0.0006	0.06	0.74
7	-0.0010	0.0003	0.03	0.45

Table A.2. (Continued)

Stage 2(b): Reduced Form regression against Selected PCA coefficients (eigenfuels)
Selection vector: 1 2 3 5 6 9 10 11

	SS	DF	Mean Sq
Mean	0.0000	1.0000	0.0000
Model	1.0107	8.0000	0.1263
Error	0.6832	897.0000	0.0008
Total	1.6939	906.0000	0.0019

F Value: 165.88
R-square: 0.5967

Predictive correlation to dependent variable: 0.7725

	Estimate	Std Err	t-value
Intercept	0.00000	0.00092	0.00000
EigFuel 1	-0.01325	0.00046	28.94845
EigFuel 2	-0.00384	0.00053	7.27071
EigFuel 3	-0.01230	0.00071	17.43712
EigFuel 5	-0.00693	0.00095	7.31320
EigFuel 6	0.00281	0.00122	2.30278
EigFuel 9	0.01112	0.00299	3.72539
EigFuel 10	-0.01368	0.00398	3.43835
EigFuel 11	-0.03749	0.00543	6.90188

SS contributions by fuel property: Simplify.m algorithm.
(regression coefficients in physical units)

Fuel Property	Reg Coeff	Model SS	Pct SS	F Ratio
1	-0.0001	0.0980	9.70	128.73*
2	-0.0000	0.1124	11.12	147.56*
3	-0.0041	0.0685	6.77	89.89*
4	0.0001	0.0264	2.61	34.65*
5	0.0036	0.3215	31.81	422.09*
6	-0.0000	0.0976	9.66	128.17*
7	0.7520	0.2217	21.93	291.08*
8	0.0049	0.0001	0.01	0.07
9	0.0002	0.0464	4.59	60.97*
10	-0.0003	0.0050	0.49	6.53*
11	0.0000	0.0125	1.24	16.42*
12	-0.0000	0.0007	0.07	0.90
13	99.9999	1.0107	100.00	1327.05*

Table A.3. PCR+ Regression Model For PM Emissions

Stage 1: Regression against engine dummy variables

Mean	4266.7101	1.0000	4266.7101
Model	232.8442	40.0000	5.8211
Error	10.2568	865.0000	0.0119
Total	4509.8111	906.0000	4.9777

F Value: 490.92
R-square: 0.9578

Stage 2(a): Regression against PCA coefficients (eigenfuels)

	SS	DF	Mean Sq
Mean	0.0000	1.0000	0.0000
Model	3.6466	12.0000	0.3039
Error	6.6102	893.0000	0.0074
Total	10.2568	906.0000	0.0113

F Value: 41.05
R-square: 0.3555

Predictive correlation to dependent variable: 0.5963

	Estimate	Std Err	t-value
Intercept	-0.00000	0.00286	0.00000
EigFuel 1	-0.02298	0.00143	16.10069
EigFuel 2	-0.01191	0.00165	7.22482
EigFuel 3	-0.01904	0.00220	8.65896
EigFuel 4	0.01547	0.00280	5.52210
EigFuel 5	0.01134	0.00296	3.83740
EigFuel 6	-0.00473	0.00380	1.24216
EigFuel 7	-0.00933	0.00445	2.09752
EigFuel 8	0.03149	0.00753	4.18202
EigFuel 9	-0.01235	0.00931	1.32647
EigFuel 10	0.01433	0.01240	1.15557
EigFuel 11	-0.09699	0.01693	5.72800
EigFuel 12	0.05597	0.04333	1.29170

SS contributions by eigenvector

Eigenvector	Reg Coeff	Model SS	Pct SS	F Ratio
1	-0.0230	1.9189	52.62	259.23*
3	-0.0190	0.5550	15.22	74.98*
2	-0.0119	0.3864	10.60	52.20*
11	-0.0970	0.2429	6.66	32.81*
4	0.0155	0.2257	6.19	30.49*
8	0.0315	0.1295	3.55	17.49*
5	0.0113	0.1090	2.99	14.73*
7	-0.0093	0.0326	0.89	4.40*
9	-0.0123	0.0130	0.36	1.76
12	0.0560	0.0124	0.34	1.67
6	-0.0047	0.0114	0.31	1.54
10	0.0143	0.0099	0.27	1.34

Table A.3. (Continued)

Stage 2(b): Reduced Form regression against Selected PCA coefficients (eigenfuels)
Selection vector: 1 2 3 4 5 7 8 11

	SS	DF	Mean Sq
Mean	0.0000	1.0000	0.0000
Model	3.5999	8.0000	0.4500
Error	6.6569	897.0000	0.0074
Total	10.2568	906.0000	0.0113

F Value: 60.63
R-square: 0.3510

Predictive correlation to dependent variable: 0.5924

	Estimate	Std Err	t-value
Intercept	-0.00000	0.00286	0.00000
EigFuel 1	-0.02298	0.00143	16.08003
EigFuel 2	-0.01191	0.00165	7.21555
EigFuel 3	-0.01904	0.00220	8.64785
EigFuel 4	0.01547	0.00281	5.51502
EigFuel 5	0.01134	0.00296	3.83248
EigFuel 7	-0.00933	0.00446	2.09483
EigFuel 8	0.03149	0.00754	4.17665
EigFuel 11	-0.09699	0.01695	5.72065

SS contributions by fuel property: Simplify.m algorithm.
(egression coefficients in physical units)

Fuel Property	Reg Coeff	Model SS	Pct SS	F Ratio
1	-0.0013	0.2645	7.35	35.64*
2	-0.0000	0.3282	9.12	44.23*
3	-0.0052	0.1028	2.86	13.85*
4	0.0002	0.0137	0.38	1.85
5	0.0085	1.1529	32.02	155.35*
6	-0.0001	0.1595	4.43	21.50*
7	-0.8377	0.2228	6.19	30.02*
8	-0.0355	0.2934	8.15	39.53*
9	0.0006	0.3310	9.20	44.60*
10	-0.0001	0.0771	2.14	10.38*
11	0.0003	0.1812	5.03	24.42*
12	0.0000	0.4729	13.14	63.72*
13	99.9999	3.5999	100.00	485.08*

Table A.4. PCA Decomposition of Commercial Diesel Fuels

PCA Eigen variables

	1	2	3	4	5	6	7	8	9	10
Cetane Num	0.223	-0.541	0.324	-0.001	0.034	-0.696	0.189	-0.090	0.133	-0.073
Density	-0.405	0.228	-0.226	0.118	0.046	-0.200	0.180	0.028	0.791	-0.136
Viscosity	-0.394	-0.287	0.093	0.088	0.187	0.217	-0.167	-0.784	-0.022	-0.147
Sulfur ppm	0.034	0.152	0.419	0.828	-0.326	0.081	0.043	0.007	-0.002	-0.009
Totl Arom	-0.344	0.298	-0.343	0.142	-0.129	-0.596	-0.030	-0.136	-0.513	-0.002
IBP	-0.028	-0.465	-0.411	0.008	-0.736	0.083	-0.218	0.017	0.122	0.048
T10	-0.345	-0.400	-0.167	0.173	0.153	0.226	0.647	0.306	-0.261	-0.094
T50	-0.394	-0.251	0.127	0.149	0.289	-0.090	-0.498	0.374	0.034	0.513
T90	-0.370	0.017	0.407	-0.271	-0.207	-0.009	-0.240	0.320	-0.095	-0.642
FBP	-0.317	0.146	0.405	-0.383	-0.383	0.020	0.361	-0.149	0.009	0.518
Eigenvalues	4.783	1.740	1.341	0.936	0.712	0.216	0.109	0.095	0.048	0.020
Pct Variance	47.834	17.401	13.410	9.359	7.122	2.157	1.088	0.951	0.478	0.199
Cumulative Pct	47.834	65.236	78.645	88.005	95.127	97.283	98.371	99.323	99.801	100.000

PCA Eigen variables in physical units: Deltas

	1	2	3	4	5	6	7	8	9	10
Cetane Num	0.625	-1.515	0.906	-0.003	0.095	-1.948	0.528	-0.253	0.372	-0.204
Density	-0.003	0.002	-0.002	0.001	0.000	-0.002	0.002	0.000	0.007	-0.001
Viscosity	-0.120	-0.087	0.028	0.027	0.057	0.066	-0.051	-0.239	-0.007	-0.045
Sulfur ppm	2.060	9.139	25.121	49.655	-19.561	4.884	2.563	0.421	-0.095	-0.512
Totl Arom	-1.749	1.516	-1.743	0.723	-0.657	-3.032	-0.151	-0.691	-2.610	-0.012
IBP	-0.638	-10.629	-9.382	0.183	-16.831	1.892	-4.980	0.390	2.792	1.094
T10	-6.093	-7.068	-2.946	3.053	2.695	3.986	11.414	5.393	-4.599	-1.659
T50	-6.042	-3.844	1.941	2.287	4.437	-1.375	-7.643	5.739	0.518	7.872
T90	-5.597	0.253	6.168	-4.111	-3.130	-0.140	-3.632	4.844	-1.435	-9.723
FBP	-5.365	2.466	6.867	-6.498	-6.487	0.345	6.116	-2.530	0.150	8.771
Eigenvalues	4.783	1.740	1.341	0.936	0.712	0.216	0.109	0.095	0.048	0.020
Pct Variance	47.834	17.401	13.410	9.359	7.122	2.157	1.088	0.951	0.478	0.199
Cumulative Pct	47.834	65.236	78.645	88.005	95.127	97.283	98.371	99.323	99.801	100.000

Table A.5. PCA Decomposition of Clear EPA Experimental Fuels

PCA Eigen variables

	1	2	3	4	5	6	7	8	9	10
NatCetane	0.129	-0.593	0.151	-0.295	0.049	-0.450	0.152	-0.143	0.435	-0.291
Density	0.304	0.466	-0.063	0.096	0.234	0.241	0.136	-0.502	0.502	-0.207
Viscosity	0.428	-0.038	-0.129	-0.117	0.251	0.230	-0.007	0.415	-0.293	-0.641
Sulfur	0.074	0.242	0.586	-0.683	-0.253	0.240	-0.023	-0.008	-0.047	0.045
Aromatics	0.192	0.522	0.232	0.201	-0.147	-0.700	0.059	0.257	-0.015	-0.122
IBP	0.306	0.001	-0.513	-0.173	-0.684	-0.091	-0.049	-0.309	-0.183	-0.079
T10	0.403	0.031	-0.305	-0.240	0.104	-0.003	-0.148	0.460	0.405	0.527
T50	0.417	-0.077	0.079	-0.094	0.440	-0.218	0.124	-0.381	-0.518	0.367
T90	0.344	-0.207	0.353	0.365	-0.124	0.084	-0.736	-0.115	0.057	-0.007
FBP	0.342	-0.223	0.269	0.392	-0.326	0.276	0.611	0.153	0.034	0.164
Eigenvalues	4.912	2.136	1.200	0.849	0.402	0.250	0.101	0.079	0.048	0.025
Pct Variance	49.117	21.358	11.996	8.486	4.019	2.502	1.007	0.791	0.477	0.248
Cumulative Pct	49.117	70.475	82.471	90.957	94.975	97.477	98.484	99.275	99.752	100.000

PCA Eigen variables in physical units: Deltas

	1	2	3	4	5	6	7	8	9	10
NatCetane	0.774	-3.553	0.908	-1.768	0.293	-2.696	0.913	-0.857	2.610	-1.742
Density	0.005	0.008	-0.001	0.002	0.004	0.004	0.002	-0.009	0.009	-0.004
Viscosity	0.291	-0.026	-0.088	-0.079	0.171	0.156	-0.005	0.283	-0.199	-0.436
Sulfur	52.696	171.455	415.192	-483.558	-179.185	170.309	-16.500	-5.611	-33.054	31.662
Aromatics	1.731	4.692	2.087	1.810	-1.320	-6.294	0.531	2.308	-0.138	-1.097
IBP	14.428	0.038	-24.170	-8.142	-32.245	-4.310	-2.332	-14.589	-8.619	-3.707
T10	15.839	1.209	-11.975	-9.425	4.087	-0.108	-5.819	18.041	15.911	20.675
T50	15.424	-2.851	2.924	-3.489	16.288	-8.065	4.593	-14.113	-19.184	13.564
T90	14.108	-8.482	14.486	14.981	-5.093	3.463	-30.160	-4.696	2.352	-0.268
FBP	13.400	-8.758	10.529	15.381	-12.776	10.819	23.943	5.991	1.346	6.413
Eigenvalues	4.912	2.136	1.200	0.849	0.402	0.250	0.101	0.079	0.048	0.025
Pct Variance	49.117	21.358	11.996	8.486	4.019	2.502	1.007	0.791	0.477	0.248
Cumulative Pct	49.117	70.475	82.471	90.957	94.975	97.477	98.484	99.275	99.752	100.000

APPENDIX B

THE MULTIPLICITY OF MULTIPLE REGRESSION

APPENDIX B. THE MULTIPLICITY OF MULTIPLE REGRESSION

B.1 INTRODUCTION

Regression of diesel emissions on fuel variables is a classic example of the widespread use of multiple regression to express a response variable as a function of predictor variables. Though the primary emphasis is on prediction, regression analysis usually involves another very important process, that of selecting the variables to be included in the regression equation. This feature of regression is so commonplace that its implications are easily overlooked, especially in an almost "automatic" environment in which terms in the equation are removed if their coefficients fail to attain the ubiquitous p=0.05 significance level.

The theory that forms the basis for computations and demonstrations in this paper is based on two previous publications: SAE Technical Paper 2000-01-1961 [Ref 1] and ORNL Report ORNL/TM-2000/5 [Ref 2]. The data used in this appendix is part of the database recently analyzed by EPA for purposes of developing a model for diesel fuel effects on emissions. Variables and their nomenclature are given in Addendum I.

Two aspects of traditional regression procedures are of concern. First, tests of significance apply only to variables *included* in the regression equation. Variables are dropped from the initial list if they fail to satisfy the p=0.05 significance criterion. There seems to be no equivalent paradigm for *adding* a variable not included in the original list of candidate predictors.[16] Secondly, there is a multiplicity of ways in which the equation can be simplified by removing one or more of the candidate variables that make up the original slate of predictors. Often a large number of these selections satisfy the significance criterion and yield essentially the same performance as judged by the classic R-Square criterion.

Mathematically, a set of N predictors gives rise to $2^N - 1$ subsets of those predictors. Each subset can be used as the basis of a prediction equation, and the problem posed to the statistical analyst is to find the "best subset model." As noted above, however, the "best" solution may be little better than a number of "next best" solutions.

This paper examines all subset models for NO_x for the EPA dataset for Tech Group T. The $2^{12} - 1 = 4095$ "all possible regressions" include all equations in which the 12 variables occur singly, then as all possible pairs of variables, and so on up to the all-inclusive case in which all 12 predictor variables are in the model. The principal thrust is to reveal the anomalies that can occur when the predictor variables are substantially correlated, as they are in the subject data set and as they usually are unless the data are collected according to an orthogonal experiment design.

In particular, for correlated variables, it will be shown that:

[16] It may be argued that stepwise regression provides a procedure by which terms can be entered into a model as well as removed from the model. The stepwise algorithm does not really qualify, however, because a set of variables is specified at the outset, and variables to be entered into the model are drawn from this set. Our concern here is that the stepwise process carries with it the implication that there is an *excess* of predictors, not a *deficiency*. In short, the variable-selection process is one of "culling" rather than augmentation. The distinction is admittedly a fine one, but one that has philosophical merit.

1. A relatively large number of the "all possible regressions" perform almost equally well, even though different predictor variables are included in the subsets and even though all variables, in each case, are claimed to be "significant."
2. A given predictor variable can exhibit widely different t-ratios and corresponding "p" (probability) values, depending on "the company it keeps" in a particular subset. This anomaly raises the question of what is the "true" significance of that variable.
3. Variables going by the same name in the various subsets are not really the "same" variables. In any subset, any given variable has "aliases" that depend on the variables included in and excluded from that subset.
4. Though stepwise regression provides means, either automatic or interactive, for variable selection, it does not remedy the above concerns.

By redefining the predictor variables as linearly independent *eigenvectors*, the method we refer to as PCR+ (Principal Components Regression Plus) eliminates the objections noted above for stepwise regression. The solutions obtained are subject to less multiplicity and are free of aliasing. The t-ratios and corresponding probabilities are stable and unaffected by other variables included in a subset model. In subsequent discussion, stepwise regressions are referred to as regressions in *P-Space* (P for "property"). Similarly, PCR+ regressions are referred to as regressions in *E-Space* (E for "eigenvector").

The position is taken that every data set has its own orthogonal basis and that analysis in the corresponding vector space is both unique and the *only* solution that deals with multicollinearity in a mathematically rigorous way. However, it is recognized and demonstrated in a later section that the multiple stepwise solutions correlate highly with each other and with the eigenvector solution so far as *predicted response* is concerned. On the other hand, a particular stepwise solution may provide little guidance in identifying the *factors* that influence emissions, inasmuch as many different slates of variables can yield the same predictive performance.

B.2 SIGNIFICANCE AMBIGUITY

The data set under consideration consists of 480 NO_x test results for engines comprising Tech Group T in the EPA database. To eliminate the effects of engines on emissions, the mean logarithmic emission level for each engine is subtracted from log emissions for every test performed with that engine. The resulting "residuals," referred to subsequently as "LogRes," are construed as expressing the net effect of fuel variables on NO_x emissions with zero intercept.

Table B. 1 gives the regression coefficients, their standard errors and associated t-ratios when all P-Space predictor variables are included in the subset. Note that seven of the fuel variables are indicated to have significance at p=0.05 or better. Accordingly, these seven variables were retained and the engine-corrected residuals, as defined above, were regressed on those seven fuel variables. Results for the second regression are shown in Table B.2.

TABLE B.1. REGRESSION OF LOGRES ON FUEL VARIABLES

R-Square = 0.6361

	Regression Coefficient	Coefficient Std Error	t-Ratio
NatCet	-0.0077	0.0134	0.5779
NatCet^2	-0.0042	0.0135	0.3110
CetDif	-0.0289	0.0047	6.1491*
CetDif^2	0.0127	0.0046	2.7813*
Arom	0.0324	0.0062	5.2428*
Arom^2	-0.0098	0.0054	1.8082
Spc Grv	0.0106	0.0033	3.1955*
Oxygen	0.0053	0.0014	3.6766*
T10	0.0104	0.0023	4.5187*
T50	-0.0104	0.0034	3.0581*
T90	0.0021	0.0022	0.9460
Sulfur	-0.0021	0.0015	1.4289

*Significant at p=0.05 or better

TABLE B.2. RECOMPUTED LOGRES REGRESSION IN P-SPACE

R-Square = 0.6131

	Regression Coefficient	Coefficient Std Error	t-Ratio
CetDif	-0.0273	0.0047	5.7550*
CetDif^2	0.0122	0.0046	2.6244*
Arom	0.0248	0.0023	10.6099*
Sp Grav	0.0203	0.0024	8.6138*
Oxygen	0.0055	0.0015	3.7416*
T10	0.0103	0.0021	4.8188*
T50	-0.0173	0.0023	7.5659*

*Significant at p=0.05 or better

The t-ratios for those variables considered significant are compared below (Table B.3), before and after simplification and recomputation. Note that the t-ratios are different in the two regressions, in some instances dramatically so.

TABLE B.3. COMPARISON OF T-RATIOS BEFORE AND AFTER SIMPLIFICATION

	Before Recomputation	After Recomputation
CetDif	6.1491	5.7550
$CetDif^2$	2.7813	2.6244
Arom	5.2428	10.6099
Sp Grav	3.1955	8.6138
Oxygen	3.6766	3.7416
T10	4.5187	4.8188
T50	3.0581	7.5659

This occurrence, of course, is no surprise to those accustomed to performing regression analyses. It is evident that significance is transferred from one fuel property to another, via a mechanism we refer to as *aliasing*, about which more will be said later in this report. Our object is to explore these apparent ambiguities in computed significance levels by computing R-squares and t-ratios for all possible subsets of the variables.

Let us now perform the LogRes regression in E-Space (See Table B.4).

TABLE B.4. REGRESSION OF LOGRES ON EIGENVECTORS

R-Square = 0.6361

Eigenvector	Regression Coefficient	Coefficient Std Error	t-Ratio
1	-0.0150	0.0007	22.6119*
2	-0.0007	0.0008	0.8607
3	-0.0152	0.0010	14.9292*
4	-0.0013	0.0013	0.9901
5	-0.0072	0.0014	5.1735*
6	0.0037	0.0017	2.2364*
7	0.0029	0.0021	1.3674
8	0.0063	0.0034	1.8693
9	0.0138	0.0044	3.1707*
10	0.0237	0.0064	3.6843*
11	-0.0352	0.0079	4.4290*
12	0.0049	0.0190	0.2596

*Significant at p=0.05 or better

Seven eigenvectors, numbers 1, 3, 5, 6, 9, 10 and 11 are significant at the 0.05 significance level. As in the case for stepwise regression, we recompute the regression coefficients with just those seven vectors retained (see Table B.5).

TABLE B.5. RECOMPUTED LOGRES REGRESSION IN E-SPACE

R-Square = 0.6305

Eigenvector	Regression Coefficient	Coefficient Std Error	t-Ratio
1	-0.0150	0.0007	22.5605*
3	-0.0152	0.0010	14.8953*
5	-0.0072	0.0014	5.1618*
6	0.0037	0.0017	2.2314*
9	0.0138	0.0044	3.1635*
10	0.0237	0.0064	3.6759*
11	-0.0352	0.0080	4.4190*

*Significant at p=0.05 or better

Note that the regression coefficients are identical in the two regressions, and, as shown below, there are only minute differences in the t-ratios whether all eigenvectors are used as predictors or only those cited as being significant at the 0.05 level. The stability of the t-ratios computed before and after simplification is evident in Table B.6.

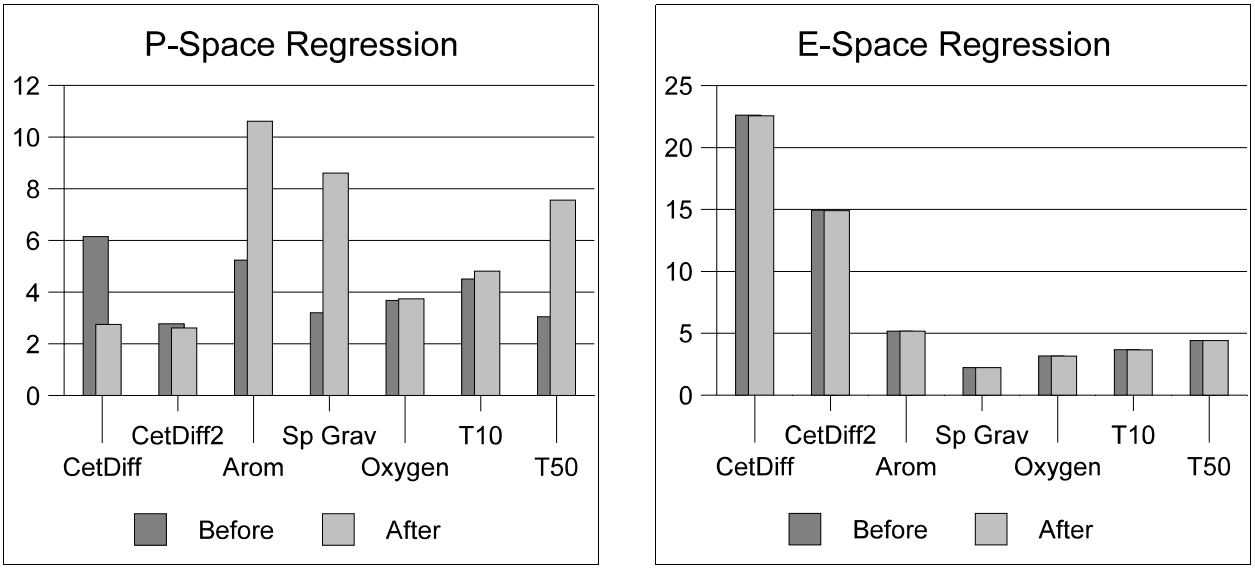
TABLE B.6. COMPARISON OF T-RATIOS BEFORE AND AFTER SIMPLIFICATION

Eigenvector	Before Recomputation	After Recomputation
1	22.6119	22.5605
3	14.9292	14.8953
5	5.1735	5.1618
6	2.2364	2.2314
9	3.1797	3.1635
10	3.6843	3.6759
11	4.4290	4.4190

The t-ratios for the 7-variable regression are all slightly less than those for the regression using all 12 variables. The reason is that, with five variables removed, the error SS is slightly higher than when all twelve eigenvectors are admitted to the model. Otherwise, the significance levels are unchanged and are certainly more stable than in the case of stepwise regression. This fact gives us assurance that the significance levels computed are "real" and can be trusted as representing the true significance of the various regression coefficients.

Figure B.1 displays graphically the difference in the behavior of stepwise and PCR+ regressions before and after the dropping of non-significant terms and recomputing the regression equation.

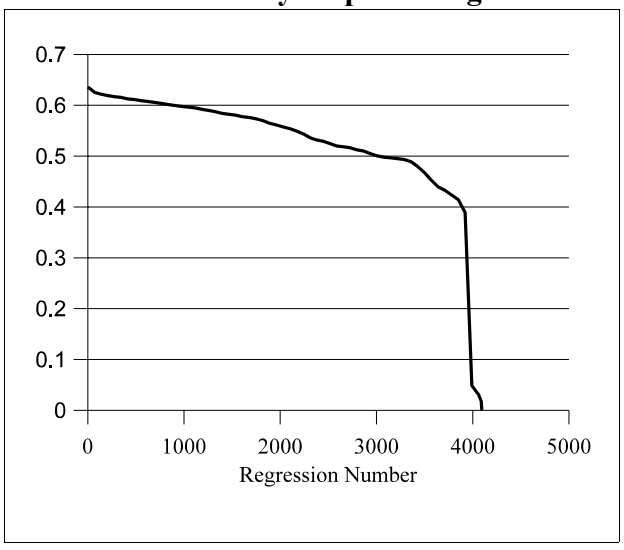
Figure B.1. Significance of Terms Before and After Recomputation



B.3 "ALL POSSIBLE" REGRESSIONS

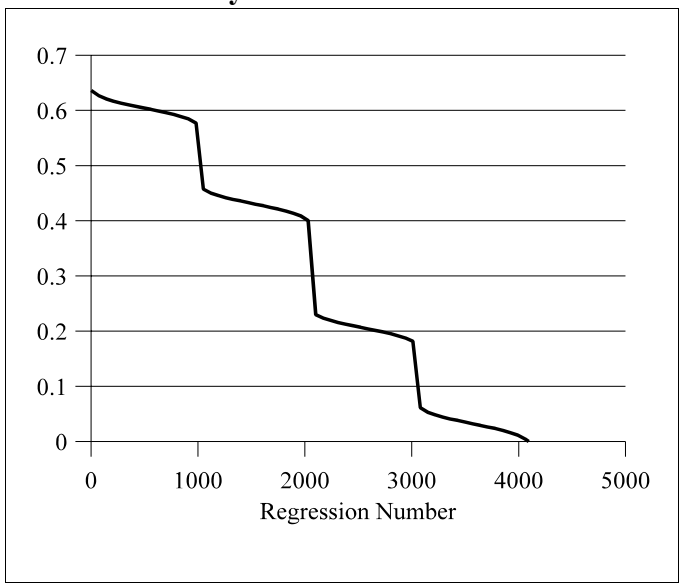
Though the notion of R-Square is subject to criticism as a measure of the efficacy of a regression model, it serves well to rank the 4095 "all possible regressions" according to how well any given model conforms to the data. Accordingly, every one of the 4095 regressions was performed and the resulting R-Squares stored for further reference. Figure B.2 is a plot of the 4095 values arranged in descending order. Note that a large fraction of the possible R-Squares are clustered near the maximum attainable with all variables included in the regression equation. The curve falls off only gradually as one goes to the right, until one approaches R-Square = 0.4, approximately.

Figure B.2. Ordered Plot of R-Square for All Subset Models by Stepwise Regression



For comparison, a similar analysis was made of all possible regressions in E-Space – that is, the space in which the response variable is regressed on the eigenvectors of the system. These results are shown in Figure B.3. As for stepwise, a relatively large fraction of the R-Squares fall in the vicinity of the maximum possible with all eigenvectors included in the model. Note, however, that as one proceeds to the right of the plot, R-Square undergoes "steps" – that is, sudden decreases in R-Square. It will become evident that these steps occur because of large changes made when one or more major eigenvectors are dropped from the subsets. In stepwise regression, no such drastic changes occur, because the change from one fuel property to another does not, in general, produce large changes in the value of R-Square.

Figure B.3. Ordered Plot of R-Square for All Subset Models by PCR+



At this point the objection might legitimately be made that the gradual decay of R-Square as one proceeds along the curve is because many of the subsets contain fuel properties that are statistically insignificant and that contribute little to R-Square. Accordingly, one might *expect* many subsets to exhibit essentially the same R-Square simply because non-significant and insubstantial variables are not removed.

To check the validity of this argument, we performed another regression on only those subsets in which there were no t-ratios *less than 1.96*. By implication, all terms in these subsets are allegedly significant at 0.05. These results are shown in Figure B.4 for the stepwise (P-Space) regressions and in Figure B.5 for the PCR+ (E-Space) regressions. For stepwise, the number of subsets was reduced from 4095 to 499 and for PCR+ from 4095 to 76. The curves have essentially the same characteristics as before and, in the case of OLM, still show a relatively large number of possible solutions near the maximum R-Square = 0.6361.

Figure B.4. Ordered Plot of R-Square for All Subset Models with p>=0.05

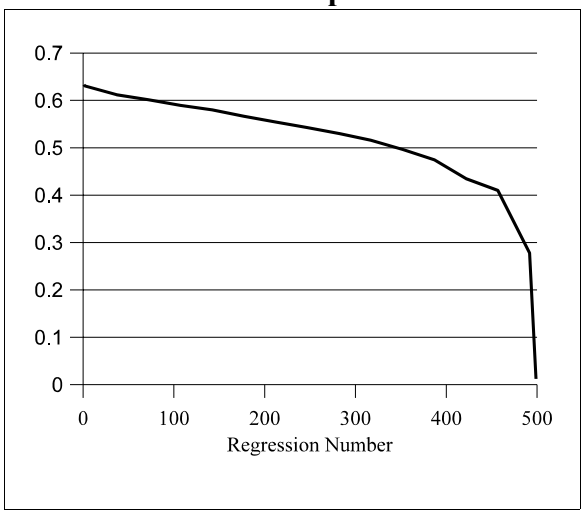
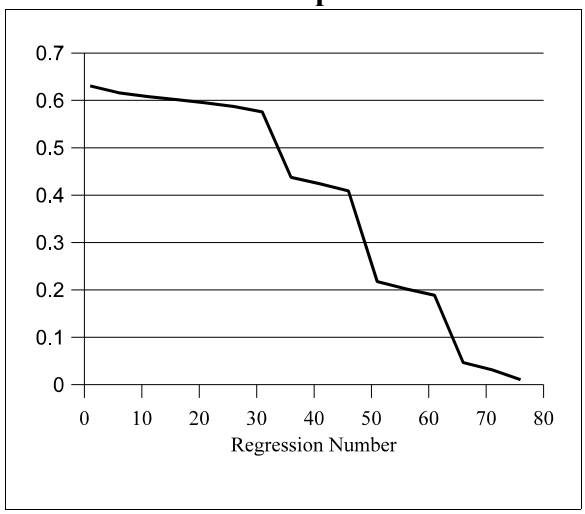


Figure B.5. Ordered Plot of R-Square for all PCR+ Models with p>=0.05



The question now becomes: of all the solutions for which R-Square is nearly maximum, what variations occur in the actual fuel properties or eigenvectors included in the model? For this purpose, a threshold was set at R-Square >= 0.60. In other words, we considered only those solutions that were near the best fit "possible" and that retained terms that, so far as the regression analysis showed, were all "significant" at the 0.05 significance level. This collection of solutions was then listed to show what fuel properties or what eigenvectors were included. The listing for the stepwise regressions are shown in Addendum II and for the PCR+ regressions in Addendum III. The implications of the analysis are summarized in Table B.7 below. The table can be interpreted as follows. Of the 76 subsets for which R-Square >= 0.6 and all fuel properties are claimed to be "significant," there is nearly a 50-50 chance that Natural Cetane is or is not included in the model (32 in, 44 out). Likewise for T50 (36 in, 40 out). In other words, there are models that give essentially the same performance and yet contain a quite different slate of terms in the equation. CetDiff is quite consistent (74 out of 76), and so is Arom (65 out of 76). Still, if any *one* subset model is selected, it is quite problematical what variables will be included in the model.

TABLE B.7. DISTRIBUTION OF "SIGNIFICANT" FUEL PROPERTY VARIABLES IN THE 76 SUBSETS FOR WHICH R-SQUARE >= 0.60

	Included in subset	Excluded from subset
NatCet	32	44
NatCet^2	27	49
CetDif	74	2
CetDif^2	47	29
Arom	65	11
Arom^2	27	49
Sp Grv	57	19
Oxygen	46	30
T10	55	21
T50	36	40
T90	2	74
Sulfur	13	63

Coincident with this finding is the fact that any given fuel property exhibits a wide range of significance in those subsets that include that property as a predictor (see Table B.8). Though all t-values exceed 1.96, as they were constrained to do, there is wide dispersion in the values computed for different subset models. This fact leaves open what the "true" significance of any of the properties is.

For comparison, we present similar statistics for the 17 eligible subsets from PCR+ regression (see Table B.9). Note that in this case only seven of the eigenvectors are listed because, unlike the corresponding case in P-Space, eigenvectors once rejected *stay* rejected in all subsets. In reality, there is only *one* choice for the "best" model, and that is the one with all seven vectors retained.

TABLE B.8. DISTRIBUTION OF t VALUES FOR CANDIDATE REGRESSION SOLUTIONS

	Maximum	Minimum	Mean	StdDev
NatCet	11.4745	4.3723	8.0473	2.3155
NatCet^2	11.5101	4.6534	7.9490	2.2663
CetDif	11.6278	5.2118	7.8545	2.5262
CetDif^2	9.4516	2.0982	2.8563	1.4275
Arom	15.1518	5.2587	9.2980	2.5013
Arom^2	8.1228	1.9793	4.5092	2.6496
Sp Grav	8.9218	2.0219	4.7796	2.3054
Oxygen	3.7636	2.0729	3.0301	0.5258
T10	5.4874	2.4364	3.9741	0.7779
T50	7.7970	1.9655	4.4479	2.2258
T90	4.3498	2.0079	3.1788	1.6559
Sulfur	3.2669	1.9658	2.4559	0.4097

TABLE B.9. DISTRIBUTION OF SIGNIFICANT EIGENVECTORS IN THE 17 SUBSETS FOR WHICH R-SQUARE >= 0.6

Eigenvector	Included in subset	Excluded from subset
1	17	0
3	17	0
5	14	3
6	9	8
9	10	7
10	11	6
11	11	6

Coincident with this finding, one will note that the scatter of the t-values for the coefficients of the retained eigenvectors is very small (see Table B.10). As pointed out earlier, the small differences in t-ratios is the result of small changes in the residual SS that result when vectors are entered into or removed from the model. The much larger range of t-values under stepwise regression is clearly the result of some other phenomenon. This aspect of regression in P-Space is the subject of the next section of this report.

TABLE B.10. DISTRIBUTION OF t VALUES FOR CANDIDATE REGRESSION SOLUTIONS

Eigenvector	Maximum	Minimum	Mean	StdDev
1	22.5605	21.7722	22.0791	0.2376
3	14.8953	14.3748	14.5774	0.1569
5	5.1618	4.9814	5.0614	0.0546
6	2.2314	2.1540	2.1860	0.0250
9	3.1635	3.0529	3.1008	0.0365
10	3.6759	3.5485	3.6028	0.0429
11	4.4190	4.2658	4.3405	0.0483

B.4. ALIASING OF CORRELATED VARIABLES

The peculiarities and anomalies that accompany stepwise regression can be explained by a concept called *aliasing*. The *alias matrix* is defined in Reference 2, Appendix A, and the phenomenon of aliasing is demonstrated in Appendix C to this report. In more common language, one can say that, when variables are highly correlated, any variable *by name*, such as NatCet or Arom, is actually a "mixture" of that variable and other variables with which it is correlated. It is that mixture that gives rise to the eigenvectors, which are mathematically independent and statistically uncorrelated.

Though alias structure can be elucidated by the alias matrix, it may be helpful to view aliasing in terms of the regression of a selected fuel property on the other fuel properties in the model. As was shown in Reference 2, it is not uncommon for as much as 90% of the variation in a fuel property to be "shared" with other properties in the set of candidate predictor variables. In a sense, then, a particular fuel property *by name* actually can be thought of as a *different variable* according to what variables are included in the regression model. This difference in the identity of the named variable is what causes the significance of that named variable to seem to change from subset to subset.

B.5. "EQUIVALENT" REGRESSION MODELS

The diversity of stepwise models that yield essentially the same R-Square suggests that, so far as prediction is concerned, these models are "equivalent." If so, it would seem to make little difference which of the candidate models is selected by a variable-selection procedure, such as stepwise regression. It must be kept in mind, however, that the fact that two models show the same *aggregate* performance does not necessarily assure that the two models will yield the same point-by-point predictions.

To explore this possibility, each of the 76 candidate stepwise models was used to compute the *predicted* emissions for each of the 480 cases in the Tech Group T data set. The correlation matrix for the resulting 480 x 76 matrix was then computed. The 2850 correlation coefficients below the principal diagonal of the correlation matrix exhibit the correlations for all pairs of subset models.

A histogram of these 2850 correlation coefficients is shown in Figure B.6. Note that approximately half of the 2850 pairs of subset models are correlated at 0.98 or above. Even the minimum is 0.94. Thus one might conclude that, as far as prediction is concerned, *it really does not matter* which subset model is selected, so long as all terms in the model are said to be significant and R-Square is near the maximum possible R-Square.

How do the multiple stepwise models compare with the best PCR+ model? This question can be answered by computing the correlation between the PCR+ predictions and each of the 76 possible Stepwise predictions. Results of this analysis are shown in Figure B.7. Approximately half of the correlations are at or above 0.98 and the minimum correlation is 0.97. Thus, it might be concluded that, so far as prediction is concerned, any of the stepwise models make about the same predictions as the PCR+ model.

Figure B.6 Distribution of Correlations Between Stepwise Subset Models

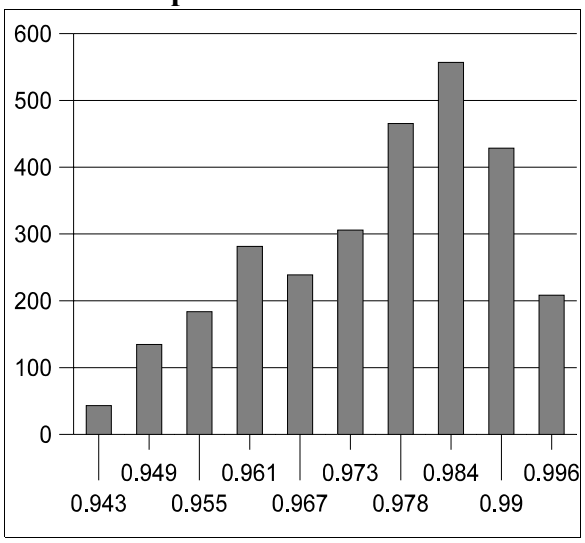
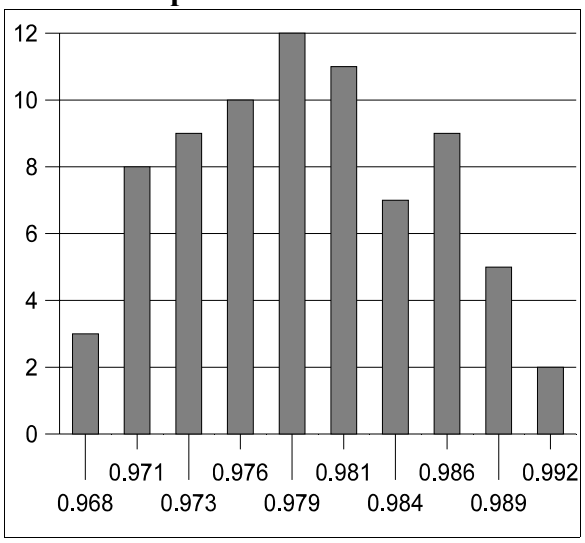


Figure B.7 Distribution of Correlations Between Stepwise and Best PCR+ Models



Two caveats are still to be observed, however. First, the correlations apply *only to the observations in the data set,* although the prediction model, whatever its form, is usually assumed to be valid for interpolation and extrapolation within the statistical confidence bounds set by the design matrix and the overall error variance. Secondly, the fact that predictions from two models correlate highly does not necessarily preclude the possibility of "lack of fit."

Two caveats are still to be observed, however. First, the correlations apply *only to the observations in the data set,* although the prediction model, whatever its form, is usually assumed to be valid for interpolation and extrapolation within the statistical confidence bounds set by the design matrix and the overall error variance. Secondly, the fact that predictions from two models correlate highly does not necessarily preclude the possibility of "lack of fit."

The use of a regression model to predict emissions for fuel vectors *not in the data set* is actually its main claim to fame. If the points in the data set were all that were of interest, there would be little point in computing a regression model. Mathematically, the procedure can be thought of as a means for extending the *domain* of a function from a finite set of isolated points to a continuum in multidimensional space. It can be readily shown that there exists an infinite number of "models" that yield identically the same response at points in the original point set but which diverge greatly from each other in the region outside those points. It is for this reason that statistical analysts often do "validation testing" by dividing the data set into two parts. One part, often referred to as the "training set," is used for computing the regression equation. Then that model is used to predict response for the points in the second part of the data, often referred to as the "test set." Bootstrap sampling is another version of such validation testing.

The term "lack of fit" has a specific connotation in regression analysis. It refers to "patterns of residuals" that result when predicted responses are compared with observed responses. For example, suppose a set of data actually follows a parabolic curve but was fit by a straight-line function. In such an instance, there is a tendency for points in the middle of the range of the data to fall systematically on one side of the predictive line, while points near the extremes of the range of data tend to fall on the other side of the predictive line. In the case of diesel emissions, for example, it is possible for the effects of such variables as NatCet or Arom to "level off" when increased beyond a certain value. The phenomenon can be likened to "saturation" in a photographic emulsion: beyond a certain exposure level, additional exposure can do little to cause further

blackening of the medium. Though the "lack of fit" component may make little difference in an aggregate measure such as R-Square, it could introduce appreciable error in regions where response is subject to curvature away from the predicted straight line.

Finally, and most important, is the fact that a given stepwise equation may list fuel properties that *appear* to be the driving forces for emissions while another stepwise equation may list quite different variables yet still give essentially equivalent predictions. It is this ambiguity that provides the strongest argument in favor of PCR+, which lists *combinations* of variables that drive the emission response, combinations that are invariant. Moreover, because the contributions of these combinations can be further broken down into contributions of the individual fuel properties that make up the eigenvectors, PCR+ discloses the most likely relative importance of those properties.

B.6. VARIABLE SELECTION: COMMENTS AND CAVEATS

As noted in the introduction, the selection of variables from a slate of candidates has become as much a part of multiple regression analysis as the least-squares principle for minimizing the error SS. It has commanded a sizable portion of the literature on regression [Ref 3] and has spawned a number of procedures, a few of which have enjoyed extensive application. One of these procedures, of course, is stepwise regression. In spite of caveats and warnings, the methodology endures.

Here is what the SYSTAT manual [Ref 4] has to say:

> Stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need automated stepwise regression to solve a particular problem, it is almost certain that you do not. Professional statisticians rarely use automated stepwise regression because it does not necessarily find (a) the "best" fitting model, (b) the "real" model, or (c) alternative "plausible" models. Furthermore, the order in which variables enter or leave a stepwise program is usually of no theoretical significance. You are always better off thinking about why a model could generate your data and then testing that model. {Page 185}

And, in discussing the validity of measures of significance, which we have shown to be highly unstable when predictors are correlated, the manual says:

> Stepwise regression programs are the most notorious source of pseudo "p-values" in the field of automated data analysis. Stepwise regression programs that print pseudo F-tests and p-values invite abuse; statisticians seem to be the only ones who know these are not "real" p-values.

A number of researchers have attempted to deal with the significance issue. For example, we cite Wilkinson [Ref 5], Wilkinson and Dallal [Ref 6], and Rencher and Pun [Ref 7]. In his 1979 paper, Wilkinson asserts:

> Stepwise regression has a controversial role in statistical data analysis. Since the introduction of various automated techniques for selecting the "best" subset of a set of predictor variables in a multiple regression, researchers have been warned about their indiscriminate use. The primary reason for this caution is that for any subset selection procedure based on inspection of the sample data, the usual F statistic for testing the significance of the multiple correlation is biased.

Wilkinson and Dallal then develop, by simulation, revised tables for assessing the significance of R-Square under an "F-to-Enter" stopping rule. The thrust of this and later work is that the usual statistical tables for R-Square and F are not applicable in the environment of subset selection and give an overly optimistic assessment of the statistical significance of the model.

A very striking affirmation of the caveats regarding subset regression is provided by Flack and Chang [Ref 8]. In a simulation study, they mixed "authentic" and "noise" predictor variables and performed a large number of regressions to see how often the authentic variables would be included in the selected subset as opposed to how often the noise variables would be included. The "authentic" variables were constructed so that they would have a correlation coefficient of 0.50 with the response variable; the noise variables had zero correlation. Surprisingly, the selected subsets contained noise variables about as often as they contained the authentic variables.

It is clear that variable selection is not a problem if the predictor variables are independent. As has been shown earlier, difficulty arises when the variables are correlated to any appreciable degree. Many of the methods devised to "fix" this problem seem to have missed the real culprit, the correlation among the predictor variables. It is this correlation that makes for the variability of the "p-value" for a given variable depending upon what other variables accompany it.

It is our contention that there *is no* fix other than orthogonalization. PCR+ methodology reveals that *every* data set has its own orthogonal basis and that significance can be defined rigorously *only* for the orthogonal variables that come from PCA of the predictor variable design matrix. Once these orthogonal effects have been identified and evaluated, the relative importance of the roles played by the original variables can be determined by a method for partitioning the model SS, as discussed elsewhere. Moreover, the chosen subset model *for the eigenvectors* can be re-expressed in terms of the original variables to produce, in one operation, a model that performs at least as well as any subset model selected by stepwise means. And, most important of all, perhaps, is the fact that the model derived in this way provides an ordering of the importance of the original variables. As has been shown in our "all possible regressions" exercise, there are many models that perform equally well so far as prediction is concerned, but which show wide diversity in the variables seemingly emphasized in the model.

B.7. A SHORT-CUT TO VARIABLE SELECTION

As has been demonstrated, the significance level for a variable correlated with other variables is not an invariant property when computed by conventional Stepwise means. Moreover, the overall significance of a particular subset model is not amenable to conventional tests of significance for R-Square and F. The PCR+ method, on the other hand, provides stable and essentially invariant measures of significance, though these measures apply not to individual fuel properties but to weighted combinations of those properties. This seeming difficulty can be remedied by a simple stratagem.

Suppose that a response variable y is regressed on N predictor variables x_1, x_2, ... x_N, and that the predictor variables are correlated to varying degree. One can compute the correlation matrix for these variables and by means described in Reference 2, can compute the eigenvalues and eigenvectors of that correlation matrix. The P-Space coefficients can be transformed to the equivalent E-space coefficients simply by multiplying the P-Space coefficients by the transpose of the eigenvector matrix.

Once in E-Space, one can apply tests of significance to the E-Space coefficients rigorously and without the need for iterations as required in the application of stepwise methodology. Eigenvectors that are insignificant and/or insubstantial can be dropped from the regression equation; then the remaining terms can be re-expressed in terms of the original fuel property variables.

At this point the objection might be raised that the resulting equation still contains *all* of the original, fuel-property variables. This fact, however, is irrelevant, because significance was established on the basis of the transformed *vectors*, and the fuel-property values serve only to compute the weights of the eigenvectors retained in the model. All necessary adjustments for the *relative* significance of the property variables are

already implicit in the significance tests of the vectors in E-Space. The fact that a few more terms might be required in the regression equation is not particularly relevant in the present age of computational ability.

If desired, for purposes of brevity or for a greater comfort level with the new approach, the vectors can be "pruned," by the method set forth in Reference 2, to remove any of the fuel property variables that contribute little to the prediction capability of the model. The model SS can be partitioned two ways: (1) according to the contribution to the SS by each of the eigenvectors, or (2) according to the contribution to the SS by each of the original variables. The latter is accomplished by means of a Matlab program called Simplify.m that is included in Reference 2, page 96, and discussed in Appendix D to this report. Demonstration of the procedure, as applied to the data of Tech Group T, is given in Addendum IV.

At this point, one can return to P-Space with a smaller number of variables. As before, the procedure can begin with conventional Stepwise regression, and, since there is no longer any need to transform to E-Space, the Stepwise-derived equation can stand as the "best" choice, having been selected by a method based on independent and uncorrelated "reformulated" variables. Moreover, the equation is operationally the same whether expressed in P-Space or E-Space and will give equivalent predictions for either the observations in the original data set of observations or for interpolated or extrapolated predictions at points in the X-Space *not* in the original data set.

B.8. SUMMARY AND CONCLUSIONS

This appendix has explored all possible "solutions" to the problem of fitting the "best" fuel-response model to data representing engine Tech Group T in the EPA database. Of the 4095 possible "subset models" we found 76 models for which all terms are "significant" as judged by conventional tests of significance at the 0.05 level and for which R-Square was between 0.6000 and 0.6361. We conclude that these different forms of the model are essentially equivalent, so far as prediction goes. The correlation coefficients for all pairs of the models are equal or greater than 0.94, and all the models correlated with the full eigenvector model at 0.96 or better.

We also have seen that an eigenvector-based model is capable of predictions equal to or better than the "best" subset model selected by other means. Moreover, the eigenvector-based model offers a plausible selection of the original variables believed to represent more truthfully their relative importance. A simple approach to deriving such a model is presented and illustrated in Addendum IV.

Inasmuch as the method starts and ends with regression in fuel-property space, it requires relatively little departure from conventional stepwise regression. The methodology takes only a brief detour into E-Space and provides a quick and efficient way to avoid the multiple steps and possibly ambiguous conclusions of stepwise regression.

Finally, it is concluded that resolving a data set into its innate independent variables, as represented by the eigenvectors of the set, is the *only* way that multicollinearity can be truly eliminated. No stepwise solution, however arrived at, can nullify the fact that any given variable is aliased with other variables with which it is correlated. This aliasing changes from subset to subset, whereas a model based on independent entities is robust regardless of what subset of those entities is included in the model. Moreover, the eigenvectors are monotonically ordered, whereas the order of importance exhibited in P-Space is dependent on what variables are included in the subset model.

It is recommended, therefore, that wherever possible variable selection in a regression model should be guided by eigenvector decomposition of the data set. If desired, the resulting model can be transformed back into the space of the original variables. Though the resulting model uses variables that are aliased with each

other, that aliasing is the one most representative of the relative importance of those variables in their effect on response.

B.9. REFERENCES

1. McAdams, H. T., R.W. Crawford and G.R.Hadder. 2000. *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*, SAE Technical Paper 2000-01-1961.

2. McAdams, H.T., W.R. Crawford and G.R. Hadder. 2000. *A Vector Approach toRegression Analysis and Its Application to Heavy-Duty Diesel Emissions*, ORNL/TM-2000/5, Oak Ridge National Laboratory, Oak Ridge, TN. November.

3. Hocking, R. R. 1983. *Developments in Linear Regression Methodology: 1959-1982*. Technometrics, 25, 219-230.

4. Wilkinson, Leland. 1990. *SYSTAT: The System for Statistics*, Evanston, IL, SYSTAT, Inc.

5. Wilkinson, L. 1979. *Tests of Significance in Stepwise Regression*. Psychological Bulletin, 86, 168-174.

6. Wilkinson, L. and G. E. Dallal. 1982. *Tests of Significance in Forward Selection Regression with an F-to-enter Stopping Rule*. Technometrics, 24, 25-28.

7. Rencher, A.C. and F. C. Pun. 1980. *Inflation of R-Squared in Best Subset Regression*. Technometrics, 22, 49-54.

8. Flack, V.F. and P. C. Chang. 1987. *Frequency of Selecting Noise Variables in Subset Regression Analysis. A Simulation Study.* The American Statistician, 37, 152-155.

Addendum I

NOMENCLATURE AND NOTATION

The following is a list of abbreviations used for fuel property variables throughout this report.

Arom	Aromatics
$Arom^2$	Aromatics squared
CetDif	Cetane Improver
$CetDif^2$	Cetane Improver squared
F	Ratio of two mean squares
logres	Logarithm of engine-corrected residuals
Max	Maximum
Min	Minimum
NatCet	Natural Cetane
$NatCet^2$	Natural Cetane squared
OLS	Ordinary Least Squares
Oxygen	Oxygen
PCR	Principal Components Regression
PCR+	Principal Components Regression as applied in References [1] and [2]
Reg Coef	Regression coefficient
Sp Grv	Specific Gravity
SS	Sum of squares
Std Dev	Standard Deviation
Std Err	Standard error
Sulfur	Sulfur
t	Ratio of a statistic to its standard error
T10	Temperature for 10% distillation
T50	Temperature for 50% distillation
T90	Temperature for 90% distillation

Addendum II

Stepwise (P-SPACE) LIST OF REGRESSION SUBSETS FOR WHICH R-SQUARE >= 0.6 AND ALL TERMS ARE SIGNIFICANT AT 0.05

NC	NC2	CD	CD2	AR	AR2	DENS	OXY	T10	T50	T90	SULF	R^2
0	0	1	0	1	0	1	0	1	1	0	1	0.6020
0	0	1	0	1	0	1	1	1	1	0	0	0.6075
0	0	1	0	1	0	1	1	1	1	0	1	0.6116
0	0	1	0	1	1	1	0	1	1	0	0	0.6023
0	0	1	0	1	1	1	0	1	1	0	1	0.6070
0	0	1	0	1	1	1	1	1	1	0	0	0.6127
0	0	1	0	1	1	1	1	1	1	0	1	0.6166
0	0	1	1	1	0	1	0	1	1	0	0	0.6017
0	0	1	1	1	0	1	0	1	1	0	1	0.6077
0	0	1	1	1	0	1	1	0	1	0	1	0.6024
0	0	1	1	1	0	1	1	1	1	0	0	0.6131
0	0	1	1	1	0	1	1	1	1	0	1	0.6183
0	0	1	1	1	1	1	0	1	1	0	0	0.6072
0	0	1	1	1	1	1	0	1	1	0	1	0.6130
0	0	1	1	1	1	1	1	0	0	1	1	0.6019
0	0	1	1	1	1	1	1	1	1	0	0	0.6187
0	0	1	1	1	1	1	1	1	1	0	1	0.6236
0	1	0	1	1	0	1	1	1	1	0	0	0.6042
0	1	1	0	0	1	1	1	1	0	0	0	0.6017
0	1	1	0	0	1	1	1	1	1	0	0	0.6066
0	1	1	0	1	0	0	0	1	0	0	0	0.6081
0	1	1	0	1	0	0	1	0	0	0	0	0.6012
0	1	1	0	1	0	0	1	1	0	0	0	0.6155
0	1	1	0	1	0	1	0	0	0	0	0	0.6050
0	1	1	0	1	0	1	0	1	0	0	0	0.6114
0	1	1	0	1	0	1	0	1	1	0	0	0.6160
0	1	1	0	1	0	1	1	0	0	0	0	0.6086
0	1	1	0	1	0	1	1	1	1	0	0	0.6256
0	1	1	1	0	1	1	1	0	0	0	0	0.6011
0	1	1	1	0	1	1	1	1	0	0	0	0.6061

Addendum II (Cont.)

NC	NC2	CD	CD2	AR	AR2	DENS	OXY	T10	T50	T90	SULF	R^2
0	1	1	1	0	1	1	1	1	1	0	0	0.6104
0	1	1	1	0	1	1	1	1	1	1	0	0.6137
0	1	1	1	1	0	0	0	0	0	0	0	0.6019
0	1	1	1	1	0	0	0	1	0	0	0	0.6122
0	1	1	1	1	0	0	1	0	0	0	0	0.6076
0	1	1	1	1	0	0	1	1	0	0	0	0.6208
0	1	1	1	1	0	1	0	0	0	0	0	0.6110
0	1	1	1	1	0	1	0	0	0	0	1	0.6142
0	1	1	1	1	0	1	0	1	0	0	0	0.6162
0	1	1	1	1	0	1	0	1	1	0	0	0.6201
0	1	1	1	1	0	1	1	0	0	0	0	0.6156
0	1	1	1	1	0	1	1	0	0	0	1	0.6187
0	1	1	1	1	0	1	1	1	1	0	0	0.6307
0	1	1	1	1	1	0	1	1	1	0	0	0.6265
1	0	0	1	1	0	1	1	1	1	0	0	0.6041
1	0	1	0	0	1	1	1	1	0	0	0	0.6018
1	0	1	0	0	1	1	1	1	1	0	0	0.6061
1	0	1	0	1	0	0	0	1	0	0	0	0.6071
1	0	1	0	1	0	0	1	1	0	0	0	0.6150
1	0	1	0	1	0	1	0	0	0	0	0	0.6051
1	0	1	0	1	0	1	0	1	0	0	0	0.6120
1	0	1	0	1	0	1	0	1	1	0	0	0.6159
1	0	1	0	1	0	1	1	0	0	0	0	0.6087
1	0	1	0	1	0	1	1	1	0	0	0	0.6183
1	0	1	0	1	0	1	1	1	1	0	0	0.6255
1	0	1	0	1	1	0	0	1	0	0	0	0.6110
1	0	1	0	1	1	0	1	1	0	0	0	0.6186
1	0	1	1	0	1	1	1	0	0	0	0	0.6003
1	0	1	1	0	1	1	1	1	0	0	0	0.6059
1	0	1	1	0	1	1	1	1	1	0	0	0.6097
1	0	1	1	1	0	0	0	0	1	0	0	0.6031
1	0	1	1	1	0	0	0	1	0	0	0	0.6108
1	0	1	1	1	0	0	1	0	0	0	0	0.6049
1	0	1	1	1	0	0	1	0	1	0	0	0.6081
1	0	1	1	1	0	0	1	1	0	0	0	0.6199
1	0	1	1	1	0	1	0	0	0	0	0	0.6108

Addendum II (Cont.)

NC	NC2	CD	CD2	AR	AR2	DENS	OXY	T10	T50	T90	SULF	R^2
1	0	1	1	1	0	1	0	0	0	0	1	0.6140
1	0	1	1	1	0	1	0	1	0	0	0	0.6165
1	0	1	1	1	0	1	0	1	1	0	0	0.6199
1	0	1	1	1	0	1	1	0	0	0	0	0.6154
1	0	1	1	1	0	1	1	1	0	0	0	0.6238
1	0	1	1	1	0	1	1	1	1	0	0	0.6304
1	0	1	1	1	1	0	0	1	0	0	0	0.6151
1	0	1	1	1	1	0	1	1	0	0	0	0.6239
1	0	1	1	1	1	1	0	1	1	0	0	0.6230
1	0	1	1	1	1	1	1	1	1	0	0	0.6336

Addendum III

PCR+ (E-SPACE) LIST OF REGRESSION SUBSETS FOR WHICH R-SQUARE >= 0.6 AND ALL TERMS ARE SIGNIFICANT AT 0.05

Eigenvector											
1	2	3	4	5	6	7	8	9	10	11	12	R^2
1	0	1	0	1	0	0	0	1	0	0	0	0.6008
1	0	1	0	0	1	0	0	0	1	1	0	0.6018
1	0	1	0	1	0	0	0	0	1	0	0	0.6035
1	0	1	0	1	1	0	0	1	0	0	0	0.6047
1	0	1	0	0	0	0	0	1	1	1	0	0.6058
1	0	1	0	1	1	0	0	0	1	0	0	0.6074
1	0	1	0	1	0	0	0	0	0	1	0	0.6082
1	0	1	0	0	1	0	0	1	1	1	0	0.6097
1	0	1	0	1	0	0	0	1	1	0	0	0.6114
1	0	1	0	1	1	0	0	0	0	1	0	0.6121
1	0	1	0	1	1	0	0	1	1	0	0	0.6152
1	0	1	0	1	0	0	0	1	0	1	0	0.6161
1	0	1	0	1	0	0	0	0	1	1	0	0.6188
1	0	1	0	1	1	0	0	1	0	1	0	0.6200
1	0	1	0	1	1	0	0	0	1	1	0	0.6227
1	0	1	0	1	0	0	0	1	1	1	0	0.6266
1	0	1	0	1	1	0	0	1	1	1	0	0.6305

Addendum IV

A SHORTCUT TO VARIABLE SELECTION

The table below, shown in the text of this report as Table B.1, gives the results of regressing the log NO_x residuals on the full slate of 12 fuel-property variables in P-Space. By "residuals" is meant the residuals obtained after correcting log NO_x observations for engine bias.

REGRESSION OF LOG NO_x RESIDUALS (LOGRES) ON FUEL VARIABLES

R-Square = 0.6361

Eigenvector	Regression Coefficient	Coefficient Std Error	t-Ratio
NatCet	-0.0077	0.0134	0.5779
$NatCet^2$	-0.0042	0.0135	0.3110
CetDif	-0.0289	0.0047	6.1491*
$CetDif^2$	0.0127	0.0046	2.7813*
Arom	0.0324	0.0062	5.2428*
$Arom^2$	-0.0098	0.0054	1.8082
Sp Grav	0.0106	0.0033	3.1955*
Oxygen	0.0053	0.0014	3.6766*
T10	0.0104	0.0023	4.5187*
T50	-0.0104	0.0034	3.0581*
T90	0.0021	0.0022	0.9460
Sulfur	-0.0021	0.0015	1.4289

*Significant at p=0.05 or better

As noted in the text, the computed significance values change when one removes from the slate of variables those indicated above to be non-significant. Moreover, it has been shown elsewhere in this report that the "p-values" corresponding to the t-ratios are not to be trusted, inasmuch as they change from subset to subset of the predictor variables.

The table below gives the results of regressing the log NO_x residuals on the full slate of 12 eigenvectors of the correlation matrix. It has also been pointed out that the E-Space coefficients can be computed directly from the P-Space coefficients simply by multiplying those coefficients by the transpose of the matrix of eigenvectors. Tests of significance for the E-Space coefficients can be performed straightforwardly just as they were in P-Space.

In E-Space, as has been shown earlier, the t-ratios are stable and hence more reliable than the t-ratios computed in P-Space. Note that five of the eigenvectors - namely, 2, 4, 7, 8 and 12 - are decidedly nonsignificant according to the 0.05 significance level.

REGRESSION OF LOG NO_x RESIDUALS (LOGRES) ON EIGENVECTORS

R-Square = 0.6361

Eigenvector	Regression Coefficient	Coefficient Std Error	t-Ratio
1	-0.0150	0.0007	22.6119*
2	-0.0007	0.0008	0.8607
3	-0.0152	0.0010	14.9292*
4	-0.0013	0.0013	0.9901
5	-0.0072	0.0014	5.1735*
6	0.0037	0.0017	2.2364*
7	0.0029	0.0021	1.3674
8	0.0063	0.0034	1.8693
9	0.0138	0.0044	3.1707*
10	0.0237	0.0064	3.6843*
11	-0.0352	0.0079	4.4290*
12	0.0049	0.0190	0.2596

*Significant at p=0.05 or better

The SS partitioning according to eigenvectors can be computed quite simply as

diagd.*coef.*coef*(n-1)

where diagd denotes the eigenvalues of the system in the form of a column vector and coef denotes the regression coefficients, also expressed as a column vector. The indicated multiplication is understood to mean multiplying all entries on a given line, line by line, as shown on the following page. The Analysis of Variance table, as computed for the aggregate model, is:

	SS	DF	MS
Model	0.7029	12	0.0586
Error	0.4021	467	0.0009
Total	1.1050		

Note that the sum of the contributions for the 12 eigenvectors checks the Model SS as given by Analysis of Variance.

Eigenvector	diagd		coef		coef	n-1	SS
1	4.0687	*	-0.0150	*	-0.0150	* 479 =	0.4402
2	2.7965	*	-0.0007	*	-0.0007	* 479 =	0.0006
3	1.7353	*	-0.0152	*	-0.0152	* 479 =	0.1919
4	1.0786	*	-0.0013	*	-0.0013	* 479 =	0.0008
5	0.9281	*	-0.0072	*	-0.0072	* 479 =	0.0230
6	0.6572	*	0.0037	*	0.0037	* 479 =	0.0043
7	0.4040	*	0.0029	*	0.0029	* 479 =	0.0016
8	0.1600	*	0.0063	*	0.0063	* 479 =	0.0030
9	0.0947	*	0.0138	*	0.0138	* 479 =	0.0087
10	0.0434	*	0.0237	*	0.0237	* 479 =	0.0117
11	0.0285	*	-0.0352	*	-0.0352	* 479 =	0.0169
12	0.0050	*	0.0049	*	0.0049	* 479 =	0.0001
					Sum	.....	0.7029

We wish now to partition the SS for each eigenvector into the contributions *to that eigenvector's SS* by each of its components. These components are, of course, the fuel-property loadings. This partitioning is accomplished by means of the previously mentioned Matlab program Simplify.m written for that purpose and included for reference in Addendum IV.

The first three columns of the following table pertain to the analysis when all 12 eigenvectors are included in the model. The columns give, respectively, the SS contribution for each of the fuel properties, those contributions expressed as a percent of the total model SS, and an F-ratio computed as the ratio of the fuel-property SS to the corresponding mean error SS. The last three columns pertain to the analysis when only those eigenvectors significant at 0.05 are retained.

	All 12 Eigenvectors			Subset of 7 significant Eigenvectors		
	SS	% SS	F	SS	% SS	F
NatCet	0.1062	15.1130	123.37*	0.1047	15.0229	121.01*
NatCet^2	0.1059	15.0677	123.00*	0.1146	16.4526	132.53*
CetDif	0.0663	9.4370	77.04*	0.0641	9.2010	74.12*
CetDif^2	0.0098	1.4011	11.44*	0.0098	1.4132	11.38*
Arom	0.1920	27.3149	222.98*	0.1754	25.1763	202.80*
Arom^2	0.0798	11.3593	92.73*	0.0639	9.1725	73.89*
Sp Grv	0.1194	16.9831	138.64*	0.1516	21.7611	175.29*
Oxygen	0.0033	0.4672	3.81	0.0004	0.0550	0.44
T10	0.0167	2.3736	19.38*	0.0107	1.5384	12.39*
T50	0.0000	0.0049	0.04	0.0000	0.0054	0.04
T90	0.0029	0.4114	3.36	0.0014	0.1978	1.59
TSulfur	0.0005	0.0670	0.55	0.0000	0.0038	0.03
Total	0.7029	100.0000		0.6967	100.0000	

* Significant at p=0.05 or better per F>=3.85

Significant terms are: NatCet, NatCet^2, CetDif, CetDif^2, Arom, Arom^2, Dens, and T10. Selection of these variables is based on the last three columns of the table – that is, the selection was made *after rejection of the nonsignificant eigenvectors as a whole*. This is considered to be appropriate, because rejection of eigenvectors 2, 4, 7, 8 and 12 eliminates the "nearly significant" terms T50 and Oxy, each of which contribute

minimally to the model SS. All of the terms retained also satisfy the substantiality criterion of contributing more than 1% to the model SS.

The following table gives the P-Space partitionings for each of the 12 eigenvectors taken singly.

Eigenvector:	**1**	**2**	**3**	**4**	**5**	**6**
NatCet	0.0563	0.0001	0.0085	0.0000	0.0003	0.0001
NatCet^2	0.0570	0.0001	0.0083	0.0000	0.0003	0.0001
CetDif	0.0167	0.0000	0.0745	0.0000	0.0002	0.0000
CetDif^2	0.0116	0.0000	0.0768	0.0000	0.0004	0.0000
Arom	0.0921	0.0000	0.0022	0.0000	0.0000	0.0002
Arom^2	0.0788	0.0000	0.0028	0.0000	0.0000	0.0002
Sp Grv	0.0922	0.0000	0.0005	0.0000	0.0000	0.0001
Oxygen	0.0002	0.0000	0.0017	0.0006	0.0020	0.0004
T10	0.0020	0.0001	0.0017	0.0001	0.0011	0.0015
T50	0.0153	0.0002	0.0067	0.0000	0.0003	0.0001
T90	0.0132	0.0001	0.0043	0.0000	0.0000	0.0012
Sulfur	0.0047	0.0000	0.0038	0.0000	0.0184	0.0003
Sum	0.4402	0.0006	0.1919	0.0008	0.0230	0.0043

Eigenvector:	**7**	**8**	**9**	**10**	**11**	**12**
NatCet	0.0001	0.0003	0.0003	0.0000	0.0002	0.0000
NatCet^2	0.0001	0.0002	0.0004	0.0000	0.0000	0.0000
CetDif	0.0000	0.0000	0.0000	0.0057	0.0004	0.0000
CetDif^2	0.0000	0.0000	0.0000	0.0055	0.0002	0.0000
Arom	0.0002	0.0000	0.0000	0.0002	0.0095	0.0000
Arom^2	0.0005	0.0000	0.0000	0.0002	0.0065	0.0000
Sp Grv	0.0001	0.0012	0.0024	0.0001	0.0001	0.0000
Oxygen	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T10	0.0000	0.0005	0.0008	0.0000	0.0000	0.0000
T50	0.0000	0.0003	0.0043	0.0000	0.0000	0.0000
T90	0.0004	0.0005	0.0003	0.0000	0.0000	0.0000
Sulfur	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
Sum	0.0016	0.0030	0.0087	0.0117	0.0169	0.0001

The total SS computed as the sum of the 12 column sums is 0.7029, in agreement with the regression Analysis of Variance. The total SS computed as the sum of the 7 column sums of the retained eigenvectors is 0.6967, in agreement with the regression Analysis of Variance when only the retained vectors are included in the regression model. It is to be especially noted, however, that the SS contributions by the property variables should not be computed simply by adding across columns. The correct p-variable contributions for subset models containing more than one eigenvector must be computed by application of the Matlab program Simplify.m or its equivalent.

It is interesting, however, to see the corresponding *percentage* contributions of the fuel properties to *each* of the eigenvector SS *considered singly*. These are shown below. These percentages are particularly helpful in "characterizing" the eigenvectors in terms of what they represent from a practical standpoint.

Eigenvector:	**1**	**2**	**3**	**4**	**5**	**6**
NatCet	12.7940	12.0309	4.4519	0.1185	1.2091	1.9292
NatCet^2	12.9379	11.5295	4.3316	0.1263	1.2198	2.9779
CetDif	3.8004	4.9536	38.8339	0.0398	1.0450	0.0080
CetDif^2	2.6355	5.4063	40.0236	0.3772	1.9376	0.1843
Arom	20.9181	0.7848	1.1441	0.1053	0.0015	4.8949
Arom^2	17.9103	2.5042	1.4725	0.0005	0.0250	5.4075
Sp Grv	20.9374	0.0358	0.2850	0.1191	0.0402	2.8762
Oxygen	0.0416	0.0103	0.8616	76.6913	8.5893	9.9197
T10	0.4639	17.7902	0.8717	15.1153	4.6096	33.9552
T50	3.4814	24.5723	3.5035	0.0000	1.4237	3.2420
T90	3.0043	17.4293	2.2510	3.0366	0.1865	27.1028
Sulfur	1.0750	2.9527	1.9695	4.2701	79.7125	7.5021

Eigenvector:	**7**	**8**	**9**	**10**	**11**	**12**
NatCet	5.0929	8.4045	3.9583	0.0168	1.3201	48.6738
NatCet^2	7.0834	5.2236	4.4888	0.0612	0.0545	49.9653
CetDif	0.2172	0.0001	0.0201	48.9083	2.1726	0.0009
CetDif^2	0.4079	0.8891	0.0069	46.7165	1.4023	0.0126
Arom	14.1216	0.0129	0.0131	1.6219	56.0045	0.3773
Arom^2	31.5543	0.4882	0.0325	1.6125	38.2229	0.7696
Sp Grv	8.2172	38.5512	27.8311	0.5613	0.4756	0.0698
Oxygen	2.0784	1.3075	0.4817	0.0128	0.0052	0.0005
T10	1.1769	15.9848	9.5188	0.3158	0.1977	0.0001
T50	2.7654	10.8372	50.0162	0.0383	0.0222	0.0978
T90	27.0835	16.1734	3.5542	0.0397	0.1065	0.0321
Sulfur	0.2011	2.1276	0.0783	0.0949	0.0158	0.0003

For example, NatCet and its square, Arom and its square, plus Dens dominate Eigenvector #1, whereas CetDif and its square clearly dominate Eigenvector #3.

It is our contention that this "shortcut" to variable selection is not only more efficient than stepwise methods but also gives the most representative choice from among the relatively large number of choices that might be made by a "best subset" choice.

A question that might be raised is this: How does a conventional regression, based on the variables selected by this shortcut method, compare with the eigenvector model based on the retained eigenvectors?

The answer is that they will be identical under the following procedure. Perform Stepwise regression to fit the log NO_x residuals to the selected fuel properties. Then, recompute the eigenvectors and eigenvalues *for just the retained fuel properties*. Finally, use that eigenvector matrix to transform the fuel-property coefficients to eigenvector coefficients. These coefficients will be *exactly the same* as those obtained by fitting the log NO_x residuals to the eigenvector weights. Except for the additional insights obtained from the eigenvector regression, there is no need to transform the fuel-property equation into E-Space.

In summary, this *shortcut to variable selection* consists of a few simple steps:

1. By OLS, derive the "best fit" equation in P-Space *with all P-variables included*.

2. Transform the P-Space coefficients to E-Space coefficients by multiplying the P-space coefficients by the transpose of the eigenvector matrix *for all P-Variables*.
3. Apply tests of significance and/or substantiality to the eigenvector coefficients; reject those eigenvectors that do not qualify.
4. Use Simplify.m or an equivalent procedure to eliminate nonsignificant (by F test) and/or nonsubstantial P-Space variables.
5. By OLS, derive the "best fit" equation in P-Space *with just the retained P-variables included*.
6. If desired, transform this equation into E-Space as in (2) but using the eigenvector matrix for just the retained P-variables.

Note that if Step 6 is omitted, the above procedure can be viewed purely as a variable-selection process. Its only involvement with E-Space is to perform that selection by means that are unique and that avoid the possible multiplicity and ambiguity that can affect stepwise methods. The method also claims to highlight the most appropriate P-variables for inclusion in a P-Space model. *Some* stepwise models can "miss" important variables or overemphasize relatively unimportant variables because of the particular *aliasing* that accompanies that particular choice of variables.

APPENDIX C

CALCULATION OF THE ALIAS MATRIX

APPENDIX C. CALCULATION OF THE ALIAS MATRIX

The use of the alias matrix in regression analysis is generically the same as its use in applications such as fractional factorial. One may recall that, in those applications, fractional factorials were aimed at such objectives as generating a half-or quarter-factorial (or smaller fractions) that claimed to be "two-factor interaction clear." What the design did was to select and arrange treatments so that linear and first-order interactions were confounded only with interactions of higher order, the assumption being that these higher-order interactions were zero – i.e., non-existent.

In a regression context, the alias matrix is defined as

$$A = (X_{in}^{T} * X_{in})^{-1} * X_{in}^{T} * X_{out} \quad (1)$$

where X_{in} denotes those columns of the original design matrix (x-matrix) that are retained,

and X_{out} denotes those columns of the X-matrix that are rejected.

The definition given in Eq. 1 stems from a simple extension of the general linear model in which it is assumed that some essential variables were left out of the model and it is desired to know how that omission might have affected ("biased") the model that was originally thought to be adequate. The application of the alias matrix in this report begins at the other end – that is, it assumes that the model was appropriate at the outset, but some terms were deleted, whatever the reason. Then, the question is: if these terms should <u>not</u> have been excluded, how would their omission affect those that were kept?

The following is an example based on log(NO_x) data from technology group T in the EPA database. The data used have been corrected for engine effects and are actually the residuals remaining after engine effects are removed. Consequently, the intercept is zero and only the coefficients of the 12 fuel variables are relevant.

The regression coefficients in the overall model (all 12 variables) are listed on the next page. The starred terms, by number [3 4 5 7 8 9 10], are the ones identified with X_{in}. Similarly, the remaining terms, by number [1 2 6 11 12], are the ones identified with X_{out}.

REGRESSION COEFFICIENTS

	Coeff.	t-value
Natcet	-0.0077	0.5779
Natcet^2	-0.0042	0.3110
CetDiff	-0.0289	6.1491*
CetDiff^2	0.0127	2.7813*
Arom	0.0324	5.2426*
Arom^2	-0.0098	1.8082
SpGrv	0.0106	3.1955*
Oxygen	0.0053	3.6766*
T10	0.0104	4.5187*
T50	-0.0104	3.0581*
T90	0.0021	0.9460
Sulfur	-0.0021	1.4289

Now, we break the original X-matrix and coefficient vector into two disjoint subsets consisting of the variables "in" and "out" of the subset model and proceed as follows. First, we compute the alias matrix:

$$A = (X_{in}^T * X_{in})^{-1} * X_{in}^T * X_{out} =$$

	Natcet	**Natcet^2**	**Arom^2**	**T90**	**Sulfur**
CetDiff	0.0329	0.0235	-0.0749	0.1113	0.4748
CetDiff^2	-0.0637	-0.0489	0.0275	-0.1542	0.3006
Arom	-0.2101	-0.1576	1.0503	0.2269	0.0556
SpGrv	-0.7527	-0.8028	-0.1354	-0.1902	0.1727
Oxygen	-0.0178	-0.0160	-0.0001	-0.0598	-0.0553
T10	-0.0857	-0.0841	0.0736	-0.3923	-0.2146
T50	0.6747	0.6504	0.0357	0.9702	0.2848

Then, the "new" coefficients for the "in" set can be computed as:

$$C_{new} = C_{in} + A * C_{out}$$

where C_{in} denotes the subset of retained coefficients, C_{out} denotes the subset of coefficients rejected, and C_{new} denotes the coefficients of the "new" model. The coefficients computed below are those that will be obtained when a separate regression is done using just the "in" subset of predictors.

$$Cin + A * Cout =$$

CetDiff	-0.0273
CetDiff^2	0.0122
Arom	0.0248
SpGrv	0.0203
Oxygen	0.0055
T10	0.0103
T50	-0.0173

If the transformation is written out *in extenso*, one will see just how the "new" coefficients are influenced by those rejected. To verify the above assertion, the following shows the regression equation with just the "in" variables.

Regression Coefficients

	Coeff	Std Err	t-value
Intercept	0.0000	0.0014	0.0000
CetDiff	-0.0273	0.0047	5.7550
CetDiff2	0.0122	0.0046	2.6244
Arom	0.0248	0.0023	10.6099
SpGrv	0.0203	0.0024	8.6138
Oxygen	0.0055	0.0015	3.7416
T10	0.0103	0.0021	4.8188
T50	-0.0173	0.0023	7.5659

Analysis of Variance

	SS	DF	MS
Mean	0.0000	1.0000	0.0000
Model	0.6775	7.0000	0.0968
Residual	0.4275	472.0000	0.0009
Total	1.1050	480.0000	

F-ratio = 106.8693

R-square = 0.6131

One can see that the alias computation yields the same regression coefficients as are obtained by regressing the response on just the seven variables retained. Whether we speak of "bias" depends on whether the rejected terms are actually non-significant or are artifacts of the p=0.05 significance level used to reject them. In any event, the rejected terms will show up computationally in the "new" equation. A main advantage of working in orthogonal space is that the alias matrix will be a null (zero) matrix, so that terms rejected from a model will not influence the coefficients estimated for the terms that are retained.

The following presents an example of the calculations to demonstrate how information from one variable is transferred to another as variables enter or leave the equation. The example is based on data from technology group T and shows the aliasing calculations for two variables: total aromatics content and specific gravity.

We begin our demonstration by performing an ordinary regression in which all 12 fuel property variables are included as predictors (see below). Because of the fact that these emissions are deviations from a mean value, they sum to zero and have zero intercept. As is usually done, those variables satisfying the 0.05 significance level are retained; all others are rejected. In this case 7 predictors are retained. The result is one of the 4095 possible subset models. We call your attention to three variables – aromatics, density and T50 – that appear to be much more significant in the subset model than in the full model.

REGRESSION PARAMETERS BEFORE AND AFTER VARIABLE SELECTION

	All 12 Variables		7 Selected Variables		
Fuel Property	Regression Coefficient	t	Regression Coefficient	t	
NatCet	-0.0077	0.58			
NatCet 2	-0.0042	0.31			
CetDif	-0.0289	6.15*	0.0273	5.76*	
CetDif 2	0.0127	2.78*	0.0122	2.62*	
Arom	0.0324	5.24*	0.0248	10.61*	<--
Arom2	-0.0098	1.81			
Sp Grav	0.0106	3.20*	0.0203	8.61*	<--
Ox	0.0053	3.68*	0.0055	3.74*	
T10	0.0104	4.52*	0.0103	4.82*	
T50	-0.0104	3.05*	-0.0173	7.57*	<--
T90	0.0021	0.95			
Self	-0.0021	1.43			

Just what caused the change in regression coefficients and their apparent significance for these three variables? Evidently, these variables benefitted from the deletion of the five variables: NatCet, NatCet2, Arom2, T90 and Self. Exactly *ho*w this benefit comes about will be shown below.

WHAT CAUSED THE CHANGE IN AROM, SPGRAV, AND T50?

	---- Regression Coefficients ----		
Fuel Property	All Variables	Variables Eliminated	Variables Retained
NatCet	-0.0077	-0.0077	
NatCet 2	-0.0042	-0.0042	
CetDif	-0.0289		-0.0273
CetDif 2	0.0127		0.0122
Arom	0.0324		0.0248
Arom 2	-0.0098	-0.0098	
SpGrav	0.0106		0.0203
Ox	0.0053		0.0055
T10	0.0104		0.0103
T50	-0.0104		-0.0173
T90	0.0021	0.0021	
Self	-0.0021	-0.0021	

The change in the value of the retained coefficients are computed explicitly by means of a quantity called the *alias* matrix or, for reasons to be shown later, sometimes the *bias* matrix. Its effect is that each of the rejected coefficients is given a weight specific to each of the retained coefficients. The weighted sum of the rejected coefficients is then added to the retained coefficient as initially computed when all 12 variables were in the model. It will be seen that the added portion is exactly equal to the difference between the coefficients for the subset model and for the full model.

The following chart shows how the aromatic coefficient changes from 0.0324 in the full model to 0.0264 in the subset model, Note that the major source of the change is a contribution to Arom from the Arom2 term. Thus what we originally thought was an aromatic effect is now “aliased” with its square term.

SOURCE OF CHANGE IN COEFFICIENTS: Aromatics

	Model		Subset		Change
	0.0248	-	0.0324	=	-0.0076
	Coeff.		Wt.		Product
NatCet	-0.0077		-0.2101		0.0016
NatCet2	-0.0042		-0.1576		0.0007
Arom2	-0.0098		1.0503		-0.0103
T90	0.0021		0.2269		0.0005
Self	-0.0021		0.0556		-0.0001
	Total change				-0.0076

The case is not so simple for SpGrav. Here the coefficient for density is almost doubled in going from the full model to the subset model. This change is primarily attributable to NatCet and its square term. What we originally thought was the effect of density, therefore, is now aliased with the effect of natural cetane and it square.

The generalization to be made here is that similar modifications in coefficients will be made for *any* subset of the predictor variables. Relative to the full model, the coefficients in the subset model may be said to be biased, and it is for this reason that the transforming matrix is sometimes called the bias matrix. It should be noted, also, that the coefficients for any subset model can be computed directly from the coefficients for the full model without performing the least squares procedure for the subset.

SOURCE OF CHANGE IN COEFFICIENTS: SpGrav

	Model		Subset		Change
	0.0203	-	0.0106	=	0.0097

	Coeff.	Wt.	Product
NatCet	-0.0077	-0.7527	0.0058
NatCet2	-0.0042	-0.8028	0.0034
Arom2	-0.0098	-0.1354	0.0013
T90	0.0021	-0.1902	-0.0004
Self	-0.0021	0.1727	-0.0004
	Total change		0.0098

APPENDIX D

PARTITIONING THE MODEL SUM OF SQUARES

APPENDIX D. PARTITIONING THE MODEL SUM OF SQUARES

D.1. INTRODUCTION

Perhaps the most ubiquitous and enduring concern regarding the "worth" of a regression equation is the extent to which that equation "explains" the responses that constitute the data set under consideration. Traditionally, the approach to this concern has been to partition the "total" sum of the squared responses into two parts: the "model" SS and the residual or "error" SS. For succinctness these measures are customarily coalesced into a single measure usually referred to as R^2, R-Square or the Coefficient of Determination.

Let X be the "design matrix" of predictor variables and let y be the vector of observed responses. Then the response y can be written as

$$y = Xb + e = y_c + e \tag{1}$$

where b is the vector of regression coefficients obtained from fitting y to X by OLS and e is a vector that expresses the difference between y, the responses as observed, and y_c, the responses as calculated from the regression equation. Without loss of generality it can be assumed that y is "centered" – that is, has zero mean – and that X is "standardized" so that each predictor variable has zero mean and unit standard deviation.

The total SS can be written as

$$\begin{aligned} y'y &= (y_c + e)'(y_c + e) \\ &= y_c'y_c + 2\,y_c e + e'e \end{aligned} \tag{2}$$

Under the assumption that the calculated responses y and the residuals e are uncorrelated, one can concentrate on the squares of the calculated responses and how they depend on the several predictor variables that make up the regression equation. Of particular interest is the relative importance of each of the predictors in the prediction process. It is for this reason that the statistical analyst is not content just to partition error SS from model SS. Rather, the analyst seeks to partition the model SS into components associated with each of the predictor variables, with a view toward eliminating predictors that do not contribute significantly and substantially to the model SS.

D.2. ERROR PROPAGATION IN A LINEAR FUNCTION OF RANDOM VARIABLES

The regression equation fitted to a set of data takes the form of a linear combination of the predictor variables. For demonstration purposes, consider a linear function of the three variables x_1, x_2 and x_3:

$$z = a_1 x_1 + a_2 x_2 + a_3 x_3$$

in which a_1, a_2 and a_3 are constants and x_1, x_2 and x_3 are assumed to be random variables. Then,

$$\begin{aligned} \mathrm{var}(z) &= a_1^2\,\mathrm{var}(x_1) + a_2^2\,\mathrm{var}(x_2) + a_3^2\,\mathrm{var}(x_3) \\ &\quad + 2\,a_1a_2\,\mathrm{cov}(x_1x_2) + 2\,a_1a_3\,\mathrm{cov}(x_1x_3) + 2\,a_2a_3\,\mathrm{cov}(x_2x_3) \end{aligned} \tag{3}$$

If the covariances are all zero, then var(z) is simply the sum of three variances and strict additivity among the sums of squares is assured. Thus, we can write:

$$SS(z) = \text{SS attributed to } x_1 + \text{SS attributed to } x_2 + \text{SS attributed to } x_3 \quad (4)$$

Note, however, that unless the design matrix X is column-wise orthogonal, no partitioning of effects is possible without taking into account the covariances among the predictor variables. It is for this reason that such attempts face the inevitable "non-additivity" of the model SS. The implications of this non-additivity has been discussed extensively in the previous report ORNL/TM-2000/5 under the topic "The Many Faces of Sums of Squares."

D.3. SS PARTITIONING AS A VARIABLE-SELECTION PROCEDURE

Inasmuch as there exists no "true" partitioning of the model SS when the termwise components are not additive, any attempt at such partitioning must be, in a certain sense, definitional. That being the case, it is perhaps more logical to view such attempts as means for variable selection rather than as means for SS partitioning *per se*. We take the liberty, therefore, of advancing two candidate approaches to variable selection. One is based on reformulation of the regression analysis by means of our version of Principal Components Regression, referred to elsewhere as PCR+. The other is based on a procedure by which covariance effects are "distributed" among the variances in such a way that the effects of the covariances are "even-handed" so far as the predictor variables are concerned. The algorithms for the two approaches are designated Simplify.m and SortSS.m, respectively, and are listed in Section D.6 Reference. Though written in Matlab code, they can be readily implemented in any other programming language.

D.3.1 SS Partitioning by Simplify.m

In effect, this algorithm combines the effects of predictor variables according to how they participate in the principal component contributions to response. Specifically, it takes into consideration the:

- Regression coefficients for the principal components
- Eigenvalues for the principal components
- Eigenvector component applicable to each of the property variables.

As shown in the error propagation exercise in Appendix B (Addendum IV), the squares of the regression coefficients play an important role in determining the relative importance of the principal components. The eigenvalues, being simply the variances of the principal components, also clearly play a role because they quantify the range and distribution of the principal components in the data set. Our concern, however, is with expressing the model SS in terms of the original property variables rather than in terms of the principal components. It is at this juncture that the components of the eigenvectors come into play.

The procedure can best be understood by way of numerical demonstration. For this purpose, we invoke the three-variable data set from our previous report ORNL/TM-2000/5 (see Table D.1).

TABLE D.1. DEMONSTRATION DATA BASE

"Raw" Variables				Standardized x-variables			Centered y
w_1	w_2	w_3	z	xt_1	xt_2	x_3	yc
7	4	3	11.0	0.0656	0.3162	-0.7482	1.68
4	1	8	3.2	-1.9030	-1.5811	1.0332	-6.12
6	3	5	5.1	-0.5906	-0.3162	-0.0356	-4.22
8	6	1	19.1	0.7218	1.5811	-1.4608	9.78
8	5	7	9.5	0.7218	0.9487	0.6769	0.18
7	2	9	5.6	0.0656	-0.9487	1.3895	-3.72
5	3	3	5.8	-1.2468	-0.3162	-0.7482	-3.52
9	5	8	11.7	1.3781	0.9487	1.0332	2.30
7	4	5	8.0	0.0656	0.3162	-0.0356	-1.32
8	2	2	14.2	0.7218	-0.9487	-1.1045	4.88
Means							
6.9	3.5	5.1	9.32	0	0	0	0
Standard Deviations							
1.52	1.58	2.81	4.8444	1	1	1	4.8444

For simplicity, and without loss of generality, we regress the centered response data on the standardized predictor variables.

TABLE D.2. Regression of Standardized and Centered Data

	Regression Coefficient	Coefficient Std Error	t-Ratio
Intercept	0.0000	0.7188	0.0000
xt_1	3.0123	1.0275	2.9316
xt_2	0.4743	1.0674	0.4443
xt_3	-2.5598	0.7978	3.2085

	SS	DF	MS	R^2
Mean	0.00	1	0.00	
Model	180.21	3	60.07	0.8532
Error	31.00	6	5.17	
Total	211.22	10	21.12	

Partitioning of the effects of the three predictor variables is complicated by the correlations among them, as shown in the following correlation matrix.

	x_1	x_2	x_3
x_1	1.0000	0.6687	-0.1013
x_2	0.6687	1.0000	-0.2879
x_3	-0.1013	-0.2879	1.0000

Accordingly, we compute the eigenvectors and eigenvalues of this correlation matrix and re-express the problem in terms of principal components and regress the centered response on those components.

Eigenvector:	**1**	**2**	**3**	
	0.6420	-0.3847	0.6632	
	0.6864	-0.0971	-0.7207	<-- Eigenvector matrix v
	-0.3417	-0.9179	-0.2017	
Eigenvalues:	1.7688	0.9271	0.3041	<-- Eigenvalue vector d

Then, by means of the transformation

$$x_{eig} = xt * v$$

we obtain the columnwise orthogonal principal components, as listed in Table D.3. Table D.4 presents the results of the PCR+ regression. The SS partitions for pc1, pc2 and pc3 were determined by regressing the response on each of the principal components individually.

TABLE D.3. DEMONSTRATION DATA BASE EXPRESSED AS PRINCIPAL COMPONENTS

	pc1	pc2	pc3	yc
	0.5148	0.6308	-0.0335	1.6800
	-2.6600	-0.0628	-0.3309	-6.1200
	-0.5840	0.2906	-0.1566	-4.2200
	2.0478	0.9096	-0.3663	9.7800
	0.8833	-0.9912	-0.3415	0.1800
	-1.0838	-1.2086	0.4471	-3.7200
	-0.7619	1.1971	-0.4481	-3.5200
	1.1828	-1.5707	0.0218	2.3800
	0.2713	-0.0233	-0.1772	-1.3200
	0.1897	0.8283	1.3852	4.8800
Mean	0.0000	0.0000	0.0000	0.0000
Std.Dev.	1.3300	0.9628	0.5515	4.8444

TABLE D.4. REGRESSION OF CENTERED RESPONSE ON PRINCIPAL COMPONENTS

	Regression Coefficient	Coefficient Std Error	t-Ratio
Intercept	0.0000	0.7188	0.0000
pc1	3.1340	0.5697	5.5008
pc2	1.1449	0.7870	1.4548
pc3	2.1722	1.3740	1.5810

	SS	DF	MS	R_2
Mean	0.0000	1	0.0000	
Model	180.2114	3	60.0705	0.8532
pc1	156.3596	1	156.3596	
pc2	10.9360	1	10.9360	
pc3	12.9157	1	12.9157	
Error	31.0046	6	5.1674	
Total	211.2160	10	21.1216	

Let us now consider the contributions of each of the property variables xt_1, xt_2 and xt_3 to the response SS. It should be evident that if only *one* eigenvector is involved in the regression, then the disposition of the model SS among the property variables is completely determined by the components of that eigenvector. Moreover, because the eigenvector matrix is orthonormal, the SS of the components for each of the eigenvectors is unity, and, by multiplying by 100, we obtain the percent contributions of variables xt_1, xt_2 and xt_3 to the response.

These results for each of the three principal components are given in Table D.5 below.

TABLE D.5. SS PARTITIONING FOR COMPONENTS OF INDIVIDUAL EIGENVECTORS

	pc1		pc2		pc3	
	SS	% SS	SS	% SS	SS	% SS
xt_1	64.4467	41.2170	1.6182	14.7973	5.6811	43.9857
xt_2	73.6598	47.1092	0.1032	0.9434	6.7094	51.9473
xt_3	18.2531	11.6738	9.2146	84.2593	0.5253	4.0669
Sum	156.3596		10.9360		12.9157	

The above computations were performed by Simplify.m under the restriction that only single eigenvectors are involved in the regression. Note that the sums for each of the principal components agree with those obtained by separate regressions in Table D.4. It is evident, also, that for pc1 and pc3 variables xt_1 and xt_2 make major and nearly equal contributions, whereas for pc2 it is variable xt_3 that dominates.

The algorithm Simplify.m can be applied to any subset of the three principal components. In Table D.6 the partitions are exhibited for the subsets: {pc1 pc2}, {pc1 pc3} and {pc2 pc3}.

TABLE D.6. SS PARTITIONING FOR COMPONENTS OF EIGENVECTORS SUBSETS

	----{pc1 pc2}----		----{pc1 pc3}----		----{pc2 pc3}----	
	SS	% SS	SS	% SS	SS	% SS
xt_1	45.6405	27.2813	108.3967	64.0357	1.2352	5.1787
xt_2	68.2495	40.7957	35.9075	21.2125	8.4766	35.5385
xt_3	53.4057	31.9229	24.9712	14.7518	14.1400	59.2827
Sum	167.2956		169.2754		23.8517	

Finally, it is of interest to see how the property variables compare when all three principal components are retained in the model, as shown in Table D.7.

TABLE D.7. SS PARTITIONING FOR COMPONENTS WHEN ALL EIGENVECTORS ARE INCLUDED

	SS	% SS
xt_1	83.5264	46.3491
xt_2	32.1611	17.8463
xt_3	64.5239	35.8046
Sum	180.2114	100.0000

According to the above partitioning, the importance of the three property variables are ranked in the order: $xt_1 > xt_3 > xt_2$.

The mathematical basis of Simplify.m is as follows. Let v_j be the j^{th} eigenvector, let v_{ij} be the i^{th} component of that eigenvector, and let d_j be the square root of the j^{th} eigenvalue. Further, let c_j be the regression coefficient for the j^{th} eigenvector. Then for the i^{th} component, we compute the sum of squares SS_i as

$$SS_i = (c_1 v_{i1} d_1 + c_2 v_{i2} d_2 + ... + c_k v_{ik} d_k)^2 \quad (5)$$

where k denotes the total number of eigenvectors included in the model.

A word of explanation is in order concerning the Simplify.m procedure. First, it should be recognized that other partitionings of the SS are possible, as will be shown subsequently. However, it can be said that the partitioning among *property variables* by Simplify.m is one that is consistent with the partitioning among *principal components*, regardless of what subset of the principal components is retained in the model. When SS_i is summed for all components, it is found that the contributions exactly equal the total model SS as computed by regression on either the P-Space variables or the principal components. In addition, the procedure is computationally, rather than judgmentally implemented.

D.3.2 SS Partitioning by Means of SortSS.m

The algorithm SortSS.m provides a complementary approach to the partitioning of effects in a regression model. It incorporates covariant effects directly and yields a break-down capable of providing insight into aliasing effects that influence variable selection.

The logic behind SortSS.m is an extension of Equation (3), repeated below for reference:

$$\text{var}(z) = a_1^2\,\text{var}(x_1) + a_2^2\,\text{var}(x_2) + a_3^2\,\text{var}(x_3)$$
$$+ 2\,a_1a_2\,\text{cov}(x_1x_2) + 2\,a_1a_3\,\text{cov}(x_1x_3) + 2\,a_2a_3\,\text{cov}(x_2x_3) \qquad (3)$$

Interpreted generally, this equation admits a generalization that partitions effects not simply *three* ways but *six* ways, and correspondingly more ways if more than three variables are involved. It advances what is mathematically evident – namely, that the response variable may depend on synergistic or anti-synergistic effects of pairs of variables as well as on single variables acting alone. Indeed, it is possible that these combined effects may be as important, or even more important, than some of the single variables. Moreover, inasmuch as covariances can be negative, there may be negative partitionings which, none the less, make for a decomposition of effects that is strictly "additive" in the algebraic sense.

One can take the liberty, however, of "distributing" the covariant effects among the non-covariant effects, and it is conceivable that this distribution could be performed in various ways. For example, one might agree to "distribute" the covariances among the variances in such a way that the variance associated with a particular variable x_i "absorbs" those covariances x_{ij}, j not equal to i – that is, all those covariances in which the variable x_i is involved. For the three-variable problem, the results, which we shall refer to as *quasi-partitionings*, are as follows:

Component 1: $a_1^2\,\text{var}(x_1) + a_1a_2\,\text{cov}(x_1x_2) + a_1a_3\,\text{cov}(x_1x_3)$

Component 2: $a_2^2\,\text{var}(x_2) + a_1a_2\,\text{cov}(x_1x_2) + a_2a_3\,\text{cov}(x_2x_3)$

Component 3: $a_3^2\,\text{var}(x_3^2) + a_1a_3\,\text{cov}(x_1x_3) + a_2a_3\,\text{cov}(x_2x_3)$

These three quantities are additive – i.e., they produce a partitioning of the SS that adds up to the total SS – and can be thought of as providing an "even-handed" partitioning of the model SS in which the covariance between two variables is evenly shared by those two variables.

The algorithm SortSS.m breaks down the contribution of effects into single components for each of the covariant effects and for each of the non-covariant (pure square) effects in such a way that they can be re-combined as desired. The even-handed recombination is especially easy to implement and is considered one of the most meaningful. Of particular interest is the fact that SortSS.m is applicable either to the property-variable (non-orthogonal) regression or to the principal-component (orthogonal) regression.

First, we will demonstrate the algorithm as applied to the property data xt_1, xt_2 and x_3. To invoke the algorithm one needs to know only the X matrix and the regression coefficients. The result is a square matrix, in this instance of size 3 x 3. The diagonal elements are the contributions caused by non-covariant (pure square) effects; off-diagonal terms correspond to effects produced by covariances:

	xt_1	xt_2	xt_3
xt_1	81.6655	8.5981	7.0308
xt_2	8.5981	2.0247	3.1462
xt_3	7.0308	3.1462	58.9711
Sum	97.2944	13.7690	69.1481
Percent	53.9890	7.6405	38.3705

The "distributed" or "even-handed" effects are obtained simply by adding rows or columns, and the effects produced by non-covariant (pure square) terms are obtained simply by summing the diagonal.

Purely covariant effects can be determined directly or by difference and are tabulated below under the heading x-prods. It is to be noted the values tabulated represent only *half* of the covariant effects, because these effects are divided equally between the two covariant properties:

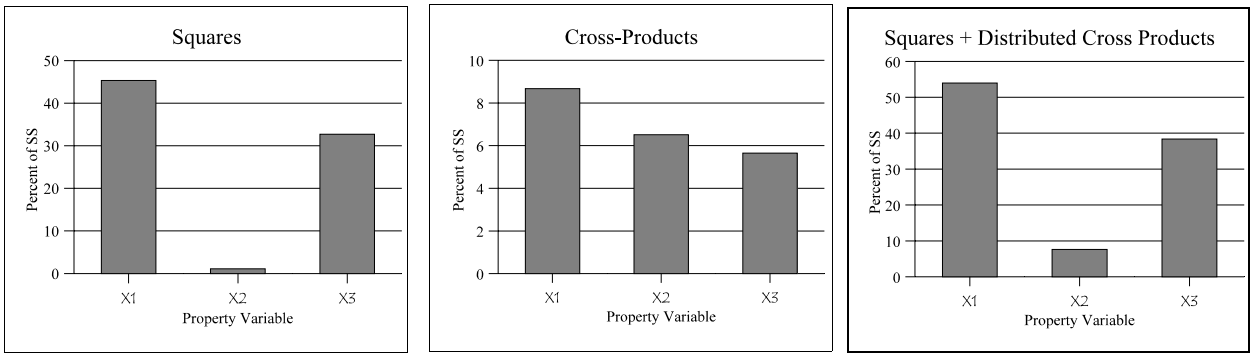
	Squares	**x-prods**	**Dist. xprods**
x_1	81.6655	15.6289	97.2944
x_2	2.0247	11.7442	13.7689
x_3	58.9711	10.1770	69.1481
Sum	142.6612	37.5501	180.2114
	(79.16%)	(20.84%)	(100.00%)

Percent of total SS and cross-prods attributed to individual variables:

	Squares	**x-prods**	**Dist. xprod**
% attributed to x_1	45.3165	8.6725	53.9890
% attributed to x_2	1.1235	6.5169	7.6404
% attributed to x_3	32.7233	5.6473	38.3706

Results are shown graphically in Figure D.1.

Figure D.1. Sum of Squares Partitioning by SortSS.m Using Demonstration Data



If the variables are ranked only on the basis of their variances (i.e., without considering the covariances), one sees that x_2 appears to have relatively little influence on response. When the covariances are distributed among the sums of squares, that distribution has relatively little effect, percentage-wise, on x_1 and x_3 but a relatively large effect on x_2. This fact reflects the large proportion of the effect of x_2 that is attributed to covariance.

How the algorithm performs in the case of an orthogonal X-matrix is illustrated by application of the algorithm to the principal-component version of the demonstration data.

	pc1	pc2	pc3
pc1	156.3596	-0.0000	-0.0000
pc2	-0.0000	10.9360	-0.0000
pc3	-0.0000	-0.0000	12.9157
Sum	156.3596	10.9360	12.9157
Percent	86.76	6.07	7.17

Comparison of the orthogonal and non-orthogonal forms of the data makes it evident that a particular design matrix need not be severely "ill-conditioned" in order for covariance effects to play a substantial role in variable selection.

D.4. SS PARTITIONING OF REAL DATA: EMISSIONS FOR TECH GROUP T

Analysis of engine-corrected data for Tech Group T emissions provides a meaningful demonstration of how SS partitioning can provide insights for variable selection.

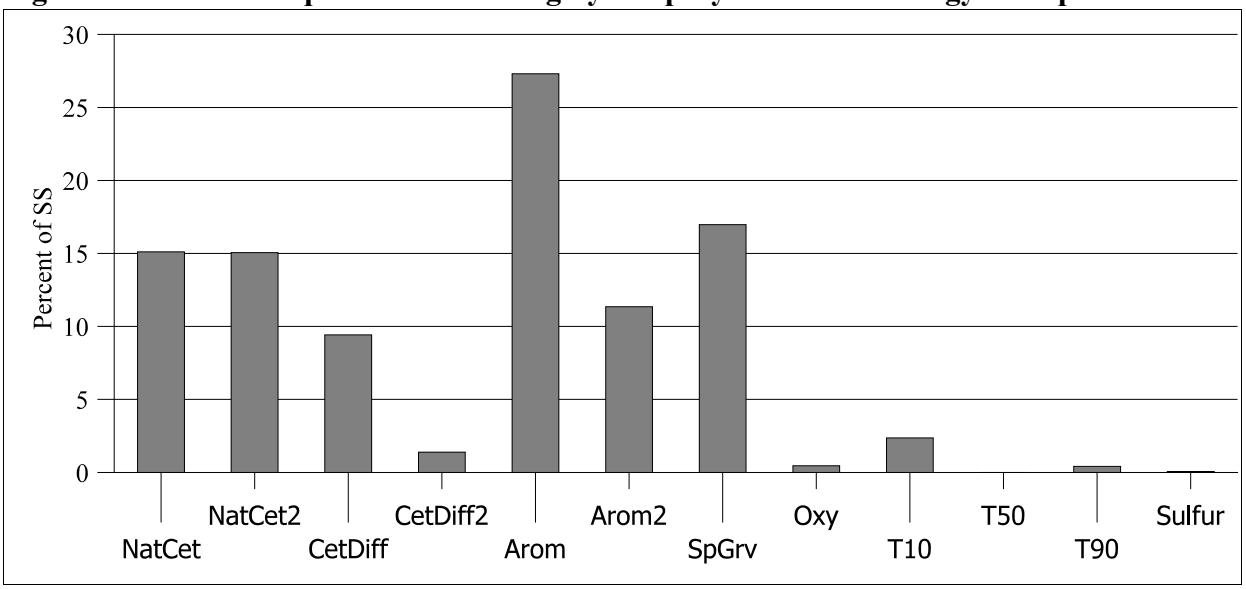
D.4.1 Partitioning of Tech Group T Data by Simplify.m

Table D.8 provides the partitioning of emission data for Tech Group T by means of the Simplify.m algorithm. Results are shown graphically in Figure D.2. According to this partitioning, NatCet plays an important role, contrary to the findings reported by EPA in their final model.

TABLE D.8. SS PARTITIONING OF TECH GROUP T DATA

NatCet	0.1062	15.1130
NatCet2	0.1059	15.0677
CetDif	0.0663	9.4370
CetDif2	0.0098	1.4011
Arom	0.1920	27.3149
Arom2	0.0798	11.3593
Dens	0.1194	16.9831
Oxy	0.0033	0.4672
T10	0.0167	2.3736
T50	0.0000	0.0049
T90	0.0029	0.4114
Sulfur	0.0005	0.0670
Sum	0.7029	100.0000

Figure D.2. Sum of Squares Partitioning by Simplify.m for Technology Group T



D.4.2 Partitioning of Tech Group T Data by SortSS.m

Because of its format and capabilities, SortSS.m can be applied either to the non-orthogonal form of the data or to the data as reformulated in terms of principal components.

D.4.2.1 Partitioning of Property Variables

Application of the SortSS.m algorithm to property data for Tech Group T is summarized in Table D.9 below and are shown graphically in Figure D.3. It is seen that CetDif and Arom dominate when viewed in terms of sums of squares only, but when covariances are distributed, the NatCet effect appears comparable to that of CetDif, and both Arom and SpGrv take greater prominence.

D.4.2.2 Partitioning for Principal Components

After the eigenvectors and eigenvalues for the Tech Group T data are obtained, the data are converted from fuel properties to principal components. When the algorithm SortSS.m is applied to the derived orthogonalized data, the results shown in Table D.10 are obtained.

Figure D.3. Sum of Squares Partitioning by SortSS.m for Technology Group T

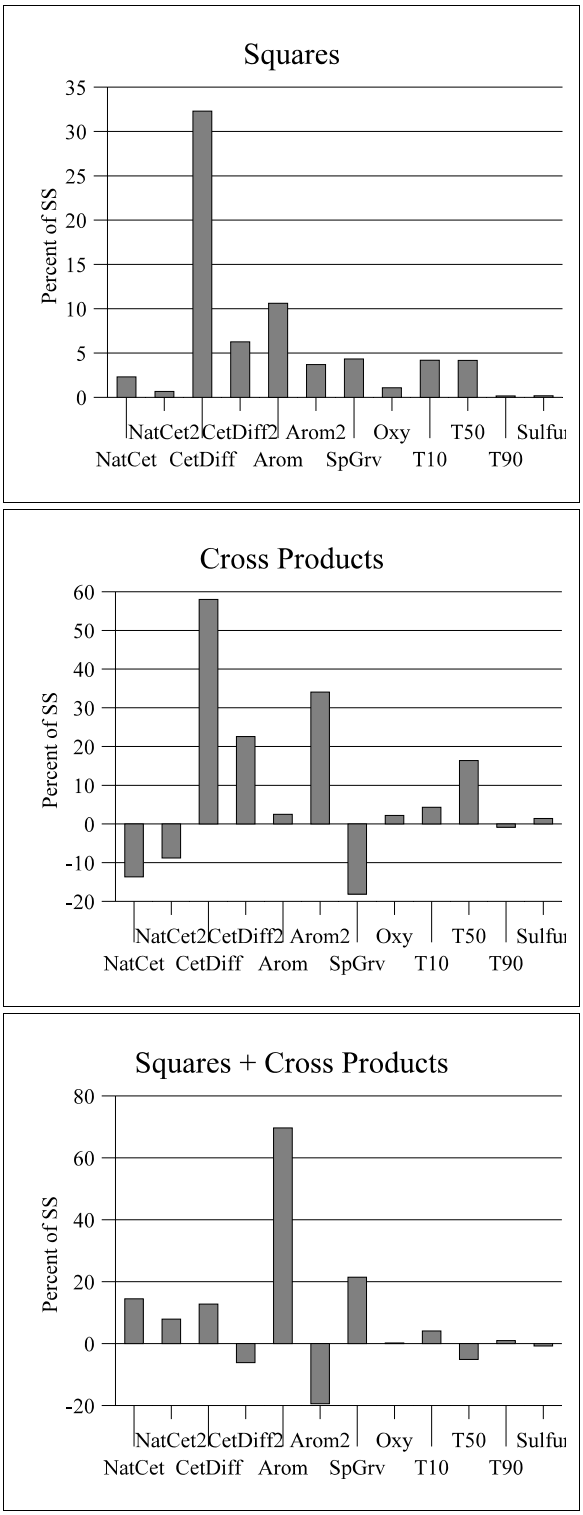


TABLE D.9. SS PARTITIONING OF TECH GROUP T DATA

Sums of Squares, Cross-Products and Distributed Products

	Pure Squares	Cross-Products	Dist. x-prods	% Squares	% x-prods	% Dist.
NatCet	0.0286	0.0731	0.1017	2.3124	-13.6583	14.4758
NatCet^2	0.0085	0.0470	0.0555	0.6840	-8.7868	7.8970
CetDif	0.4002	-0.3106	0.0895	32.3178	58.0241	12.7395
CetDif^2	0.0774	-0.1208	-0.0434	6.2509	22.5709	-6.1786
Arom	0.5030	-0.0133	0.4897	40.6220	2.4846	69.6678
Arom^2	0.0458	-0.1823	-0.1365	3.7007	34.0601	-19.4213
Sp Grv	0.0535	0.0972	0.1507	4.3227	-18.1514	21.4392
Oxygen	0.0134	-0.0117	0.0016	1.0804	2.1927	0.2332
T10	0.0519	-0.0232	0.0287	4.1916	4.3399	4.0786
T50	0.0517	-0.0877	-0.0360	4.1781	16.3814	-5.1162
T90	0.0021	0.0046	0.0067	0.1676	-0.8682	0.9565
Sulfur	0.0021	-0.0076	-0.0054	0.1720	1.4109	-0.7716
Total	1.2382	-0.5353	0.7029	100.0000	100.0000	100.0000

Note, as expected, that all cross-products vanish. Note, also, that principal components #1 and #3 dominate.

TABLE D.10. PARTITIONING FOR TECH GROUP T ORTHOGONAL DATA

Sums of Squares and Cross-Products

	Squares	x-prods	Dist.xprod.	Percent*
pc1	0.4402	0.0000	0.4402	62.6327
pc2	0.0006	-0.0000	0.0006	0.0908
pc3	0.1919	-0.0000	0.1919	27.3025
pc4	0.0008	-0.0000	0.0008	0.1201
pc5	0.0230	-0.0000	0.0230	3.2787
pc6	0.0043	0.0000	0.0043	0.6127
pc7	0.0016	-0.0000	0.0016	0.2291
pc8	0.0030	-0.0000	0.0030	0.4280
pc9	0.0087	-0.0000	0.0087	1.2315
pc10	0.0117	-0.0000	0.0117	1.6628
pc11	0.0169	-0.0000	0.0169	2.4030
pc12	0.0001	0.0000	0.0001	0.0083
Total	0.7029	-0.0000	0.7029	100.0000

*Percent applies to squares and distributed column since all cross-products are zero

D.5. RECOMMENDATION

It is recommended that the algorithm Simplify.m be used as the primary discriminant for variable selection, complemented by the use of SortSS.m for further diagnostic insight.

D.6. REFERENCE

D.6.1 Matlab Code for Simplify.m

```
function y=Simplify(vecref,valref,coef,s,n)
%Computes the SS contributions of the components of eigenfuels
%to the total contribution of any subset of the complete set of eigenfuels
%vecref is the complete set of eigenvectors - i.e., the eigenvector matrix
%valref is the complete set of eigenvalues, displayed as a column vector
%coef is the set of coefficients in a complete regression of the response
%variable on the eigenvectors, expressed as a column vector,
%the constant vector (intercept) excluded.
% s is a row vector listing the eigenfuels to be retained in the model.
%It can denote a single vector, and any combination of vectors
%up to and including the complete set.
%Note: the order of listing eigenvectors, eigenvalues and coefficients
%is arbitrary, so long as it is kept consistent throughout the computation.
% n is the number of cases.
[h,k]=size(coef)
    [a,b]=size(s)
    vec=vecref(:,s)
    val=valref(s)
    coef=coef(s)
    z=zeros(h,1)
for i=1:b,
    c=coef(i)
    cstar=c*vec(:,i)
      e=val(i)
      x=cstar*sqrt(e)
        z=z+x
    end
y1=(n-1)*z.*z
sumy1=sum(y1)
p=100*y1/sumy1
y=[[y1;sumy1] [p;sum(p)]]
```

D.6.2 Matlab Code for SortSS.m

```
function y=SortSS(x,c)
%Computes contributions to model SS by variances and covariances
%of the predictor variables.
[rr,cc]=size(x);
for i=1:cc,
    for j=1:cc,
      qi=x(:,i)*c(i);
```

```
        qj=x(:,j)*c(j);
        y(i,j)=qi'*qj;
    end
end
% This matrix can be used to compute:
% Partition of distributed sums of squares and cross-products
% Partition of sums of squares only
% Partition of cross-products only
% Percent of each by variables
% Percent of total partition due to squares
% Percent of total partition due to cross-products
```

APPENDIX E

EXPERIMENT DESIGN AND DATA ANALYSIS

APPENDIX E. EXPERIMENT DESIGN AND DATA ANALYSIS

E.1 INTRODUCTION

So widely acknowledged is the intimate relationship between experiment design and data analysis that it has become nothing less than a statistical mantra. This appendix presents this relationship as a formal duality in which experiment design and data analysis are bonded by a common calculus.

E.2. BACKGROUND

Recent studies of the effects of fuel characteristics on heavy duty diesel emissions have been documented by an SAE paper and, in further detail, as a U. S. Department of Energy report [Ref 1,2]. These studies are distinguished by their representation of fuel characteristics as *composite* quantities, or *vectors*, rather than as single, scalar quantities. A powerful adjunct of the data analysis procedures is that the analysis is performed in a vector space in which the composite predictor variables are orthogonal.

Orthogonalization of the predictor variables gives rise to new opportunities for experiment design, but it also introduces new challenges. Most importantly, it merges experiment design and data analysis in a way not previously perceived and, in doing so, brings new insights and interpretations to bear on multivariate statistical analysis.

E.3. FUNDAMENTAL METHODOLOGICAL PRINCIPLES

We begin with a fundamental postulate: for *any* data set with a design matrix of full rank, there exists an orthogonal basis consisting of combinations of the original predictor variables. Data analysis, such as multilinear regression analysis, is preferably performed in this orthogonalized space, primarily because the transformation removes all aspects of multicollinearity that may have existed in the original predictor variables. Thus one might say that, in a certain sense, *every* experiment is designed orthogonally: we have only to find its orthogonal basis. Once that basis is defined, the experiment can be analyzed *as if* it were designed as an orthogonal array at the outset. This array, of course, consists of the vectors that emerge as the orthogonal basis, and we accept those vectors as redefined variables.

Procrustean though this approach may seem, it is no more so than the conventional approach, in which each variable is treated as if it is a separate entity, even though that variable may be hopelessly entangled with one or more other variables.

At the outset, an investigator has two options. He may arbitrarily choose a set of variables believed to be the pertinent ones and then lay out a set of treatments that explores the predictor-variable space. His purpose is to select a set of treatments in such a way that the effects of each of the variables can be independently estimated. Alternatively, he may define his treatments *arbitrarily* and redefine the variables in such a way that the effects of each of the *redefined* variables can be independently estimated. In this sense, the two approaches can be said to be *duals* of each other. In the one case, the experimenter attempts to untie the knots that tie the variables together. In the other case, he accepts the knots as they are.

Though capable of being interpreted philosophically as well as mathematically, duality can be explained sufficiently for our purposes by means of a simple geometric example. For the statement "Two points determine a straight line" there exists the dual statement "Two straight lines determine a point." More generally, it can be said that "N points determine K straight lines" and that "N straight lines determine K points." The parameters N and K are interchangeable in the two contexts. Correspondingly, we may say that "Variables determine treatments" or that "Treatments determine variables." It has been custom to begin with variables and attempt to devise orthogonal treatments, rather than to begin with treatments and then attempt to devise orthogonal variables that make up those treatments.

The conventional approach, in which variables are used to determine treatments, grew out of experimental environments in which there was no difficulty in controlling the levels of one variable independently of the levels of another variable. For example, in agricultural experiments, in which the term "treatment" arose, one could apply various amounts of Fertilizer A to plots quite independently of applying the same or different amounts of Fertilizer B to those plots. If we let the notation $W = \{w_1\ w_2\ w_3\}$ denote the amounts of Fertilizer A applied to three plots and $V = \{v_1\ v_2\ v_3\}$ denote the amounts of Fertilizer B applied to the same plots, then one can define a set of treatments as the *cartesian product* of the two sets W and V:

$$W * V = \{w_1v_1\ w_1v_2\ w_1v_3\ w_2v_1\ w_2v_2\ w_2v_3\ w_3v_1\ w_3v_2\ w_3v_3\} \qquad (1)$$

Such a set of treatments allows the experimenter to determine the independent effect of each of the two fertilizers on crop yield, as well as certain other effects, known as *interactions* in the statistical literature.

The above example is clearly one in which the amounts of fertilizers A and B can be varied independently of each other. No matter how much fertilizer A is applied to a plot, that amount in no way constrains the amount of fertilizer B that can be applied to the same plot. When the experiment-design methodology is ported to another area of inquiry, however, situations may arise in which it is not possible or feasible to vary the levels of two variables independently, because there are natural forces that cause the variables to be correlated to greater or lesser extent.

An obvious case, and one somewhat similar to the fertilizer experiment, is presented in the design of *mixtures* of several materials, as in the formulation of fuel blends. If, for example, there are three components that together make up the mix, any increase in one component has to be accompanied by a corresponding decrease in one or both of the other two components, because collectively the three components must total 100%. The experimenter has *some* latitude of choice, therefore, inasmuch as he can choose which of the other two components, or what combination of the two, is to be displaced by the increase in the first component. He can *not*, however, choose the amounts of those ingredients arbitrarily.

Clearly, if two variables are *completely* correlated, there is no latitude at all for varying the level of the second variable once the level of the first variable is fixed. Similarly, if the two variables are completely *uncorrelated*, the experimenter has no constraints. What complicates the matter in partially correlated cases is the extent to which the disposition of levels of the several variables is volitional or involitional. In such cases the experimenter may attempt to "break" the association between variables, whereas in other instances he may set levels of the first variable but let the second variable "seek its own level," so to speak, by accepting whatever level that variable takes in normal refinery practice.

It is important to recognize, therefore, that when a collection of fuels is subjected to PCA, the definition of the characteristic "eigenvariables" may be at least partially an artifact of conscious design effort. A new experiment involving the same collection of fuels could very well lead to a somewhat different set of eigenvectors simply as a result of the experimenter's effort or lack of effort to attain a balanced experiment in the sense of Equation 1. Whether the experimenter does or does not attempt to "break" the association among variables, it is possible to isolate eigenvectors that are unique to the data set under consideration. The

problem of how to reconcile different sets of eigenvectors has been dealt with elsewhere [Ref 2], but the present discussion will afford further understanding of this seeming dilemma.

For present purposes it is assumed that a set of eigenvectors has been agreed upon and that we wish to perform additional experiments to explore further the specific effects of those eigenfuels.
In particular, we consider the process of experiment design as one in which the eigenvariables are held subject to the same constraints as would be the case for any original variable capable of being manipulated independently, as was the case for the separate fertilizers A and B. We tacitly assume that the transformed variables – what we call eigenvariables – are subject to such independent manipulation just as were the two fertilizers.

E.4. DEMONSTRATION AND APPLICATION OF PRINCIPLES

Rather than proceeding further with abstract development of theoretical principles, we elect to develop the essential concepts by demonstrating how these principles come into play in connection with the existing data set of diesel fuels. We begin at the point at which the fuels have been transformed into *eigenfuels*.

First, it will be informative to obtain an appreciation of the ranges and distributions of the coefficients of the eigenfuels. Statistics essential to this understanding are given in Table E.1.

Table E.1. Statistical Summary of Coefficients of Eigenfuels

Eigenfuel Number	Maximum Value	Median Value	Minimum Value	Standard Deviation
1	3.6797	0.1223	-5.3758	2.1285
2	5.1957	-0.0438	-3.5526	1.6096
3	2.5112	-0.1062	-2.8547	1.2446
4	4.2418	-0.1937	-3.2491	1.0866
5	1.3068	0.0016	-2.6806	0.8251
6	1.8804	0.0891	-1.8487	0.7507
7	2.5954	0.0228	-2.0217	0.6241
8	1.0957	-0.0114	-1.5340	0.4801
9	2.0249	-0.0360	-0.7922	0.3746
10	1.0460	0.0323	-0.6300	0.2873
11	0.4181	-0.0258	-0.5953	0.1895
12	0.4812	-0.0332	-0.6335	0.1604

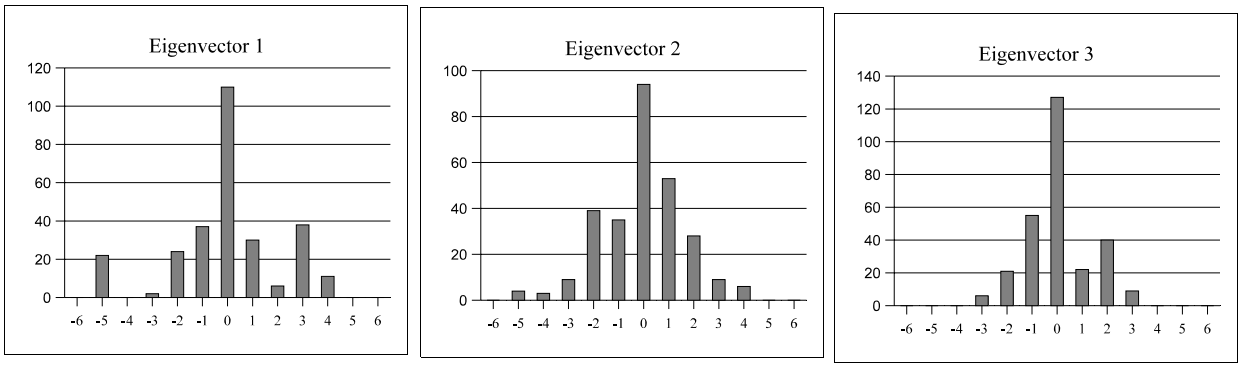
Inasmuch as the eigenfuels are arranged in decreasing order of their eigenvalues, it is seen that the range of these coefficients decreases as one moves downward in the table. Moreover, previous
regression analysis has shown that the first three eigenfuels are able to account for "most" of the response attributable to fuel characteristics. Though one might choose a different "cut-off point," it suffices, for present demonstration purposes, to confine our attention to the set of the three uppermost:

Eigenfuel No.	1	2	3
Maximum Coefficient	3.6797	5.1957	2.5112
Minimum Coefficient	-5.3758	-3.5526	-2.8547

These three are selected because, collectively, they were found to be the eigenfuels that account for "most" of the fuel influence on either NO_x or PM (particulate matter) or both.

Because these statistics were extracted from the actual fuel data set, one might conjecture that fuels having coefficients within these ranges are "realizable." However, an important caveat needs to be observed. Just because the coefficients of Eigenfuel 1 range from -5.3758 to +3.6797 overall, it can not be concluded that this range will hold uniformly over the range of coefficients for Eigenfuel 2 or Eigenfuel 3. Some insight relevant to the distribution of the eigenfuel coefficients is provided by their respective histograms in Figure E.1. However, these one-dimensional descriptions do not reveal how the coefficients of the three eigenfuels are *jointly* distributed. Cross tabulation, as is provided for in most statistical software, would provide such information, but would still not provide the insight necessary to structure an experiment design.

Figure E.1. Histogram of Eigenvector Coefficients for Fuels (280 Cases)



We suggest, instead, a somewhat different approach, one that has direct relevance to experiment design in the classical sense. Before proceeding further in this direction, however, a short tutorial regarding fuel subsets is in order.

E.4.1 Fuel Subsets and Orthogonalization

The fuel data set, expressed in terms of the coefficients of the eigenfuels, is a 280 by 12 array, the rows denoting the number of emission tests, the columns denoting the 12 eigenfuels. Note that, with regard to actual fuels tested, there is a considerable amount of redundancy, because the total number of fuels involved was considerably less than 280. Furthermore, recall that the eigenfuels were developed on the basis of the entire array of 280 entries, regardless of the fact that multiple tests may have been made of the same test fuel.

The resulting design matrix (often referred to as the X matrix) is column-wise orthogonal *because it was forced to be that way* by virtue of the principal-component analysis (PCA) employed in transforming the fuels from P-Space (the space of fuel properties) to E-Space (the space of eigenfuels). The transformation guarantees column-wise orthogonality *only if all fuels, duplicates included, are contained in the E-Space fuel data set*. It does not necessarily guarantee that any particular *subset* of the fuels will boast the column-wise orthogonality feature. Neither does it guarantee column-wise orthogonality for any given collection of fuels that might be generated by *blending* eigenfuels in various proportions. Just as in *any* experiment-design exercise, appropriate thought must be brought to bear on the process so as to produce a design having such desired properties as balance (orthogonality), appropriate *power* and so on. After all, orthogonality is a property associated with a *collection* of fuels and has no meaning for *individual* fuels.

In addition, the experiment-design process must assure that the resulting postulated blends are *realizable*, preferably at production scale. This question is considered, via demonstration, in the following parts of this

appendix. Though a question may be raised as to whether fuel properties blend linearly, the fact that the demonstration is consistent with the assumptions made in computing the eigenfuels argues in favor of the procedure. Further validation awaits further application of the methodology in real-world formulation of fuel blends.

E.4.2 Isomorphism: Basic Modeling Tool

Modeling is a means for studying how systems operate. It allows us to simulate the manipulation of elements of the system by substituting for those elements certain *alternative* elements that are much easier to manipulate. For modeling to be valid, however, there must be a one-to-one correspondence between like elements in the two systems as well as a one-to-one correspondence between the *manipulations* involved in the two systems. Note that the manipulations need not be the *same* in the two systems, only that the *results* of the manipulations must correspond in both systems.

The process described pedantically above is called *isomorphism* in mathematics. A simple example is the slide rule, or, more fundamentally, the correspondence between real numbers and their logarithms. The results of the *multiplication* of two numbers in number space is in direct one-to-one correspondence with the results of *addition* in logarithm space. The slide rule "takes it up a notch" by establishing an additional correspondence between logarithms and distances on the slide rule.

We proceed, now, to illustrate how these principles can be brought to bear on experiment design when the elements to be manipulated are *not* perceived fuel properties but weighted combinations of those properties known as eigenfuels.

E.4.3 How to Blend Fuels to Treatment Requirements

First, let us set up a very simple array of numbers that can be placed in one-to-one correspondence with a *balanced experiment design*, specifically a 2^3 factorial, defined as an experiment in which there are three variables, each set at two levels. The result, by taking the Cartesian product of the three sets of levels, is a set of eight *treatments*, as shown in Table E.2.

Table E.2. Binary Representation of a 2^3 Factorial Experiment

-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

This array is clearly columnwise orthogonal - that is, the inner product of any two columns is zero. Consequently, it serves as a model for a fuel subset having like characteristics if only such a subset, or "reasonable facsimile thereof" can be found among the fuels of the total data set or can be produced by blending fuels that are extant or are known to be producible.

To appreciate fully the general applicability of Table E.2 one needs to note that the two "levels," -1 and +1, can be taken as representing any two actual levels of real, physical quantities. For example, suppose that the

variable involved is degrees Celsius and that the two temperature levels are, numerically, 40 and 80. By the transformations (40 - 60)/20 and (60 - 40)/20 those two levels map into -1 and +1.

In a quite similar manner the eigenvector coefficients for eight fuels can be mapped into the array of Table E.2 by selecting two values for each eigenvector and mapping those values into {-1, +1}. Inasmuch as the actual coefficients -1 and +1 denote a representative range of coefficients in the fuel collection, we make our demonstration by choosing those actual values as the "levels" in our experiment. Note, however, that a more extreme or less extreme range could be employed simply by performing the necessary linear transformations as detailed above.

It is a straightforward matter to find, among the fuels of the overall data set, those eight fuels that most closely approximate the above array. One simply finds, for each treatment vector, that fuel that is the least-squares best fit from among the 280 fuels in the list. These are shown in Table E.3.

Table E.3. Approximation of a 2^3 Factorial Array By Fuels Selected from the Overall Fuel Data set

Fuel No.	Coefficient of Eigenfuel #1	#2	#3	Figure of Merit
217	-0.8829	-0.5011	-1.2429	0.5671
40	-1.4034	-0.2630	0.8515	0.8532
148	-1.7259	0.7117	-0.6660	0.8494
259	-0.2935	0.7253	1.4268	0.8699
213	0.5218	-0.2641	-1.1766	0.8952
124	0.7486	-0.9352	-0.0175	1.0501
226	0.8011	1.0553	-0.4712	0.5677
223	0.4434	1.2410	0.3738	0.8718

The FM is the squared deviation of a particular fuel from its "ideal" or "target" configuration given in Table E.2. The fuels listed are those with smallest FM, which, for perfect fit, would be zero. The "fit" of the candidate fuels to the target fuels is not very good but can be improved by blending certain fuels from among the 280 available in the fuel data set (see Table E.4).

It should be realized that the fuel combinations in Table E.4 are not the *only* ones that might be used in the blends. Indeed, multiple combinations could be used to create multiple blends, each of which satisfies the restrictions placed on the first three eigenvectors but which vary with regard to the levels of the remaining nine eigenvectors. In a sense, these multiple solutions could be considered as "replicates" from which one could obtain an estimate of the "error" resulting from the assumption that only the first three eigenvectors need to be considered.

Given that a set of fuels generates a particular "treatment," its authenticity can be easily verified, as will be demonstrated in the following discussion. For demonstration purposes, we use the first treatment blend of Table E.4, which is a mix of four fuels that have been selected to generate the treatment [-1 -1 -1] – that is, a mix such that the coefficient for each of the first three eigenfuels is exactly -1, as required by the array of Table E.2.

Listed below are the four fuels being blended, together with the coefficients for the first four eigenfuels for each. Tabulated also are the weight percentages of each fuel as proposed for the blend.

Fuel No.	Weight (%)	Coefficient of Eigenfuel #1	#2	#3	#4
43	6.14	-2.4925	-3.4119	2.4646	-0.2809
71	70.95	0.0227	-0.8238	-1.1357	-0.2019
176	16.44	-5.3020	-0.3521	-2.0348	-0.1123
114	6.47	0.1334	-2.2896	-0.1699	-0.1328

Then, the coefficient of Eigenfuel #1 *in the treatment blend* is computed as follows:

```
   (0.0614)*(-2.4925) + (0.7095)*(0.0227)
 + (0.1644)*(-5.3020) + (0.0647)*(0.1334)
 = -0.1530 + 0.0161 -0.8717 + 0.0086
 = -1
```

The "partial" contributions of each of the fuels to the coefficient of the eigenfuels in the blend are as follows:

Fuel No.	Coefficient of: Eig 1	Eig 2	Eig 3	Eig 4
43	-0.1530	-0.2095	0.1513	-0.0172
71	0.0161	-0.5844	-0.8058	-0.1432
176	-0.8717	-0.0579	-0.3345	-0.0185
114	0.0086	-0.1482	-0.0110	-0.0086
Sum	-1.0000	-1.0000	-1.0000	-0.1875

The other treatments in Table E.4 can be similarly verified.

How to compute the blend-component weights is not a matter of "smoke and mirrors," though it may appear to be somewhat circular. The components of the blend have to satisfy three constraints in order to obtain the right "levels" proposed for a given treatment. In addition, the component weights must sum to unity and every weight must be non-negative.

If it is known that a set of fuels is capable of meeting these constraints, computing the component weights is a straightforward matter of solving four equations in four unknowns. Unfortunately, we can not be assured that the equations are "compatible" until they have been solved – a "Catch 22."

How this apparent dilemma is resolved is set forth in detail in Addendum I. As shown there, finding a compatible set of four fuels may be a matter for Monte Carlo sampling. The number of combinations of 280 fuels taken four at a time is a big number, and to explore the regime exhaustively would be impractical. Instead, we take successive random samples of four fuels and compute the blending weights just as if they are legitimate. Actually, many solutions will not be. The proportions of the four fuels may add to unity only by virtue of the fact that some are negative, or the matrix of the equations may even be singular or very poorly conditioned. However, if the sum of the *absolute* weights is unity, then the mix is allowable and is tagged for retention as an admissible solution.

Much simplification is possible by eliminating duplicate fuels from the 280 x 12 list of "fuels." Many means are at our disposal for mapping the 280 x 12 array into a 280 x 1 array. One that is quite convenient is simply

to list not the 12 individual coefficients but the *sum* of those coefficients. Though it is possible that two or more fuels could have the same sum of coefficients, the possibility is remote. Accordingly, the sums were computed and sorted in ascending order. Only 76 distinct subsets emerged, the subsets being distinguished by different sums. For all fuels within each subset, each of the eigenfuel components was identical, *not* just their sums. A list of the 76 fuels obtained by this compression procedure is given in Addendum II. Any fuel in the original 280 has a correspond fuel among the 76 fuels isolated as distinct.

Table E.4. Treatment Blends in a 2^3 Factorial in Which Eigenfuels 1, 2, and 3 are the Controlled "Variables"

Treatment "Levels"	Fuels* Used in the Blend	Blend-Component Weights (%)
-1 -1 -1	43	6.14
	71	70.95
	176	16.44
	114	6.47
-1 -1 1	153	22.68
	7	35.83
	21	34.27
	161	7.22
-1 1 -1	179	19.31
	128	12.50
	141	12.68
	99	55.50
-1 1 1	262	75.58
	148	4.74
	43	12.11
	168	7.57
1 -1 -1	86	7.48
	116	47.50
	162	27.04
	142	17.99
1 -1 1	75	19.55
	31	25.47
	104	50.86
	52	4.11
1 1 -1	108	17.96
	45	19.72
	141	53.97
	175	8.35
1 1 1	213	7.47
	125	37.03
	266	50.89
	165	4.61

* Numbers refer to the test in which the fuel is used

The number of sets of four fuels that can be drawn from the total of 76 is still a big number but not big enough to overload a modern computer. Specifically, the number of combinations of n things take k at a time is

$$_nC_k = n!/(k!*(n-k)!)$$

which, for n=76 and k=4 is 1,282,975.

There are a number of ways in which these combinations can be generated. One involves four nested loops and is capable of providing an exhaustive examination of the eligibility of all possible combinations of four fuels drawn from the total collection of 76. An alternative approach abandons the notion of exhaustive search by making use of random sampling of the 76 fuels four at a time, as detailed in Addendum I.

E.5. SUMMARY AND CONCLUSIONS

It has been demonstrated that it is possible to blend fuels in such a way as to produce "treatment" blends that satisfy the requirements of factorial design. It is recognized that a given set of fuels may not be amenable to such blending, because one or more of the fuels may require negative amounts in the blend, an obvious impossibility. However, it has also been shown that, in a typical set of fuels, such as the 76 fuels in the data base, there is an acceptable probability that a set of four admissible fuels can be found among those fuels. Indeed, there are thousands of possible combinations that are acceptable (see Addendum I). The indication is, therefore, that blending producible fuels to meet experiment-design requirements is entirely feasible, given a "reasonable" number of source fuels from which to draw.

What is needed is to identify properties of the fuels that make them likely candidates and to get a feel for just how few fuels would suffice for an experiment-design collection. Certainly 76 is enough for a three-variable design, but, for favorably tailored fuels, a much smaller number should suffice. On the other hand, if one wants to control *more* than three composite properties (eigenfuels), the yield of admissible blends could be sparse, and that fact could necessitate having a sizable range of fuel properties from which to draw.

As noted earlier, in controlling three selected or "key" variables no attempt is made to control the remaining eigenvectors. Nevertheless, the actual levels of those uncontrolled eigenvectors can be computed in the same way that one computes the levels of the controlled vectors. When this augmented design matrix is pre-multiplied by its transpose, the first three variables will be seen to be columnwise orthogonal, but the other columns will most likely deviate to greater or lesser degree from that ideal. Of course, if the assumption of three key variables is correct, that departure from the ideal may be relatively unimportant. In the event that such is not the case, one can incorporate additional variables in the design, provided, of course, that solution of these more complex cases is feasible.

Eventually, the addition of more predictor variables encounters the same resource difficulties as are encountered in the usual approach to experiment design. The number of treatments, even in a two-level factorial design, increases as 2^N, where N is the number of variables. The usual means of dealing with the problem, such as fractional factorial designs and random balance designs, apply in eigenvector space just as they do in property space.

Further consideration of such matters is beyond the scope of this appendix, the purpose of which is to set forth a *methodology* for experiment design in eigenvector space and to provide tools for its implementation. That objective has been achieved. In addition, our conclusions are believed to be real rather than merely theoretical. The tailor-to-order treatment blends represent mixtures of *real* fuels known to be extant, and hence producible, by virtue of their occurrence in the fuel data base.

Finally, we have proposed that experiment design and data analysis be viewed as dual aspects of the experimental approach. Even when an experiment is highly unbalanced in terms of the original predictor variables, a set of combined variables can be defined to orthogonalize the design matrix. These are the entities that are *really* varied in the experiment, *not* the single fuel properties.

In a strict sense, one can go so far as to say that these vector variables are the *only* predictors capable of yielding correct response information. Once the data has been analyzed in those terms, however, it is possible to deduce the relative importance of each of the original variables *in that particular, orthogonalized context* by partitioning the model SS in the manner presented in Appendix D.

The full implication of the "duality" principle presents a major challenge to the usual procedures involved. Pushed to the limit, as noted above, the duality asserts that an experiment can not be properly analyzed *except* in terms of the eigenvectors for that particular experiment. If the design matrix is orthogonal, then the design basis and the eigenfuel basis will be the same. Otherwise, the analysis should follow eigenvector lines.

It is this connection that constitutes the "common calculus" alluded to earlier. The "flip side" of the argument is that, once an orthogonal set of predictor variables has been identified, these vectors are identified as the ones that *can* be independently varied, even though the individual components can not. These eigenvectors, though possessing no innate property of orthogonality, can be used as building blocks in which the experimenter is *enabled*, so to speak, to *consciously* design an orthogonal array according to the usual principles of experiment-design theory.

E.6. REFERENCES

1. McAdams, H.T., R.W. Crawford and G.R.Hadder. 2000. *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*, SAE Technical Paper 2000-01-1961.

2. McAdams, H.T., R.W. Crawford and G.R. Hadder. 2000. *A Vector Approach toRegression Analysis and Its Application to Heavy-Duty Diesel Emissions*, ORNL/TM-2000/5, Oak Ridge National Laboratory, Oak Ridge, TN. November.

Addendum I

COMBINATORIAL ASPECTS OF FUEL BLENDS SATISFYING EXPERIMENT-DESIGN REQUIREMENTS

To find fuel blends capable of satisfying formal experiment-design requirements requires finding subsets of the set of available fuels and then testing each of those fuel combinations to see if it provides a non-negative proportion of each of the fuels in the combination. For the present, there is available a set of 76 distinct fuels from which subsets of 4 (k=2 treatment levels) may be drawn. However, it is reasonable to believe that, in a practical refinery environment, the number of source fuels available for blending might be considerably less than that number. Selection of subsets may be done exhaustively or by random sampling.

The number of subsets possible in a collection of N items is 2^N, if the term *subset* is defined to include both the empty set, containing *no* items and the full set, containing *all* items. For example, given three items a, b, and c, the possible subsets are:

```
Null set   [ ]          [a, b]
           [a]          [a, c]
           [b]          [b, c]
           [c]       [a, b, c]   Full set
```

Note that the number of possible subsets containing 0, 1, 2, or 3 items are, respectively, 1, 3, 3 and 1.

This relation can be expressed in the form of a "generating function" Gf as follows:

$$Gf(3) = (s + \bar{s})^3 = s^3 + 3\, s^2\bar{s} + 3\, s\bar{s}^2 + \bar{s}^3$$

where the expansion coefficients express the number of subsets of a given type, the exponent of s denotes the number of items *included* in the set, and the exponent of $\bar{s}$ denotes the number of items *excluded* from the set.

The general form of the generating function, for any number of items n, is

$$GF(n) = (s + \bar{s})^N$$

in which the general term is

$$_nC_k\, s^k\, \bar{s}^{n-k}$$

where $_nC_k = n!/(k!\ (n-k)!)$ denotes the number of combinations of n items taken k at a time, or the number of subsets of k items that can be taken from a universe of n items.

In the present environment, there are 76 fuels capable of being taken 4 at a time to yield

$$_{76}C_4 = 76!/(4!*72!) = 1,\ 282,\ 975$$

candidate subsets. Each candidate would then have to be tested for "realizability." It is "admissible" only if the blend proportion computed for every fuel in the blend is a non-negative quantity. With as many as 76 fuels to choose from, one can expect multiple solutions for a given treatment vector.

It is at this point that the exercise can become quite computation-intensive, particularly if the number of source fuels is large. However, real life does not require that every possible combination of source fuels be examined. Consequently, we can draw samples of four numbers, each having equal likelihood of being some number between (and including) 1 to n, where n is the number of available source fuels, and interpreting those numbers as source fuel candidates for blending. They must then, of course, be tested for acceptability. Completion of the several steps requires invoking several Matlab m-files developed for that purpose.

Invoking those m-files allowed at least approximate calibration of the number of "hits" that might be expected in a particular sampling process. In an exhaustive search of all subsets of four fuels from among the 76 available, it was found that 13,002 combinations provided realizable solutions for the [-1 -1 -1] treatment vector. The "yields" of acceptable blends obtained in 5000 random samples are shown below for each of the eight treatments in the 2^3 factorial design.

Treatment	Solutions
-1 -1 -1	44
-1 -1 +1	50
-1 +1 -1	15
-1 +1 +1	12
+1 -1 -1	83
+1 -1 +1	59
+1 +1 -1	39
+1 +1 +1	12

It is to be understood that if more than three variables are included in the experiment design, the search would be more complex and would most likely have lower "yield" of admissible blends. Running times would depend on the clock speed of the computer in use. Our exploration with a relatively slow computer (133 MHz) indicated, however, that the random search approach is feasible within the present state of computer art and is compatible with source fuels likely to be available under current refinery practice. Most likely, random search would not be necessary, because refiners would have available preferred, candidate source fuels to be used in the preparation of treatment blends. Alternate choices would be required only if the initial choice of source fuels failed to give realizable solutions.

Addendum II

COMPLETE LIST OF AVAILABLE FUELS IN THE DATABASE

For purposes of the application, only the coefficients of the first four eigenfuels are used, since only these four are used in the computations of "eligible" blends.

Fuel No.	Coefficients of the First Four Eigenfuels Eig 1	Eig 2	Eig 3	Eig 4
1	-5.3758	-0.0438	-0.8549	0.3440
2	-5.3020	-0.3521	-2.0348	-0.1123
3	-4.6060	0.5508	1.6059	-0.2820
4	-2.7605	-0.9520	-1.7526	-3.2491
5	-1.8950	1.7602	-0.1361	-0.6110
6	-1.3172	2.1513	-0.3998	-1.8564
7	-1.7540	0.8290	-0.2172	-1.3220
8	-1.4034	-0.2630	0.8515	4.2418
9	-1.7259	0.7117	-0.6660	-1.4956
10	-1.1078	1.9912	-0.1062	-0.3789
11	-1.0091	1.5219	-0.0808	1.4707
12	-1.1487	1.8699	-0.0949	-0.4755
13	-2.4925	-3.4119	2.4646	-0.2809
14	-0.8829	-0.5011	-1.2429	-2.0812
15	-0.5028	0.2953	0.0288	2.3392
16	-2.4588	-3.5526	1.9260	-0.4892
17	-0.6949	0.3721	0.1878	0.7902
18	-0.5874	5.1957	0.1268	-1.4008
19	-0.0648	2.0348	0.0970	0.5424
20	-0.0097	1.8102	0.4810	1.3915
21	-0.0316	1.4770	0.4660	2.0823
22	-0.3623	3.5796	0.5198	-1.2769
23	-0.5804	-0.4821	-1.1460	-0.8984
24	-0.6596	0.2246	-0.3764	0.5720
25	-0.0480	-0.5288	-0.0072	0.2345
26	-0.2772	1.8293	1.0556	-1.0588
27	0.4434	1.2410	0.3738	2.4347
28	0.0205	-0.5829	-0.4992	0.9915
29	0.1205	-0.2126	0.0658	0.2590
30	-0.4242	-0.0332	-1.2280	1.8532

31	-0.2935	0.7253	1.4268	-0.7426
32	0.0636	-1.3437	-0.2938	0.7215
33	0.9046	-0.0200	0.1672	2.9069
34	-0.0000	-0.3579	-1.1714	-0.7230
35	0.1223	-0.4112	0.7679	0.0363
36	0.4103	0.1016	-0.2104	1.0641
37	0.6460	0.0418	-0.1723	0.0513
38	0.7634	0.2128	-0.2969	0.0606
39	0.0678	-0.7806	-1.2559	0.6989
40	0.1721.	-1.9179	-0.0548	-0.0115
41	0.0227	-0.8238	-1.1357	-0.2019
42	0.1334	-2.2896	-0.1699	-0.1328
43	1.3714	1.5913	-0.4950	-0.3507
44	0.1422	-2.3265	-0.3110	-0.1873
45	0.4562	0.1952	-0.5893	1.5945
46	0.0959	-0.8980	-1.1047	0.5254
47	0.1243	-0.6256	1.5259	-0.2042
48	0.5218	-0.2641	-1.1766	-0.5814
49	0.1072	-0.9449	-1.8843	0.4559
50	0.7486	-0.9352	-0.0175	0.0084
51	0.8291	0.6059	-0.3735	-0.6359
52	0.8011	1.0553	-0.4712	-0.9769
53	0.8276	-0.0553	-1.3228	-0.3362
54	0.1337	-1.0555	-2.3075	0.2923
55	0.3665	-0.5452	-2.0645	-1.2012
56	0.8349	-0.0855	-1.4382	-0.3808
57	1.9420	0.7178	-0.8084	-0.6141
58	0.8662	-0.2162	-1.9384	-0.5742
59	0.2249	-2.6717	-1.6319	-0.6981
60	0.2249	-2.6717	-1.6319	-0.6981
61	0.8774	-0.2631	-2.1179	-0.6436
62	2.8552	2.0274	-0.2497	0.3187
63	1.5192	0.9746	-2.8547	-1.2632
64	2.7289	2.2465	0.2730	1.2135
65	2.0673	0.1983	-2.7962	-1.3828
66	3.2832	-1.0227	-0.5131	0.0704
67	3.5080	-1.0346	-0.5263	0.1782
68	3.3282	-1.2104	-1.2313	-0.2072
69	3.3370	-1.2472	-1.3724	-0.2618
70	2.6777	-2.0109	2.3946	-0.4224
71	2.8274	-1.9202	2.5112	-1.2978
72	3.0770	2.7964	2.4038	-0.9489
73	3.5682	-1.2860	-1.4881	-0.1937
74	3.6797	-0.4406	0.2655	1.0940
75	3.5126	-1.5302	1.2258	-0.3776
76	3.6356	-1.5675	-2.5654	-0.6103

APPENDIX F

EIGENVECTORS: THEIR ROLE IN EMISSIONS

APPENDIX F. EIGENVECTORS: THEIR ROLE IN EMISSIONS

F.1. INTRODUCTION

The formulation of diesel fuels in terms of weighted combinations of fuel properties has been well demonstrated, as has also the methodology by which emissions can be expressed in terms of those combinations. The relevant publications are SAE Technical Paper 2000-01-1961 [Ref 1], presented in Paris, and ORNL Report ORNL/TM-2000/5 [Ref 2]. Additionally, Appendix F examines how to blend fuels in such a way as to meet specific requirements imposed by experiment-design considerations. The database for these studies and for the examples and demonstrations in this appendix will be alluded to as "the Paris database." Applicable nomenclature may be found in the Addendum.

Though useful for preparing blends having specified eigenvector weights, the methodology of Appendix E does not make clear what happens to a fuel when some *arbitrary* change is made to either the weights of its eigenvectors or to the weights of one or more specific fuel *properties*. Furthermore, it has not been made clear how fuel *properties* can be manipulated so as to produce a desired weight for one or more eigenvectors that play a critical role in emissions. Unless one understands the mechanisms by which fuel properties are transformed into eigenvector weights and vice versa, one might reach the conclusion that to improve emissions one must alter the weights of fuel properties *simultaneously* and in the same relation as they appear in the eigenvector of interest. This conclusion is not only erroneous but countermands the very flexibility that makes the eigenvector approach so useful.

F.2. FUEL MODIFICATIONS: HOW THEY AFFECT EMISSIONS

Let F denote a fuel that is unsatisfactory from the standpoint of emissions, and let F_m be that same fuel after it has been modified by a change in the value of one or more of the fuel properties. What effect does such a change have on the weights of the eigenvectors?

This question can be answered by a general rule that applies to *all* fuel modifications, whether those changes are made in accordance with the makeup of critical eigenvectors or not. Simply express F and F_m in terms of the eigenvectors of the system, compute the corresponding emissions by means of the applicable vector regression model, and compare. A simple example will suffice to illustrate the procedure.

Suppose that it is known that a reduction in mono- and poly-aromatics reduces NO_x emissions. Let us, therefore, reduce the values of these two properties by one unit *without changing the values of any of the other properties*, and let us assume, for purposes of argument, that such a modification is possible. How does such a change affect the weights of the eigenvectors in the regression model?

We proceed as follows. We have two fuels, the original fuel F and the modified fuel F_m.

Let x = Property values in standard units before modification
x_m = Property values in standard units after modification

Then $x-x_m$ = Change in property values due to modification

Let xeig = Eigenvector weights before modification
$xeig_m$ = Eigenvector weights after modification

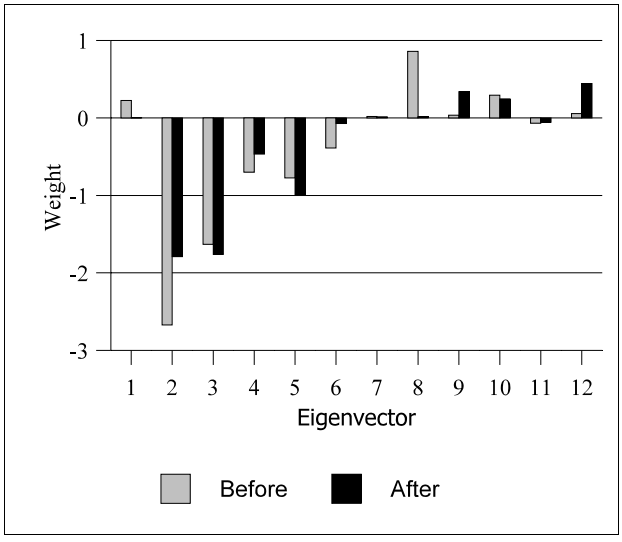
Then xeig-xeig$_m$ = Change in eigenvector weights due to modification

The changes in fuel properties (P-Space) and the results of those changes in eigenvector space (E-Space) are summarized in Table F.1. Note that although the only fuel values changed are those for the aromatics, this change propagates through *all* of the eigenvector weights. The magnitude of the change varies from one eigenvector to another, being negligible in some but substantial in others. There *is* a substantial change in the weight for Eigenvector 2, because that eigenvector loads heavily on aromatics content. Figure F.1 illustrates graphically the effects of this relatively simple perturbation on the eigenvector weights.

Table F.1. P-Space Changes and Their Effect in E-Space

Property	Before	After	Diff	Vector	Before	After	Diff
NatCet	-0.9495	-0.9495	0	1	0.2249	0.0035	0.2214
CetImpr	2.1275	2.1275	0	2	**-2.6717**	**-1.7900**	**-0.8817**
Dens	0.9459	0.9459	0	3	-1.6319	-1.7606	0.1287
Visc	-0.5788	-0.5788	0	4	-0.6981	-0.4659	-0.2322
Sulf	-0.3405	-0.3405	0	5	-0.7745	-0.9874	0.2129
Mono	**1.8634**	**0.8634**	**1.0**	6	-0.3871	-0.0702	-0.3170
Poly	**1.1597**	**0.1597**	**1.0**	7	0.0197	0.0161	0.0037
IBP	-0.1385	-0.1385	0	8	0.8606	0.0175	0.8430
T10	-0.4845	-0.4845	0	9	0.0353	0.3397	-0.3043
T50	-0.1520	-0.1520	0	10	0.2935	0.2460	0.0475
T90	0.1538	0.1538	0	11	-0.0678	-0.0555	-0.0123
FBP	0.1836	0.1836	0	12	0.0564	0.4456	-0.3892

Figure F.1. Change in Eigenvector Weights as a Result of Fuel Modification



Unless fuels are modified in a very specific way, as will be shown in the following section of this appendix, the modification produces perturbations in the weights of *all* eigenvectors, whether those eigenvectors affect

emissions or not. Clearly, if we want to be most effective in making a change to reduce emissions, we need to change the relevant *eigenvectors'* weights. However, it will be demonstrated that there are many ways to achieve improvement, some more effective than others. The objective of this paper is to clarify these predictive relationships and provide a Limit Rule for keeping emissions below some specified level.

The appendix will consider separately those property changes that affect the weights of only a *single* eigenvector and those property changes that affect the weights of *all* eigenvectors. It will illustrate a multiplicity of ways in which the weights of prominent eigenvectors can be controlled. The development will proceed, step by step, from "raw data" to final results.

In the development, two possibilities are recognized: (a) a fuel modification that changes the weights of critical eigenvectors *without* changing the weights of other eigenvectors, and (b) a fuel modification that changes the weights of the critical eigenvectors but, in doing so, also changes the weights of other eigenvectors in the system. These two kinds of fuel modification will be referred to respectively as *non-participating* and *participating*. In the following paragraphs we consider the conditions under which these two types of modifications occur and the effect that those modifications have on emissions.

Prior studies have indicated that NO_x emissions are strongly influenced by Eigenvector 2 in the Paris database, the eigenvector having the second largest eigenvalue. Accordingly, that phenomenon will provide the examples by which we demonstrate fuel modifications and their emission consequences.

F.2.1 Fuel Modifications: Non-Participating

In this section of the report, we demonstrate what happens when the weight of a specific eigenvector is changed in such a way that it does not induce changes in the weights of any of the other eigenvector weights. This type of fuel modification is referred to as *non-participating*, in the sense that changes made in the weights of selected eigenvectors do not propagate beyond those eigenvectors.

For demonstration purposes, we select from the Paris database that fuel that exhibits NO_x emissions of 5.38 gms/hp-hr. The corresponding weight for Eigenvector 2 is 2.01. Assuming that it is possible to do so, we arbitrarily subtract one unit from that eigenvector weight, making the revised weight 1.01.

Two questions are of interest:

1. What effect does this change have on emissions?
2. How does the arbitrary change in the critical eigenvector translate into changes in the fuel properties (NatCet, CetImpr, ..., T90, FBP)?

According to regression analyses reported in the cited publications,

$$\log(NO_x) = 1.5229 + 0.0344 * z_2 + \text{other terms} \qquad (1)$$

where z_2 is the weight of Eigenvector 2. Accordingly, reducing that weight by one unit is equivalent to reducing the logarithm of NO_x by 0.0344. The decrease in NO_x emissions is:

$$\exp(1.5229) - \exp(1.5229-0.0344) = 4.5855 - 4.4304 = 0.1551$$

or about 3.4% of the initial value of 4.5855 gm/hp-hr.

By hypothesis, the postulated change in the weight of Eigenvector 2 does not affect the weights of any of the other eigenvectors. What we need to know, though, is the configuration of fuel properties that makes possible such a non-participating transformation.

To translate the eigenvector change into corresponding changes in fuel properties, one proceeds as follows. Recall that the fuel-property values are transformed into eigenvector weights by the matrix multiplication

$$W = X * E$$

where X is the 280 x 12 "design matrix" and E is the 12 x 12 matrix of eigenvectors. The inverse transformation is

$$X = W * E'$$

where W is the 280 x 12 matrix of eigenvector weights and E' is the transpose of the eigenvector matrix E.

Let W_m denote the weight matrix after the modification in which the weight of the second eigenvector is reduced by one unit. Then:

$$X_m = W_m * E' \qquad (2)$$

where X_m is the X matrix as modified by the unit change in the weight of the second eigenvector. The quantity of interest is the difference between X, the original property matrix, and X_m, the modified property matrix.

Both X and X_m are matrices consisting of 280 rows (one for each fuel) and 12 columns. Upon examining the two matrices and their difference, one finds that the column-by-column differences between elements in a selected row of X and the corresponding row of X_m are identical for all rows. It suffices, therefore, to examine simply the first row of the two matrices.

These rows are compared below. For convenience of tabulation, however, the rows are displayed as columns. All values are in standard units.

	Row 1 of X, Original fuel property matrix.	**Row 1 of X_m, Modified fuel property matrix**	**Difference $(X-X_m)_{Row\ 1}$**
NatCet	-1.5391	-0.9834	**-0.5556***
CetImpr	-0.5353	-0.6788	0.1435
Sp Grv	0.9171	0.4684	**0.4488***
Visc	-1.6946	-1.5750	-0.1196
Sulfur	2.0619	1.8820	0.1799
Mono	0.1078	-0.3562	**0.4640***
Poly	2.7669	2.3492	**0.4177***
IBP	-0.4098	-0.3374	-0.0725
T10	-1.2677	-1.1543	-0.1134
T50	-1.2863	-1.2035	-0.0828
T90	-1.5227	-1.4483	-0.0743
FBP	-1.0235	-0.9753	-0.0482

It is seen that the largest changes are in NatCet, SpGrv, Mono and Poly. More significant, however, is the fact that the difference column *exactly matches the second eigenvector*.

If the weight of Eigenvector 2 had been changed by 0.5 instead of by 1.0, then the values in the difference column would be just half as great as those shown. Whatever the magnitude of the change, however, the quantities in the difference column are *proportional* to the components of Eigenvector 2. Thus it would appear that fuel modification in property space (P-Space) should be performed in accordance with the components of the critical eigenvector. This conclusion is true, however, *only* if it is mandated that there be *zero change in the weights of all but the second eigenvector*. As will be shown in the following section, this is an unnecessary restriction and one that could severely limit blending options.

The fact is that the change of one unit in the weight of Eigenvector 2 can be realized in a multitude of ways. Except for the strict case presented above, such changes may produce finite changes in the weights of the other eigenvectors. When only one eigenvector contributes to emissions, however, the changes in the weights of the other eigenvectors are irrelevant and can be largely ignored.

Of course, it is recognized that not all predictive equations are as simple as Equation (1) above. In *any* case, however, the net result of fuel modifications can be evaluated by comparing eigenvector weights before and after the modification, according to the "universal rule" enunciated at the beginning of this appendix.

F.2.2 Fuel Modification: Participating

The previous section shows that in order to change the weight of a single eigenvector without changing the weights of other eigenvectors, one must adhere to a specific regimen with regard to changes in fuel properties. When that regimen is not adhered to, changes in the weights of other eigenvectors may occur, and these changes could have either a positive, negative or neutral effect on emissions. Fuel modifications of this type will be referred to as *participating*, in the sense that changes made in the weights of specified eigenvectors also affect the weights of eigenvectors that were *not* specified.

To understand the interrelationships among changes in the weights of eigenvectors, one must understand the mathematics by which changes in the weights of fuel properties are transformed into changes in the weights of eigenvectors and vice versa. Though the requisite transformations are succinctly stated in Equation (2) above, the matrix algebra in that equation does not exhibit the actual *arithmetic* involved. The following discussion attempts to remedy that situation.

At the risk of tedium, therefore, we "begin at the beginning," with the "raw" data for fuel properties and emissions and the consequences of transformations on those data. The following paragraphs attempt to clarify, by example, how fuel modifications affect the "eigenstructure" of a fuel and how those modifications translate into emissions.

For purposes of illustration, we examine that fuel having maximum observed NO_x emissions, which was discussed abstractly above. Complication begins immediately as we transform the 12 fuel variables, each in its own peculiar units, into "standard form." First, we subtract from the value of each fuel property its mean value as computed across the entire data base. Then we divide the resulting numbers by their corresponding standard deviations as computed for the entire data base. The result is a set of 12 numbers that are dimensionless and that tend to put all fuel properties on a comparable scale, so far as the current database is concerned.

So far, the transformation is straightforward. We come now, though, to a transformation that is not quite so easy to follow but which is absolutely essential to the eigenvector approach. It is the matrix transformation of Equation (2) above.

The argument goes as follows. The 12 fuel properties are interrelated. Try as it might, a single fuel property – say T90 – can not vary independently of – say, T50. As we go from one fuel to another in the data base, if T50 goes up, T90 will tend to go up also. In other words, these two properties may be *covariant* to such an extent that they could almost be treated as a *single variable*. However, they do not *exactly* track each other, and our analysis has to allow for that fact.

It is exactly here that the concept of eigenvector arises. There are certain weighted combinations of the 12 fuel properties, combinations called eigenvectors, whose weights *can* be varied independently, even though it may *not* be possible to vary independently the values of the individual fuel properties. That being the case, we make a transformation that replaces the 12 fuel properties with 12 weighted combinations of those properties. Our present concern is with the specifics of how the property weights and the eigenvector weights are related and with how one can be transformed into the other. Pertinent data are given in Table F.2.

Given that X is a matrix of fuels and that E is a matrix of the eigenvectors of the correlation matrix of X, every fuel in X has a representation as a set of eigenvector weights or coefficients, referred to by some as principal components. Recall that in the Paris database the matrix X consisted of 280 rows (one for each fuel) and 12 columns (one for each fuel property).

At an early point in the eigenvector-analysis process, one makes the transformation

$$Y = X * E \qquad (3)$$

$$280\text{x}12 \quad 280\text{x}12 \quad 12\text{x}12$$

TABLE F.2. THREE EQUIVALENT DESCRIPTIONS OF CASE 125 DIESEL FUEL

	Physical Units	Standard Units	Eigenvector Wts.	
NatCet	39.9	-0.87	Eig.1	2.68
CetImpr	0	-0.54	Eig.2	**2.01**
Dens	0.88	1.53	Eig.3	2.39
Visc	3.08	0.76	Eig.4	-0.42
Sulf	3360	2.44	Eig.5	0.52
Mono	20.4	0.37	Eig.6	-0.32
Poly	20.4	1.85	Eig.7	0.26
IBP	197	0.84	Eig.8	0.64
T10	233	0.98	Eig.9	-0.04
T50	278	1.02	Eig.10	0.14
T90	327	0.86	Eig.11	-0.03
FBP	364	1.10	Eig.12	-0.35
NO_x	5.38	5.38		5.38

The numerics 280x12 and 12x12 denote the "size" of the matrices, the first number denoting the number of rows, the second number denoting the number of columns.

In the following discussion, we focus attention on a single fuel – i.e., a single row of the matrices X and Y.

For a single fuel, the transformation becomes

$$y = x * E \qquad (4)$$

1x12 1x12 12x12

The components of y are the weights that have to be assigned to the respective eigenvectors of E in order to describe the particular fuel represented by x (a single row with 12 columns). Moreover, if we are interested only in the second eigenvector, the transformation degenerates even further into

$$y = x * e \qquad (5)$$

1x1 1x12 12x1

where e denotes just the second column of the eigenvector matrix E. The result, y, is a 1 by 1 matrix – that is, a single row and a single column; in other words, a scalar, or just a plain number.

The fuel property vector is as follows:

-0.87 -0.54 1.53 0.76 2.44 0.37 1.85 0.84 0.98 1.02 0.86 1.10

where the columns represent, respectively, the fuel property values in standardized units. These are the numbers displayed in the "Standard Units" column of Table F.2.

The application of Equation (4) to this vector gives:

2.68 2.01 2.39 -0.42 0.52 -0.32 0.26 0.64 -0.04 0.14 -0.03 -0.35

These numbers are displayed in Table F.2 as "Eigenvector Weights."

Now, let us narrow our attention just to the single value 2.01, highlighted in Table F.2 as the "weight" or "coefficient" of Eigenvector 2, the one that we have singled out as critical.

Matrix multiplication is a row by column operation, but it will be more revealing to display x as a column alongside the column vector e. We then multiply the numbers in each row to obtain a third column containing those products.

The sum is the result of the abbreviated matrix multiplication and agrees with the value tabulated in the column headed "Eigenvector Weights" in Table F.2.

The important point to be made here is that the numbers in Column B (e in Equation 5) are *fixed*; they are the components of the second eigenvector. The numbers in A (x in Equation 5) are peculiar to the fuel under consideration; they denote respectively the standardized values of NatCet, CetImpr, ..., T90, FBP. It is clear, therefore, that there could exist many sets of numbers in column A that, when multiplied by their correspond in Column B, could produce the same sum, 2.0109.

A (x)	B (e)	A*B (y)
-0.8672	-0.5556	0.4819
-0.5353	0.1435	-0.0768
1.5314	0.4488	0.6872
0.7565	-0.1196	-0.0905
2.4428	0.1799	0.4396
0.3698	0.4640	0.1716
1.8464	0.4177	0.7712
0.8357	-0.0725	-0.0606
0.9834	-0.1134	-0.1115
1.0230	-0.0828	-0.0847
0.8586	-0.0743	-0.0638
1.0954	-0.0482	-0.0528
	Sum	2.0109

The problem is amenable to still further simplification, if we are willing to accept a "reasonable approximation." In Column B, the first, third, sixth and seventh rows exhibit the largest numbers (in an absolute-value sense). If we sum just the rows of Column C that correspond to those rows, we obtain:

	0.4819
	0.6872
	0.1716
	0.7712
Sum	2.1119

The selected rows correspond respectively to NatCet, Dens, Mono and Poly. Therefore, one could use various values of those properties and still satisfy the condition that the sum is approximately 2.01 (or any other specified value). Even if the effects of these changes propagate into the weights of other eigenvectors, no harm is done because the other eigenvectors have been dismissed as having no substantial or significant effects on emissions.

The many-to-one mapping of fuel values into eigenvector weights offers great flexibility for fuel modifications aimed at reducing emissions. In particular, it facilitates a very flexible rule for limiting emissions to a specified upper limit, as will be demonstrated in the following section.

F.3 A Limit Rule for Emissions

According to the analyses set forth in the Paris paper, it appears that NO_x is largely a function of Eigenvector 2 according to the regression line given in Equation 6 below.

$$\log(NO_x) = 1.5229 + 0.0344 * z_2 \qquad (6)$$

The key to understanding how this equation can be used to set a limit for NO_x emissions appears to reside in the quantity z_2. It is important, therefore, to understand exactly what z_2 represents and how it can be constrained.

The quantity z_2 is associated with Eigenvector 2, but it is important to keep in mind that it is a scalar, the *weight* of that vector, as computed by the procedures illustrated in the previous section. In what follows, we apply the teachings of that section to placing a limit on NO_x emissions.

First, we started with a list of fuels (there were 280 in the data set), though many of them were duplicates. Each one of those fuels was characterized by 12 pure numbers, dimensionless, and associated respectively with the 12 variables of choice: NatCet, CetImpr, ..., T90, FBP. Those numbers were obtained by subtracting the mean for each property from the property value for the fuel under consideration and dividing by the corresponding standard deviation for that property. These numbers generally would be between the values -3 and +3, but can be subject to an occasional excursion outside those limits. They represent positions in what has been termed P-Space (P for "Property").

Next, these "standardized" values of the fuel properties were transformed to eigenvector weights. As such, they serve as multipliers of the eigenvectors that serve as a basis for the space we call E-Space (E for "Eigenvector"). Geometrically, they represent positions in that space.

It is important to understand the difference between these two vector spaces; otherwise, confusion in interpreting z_2 in Equation (6) is likely to arise.

To appreciate the difference between P-Space and E-Space, one must return to those uncomplicated times when a number was just a number and nothing more.

In P-Space, one regressed log-transformed emissions on actual values of NatCet, CetImpr, ..., T90, FBP, where those quantities were expressed in either physical or standardized units. The notion of *vectors* never arose, but that did not mean that vectors were not involved. Actually the numbers used in the regression are multipliers or weights for an underlying set of *basis* vectors just as in E-Space.

The vectors in P-Space are disarmingly simple, and that is why they are lost sight of in ordinary computations. In reality, in the instance in which there are twelve properties, there are twelve vectors. Each of these vectors contains eleven zeros and a single one. The first vector has the 1 in first position, the second vector has the 1 in the second position, and so on until the last (twelfth) vector, which has the 1 in the twelfth position.

Basis Vectors in P-Space

NatCet	1	0	0	0	0	0	0	0	0	0	0	0
CetImpr	0	1	0	0	0	0	0	0	0	0	0	0
Dens	0	0	1	0	0	0	0	0	0	0	0	0
Visc	0	0	0	1	0	0	0	0	0	0	0	0
Sulf	0	0	0	0	1	0	0	0	0	0	0	0
Mono	0	0	0	0	0	1	0	0	0	0	0	0
Poly	0	0	0	0	0	0	1	0	0	0	0	0
IBP	0	0	0	0	0	0	0	1	0	0	0	0
T10	0	0	0	0	0	0	0	0	1	0	0	0
T50	0	0	0	0	0	0	0	0	0	1	0	0
T90	0	0	0	0	0	0	0	0	0	0	1	0
FBP	0	0	0	0	0	0	0	0	0	0	0	1

Clearly, these vectors make up the 12 x 12 identity matrix which, by definition, does not change anything it multiplies. So, it is easy to ignore its existence, even though without it there would be no agreed-on space in which to work.

So, in this space, we decide to regress log-transformed emissions on these pure, dimensionless numbers that form a 280 x 12 matrix, where the rows denote the 280 separate observations and the columns denote the standardized fuel property variables. For the regression equation, we have

$$\log(NO_x) = a_0 + a_1 x_1 + ... + a_{12} x_{12} \quad (7)$$

where the x-values are the standardized values of the property variables and the a-values are numbers to emerge from the regression exercise. After the fact, we may agree to eliminate some of the terms and recompute the regression equation – but eventually we end up with an equation that is supposed to predict $\log(NO_x)$, a pure number, in terms of a weighted combination of the pure numbers associated with whatever terms we retain in the regression equation. *Vectors* never enter our minds, but they are there in the background all the same. And the x-values are really just the multipliers, weights, or coefficients for those "hidden" vectors.

The situation is actually no different in E-space. We have a proposed regression equation of the form:

$$\log(NO_x) = b_0 + b_1 z_1 + ... + b_{12} z_{12} \quad (8)$$

where the z-values are just scalars, specifically the multipliers, weights, or coefficients for the set of basis vectors we call eigenvectors. As in P-space, the z-values are just numbers and nothing more. It just happens that they come into being somewhat differently than do the x-values in P-space. It is sufficient to know that these numbers have the nice characteristic of orthogonality, which makes life easier for us because it assures us that the regressors are mathematically independent and statistically uncorrelated.

This feature is the result of our choice of vectors to define our working space. Those vectors, however, play no direct role in the regression and, for the moment at least, we can forget about them, just as we did unknowingly in the previous case for P-space. And, just as we did there, we can drop out terms that we consider unimportant and retain the rest, to end up with the very simple Equation (6) displayed at the beginning of this discussion. A nice difference between the P-space and E-space exercises is that in E-space we can drop terms and not have to recompute the regression equation. That, however, is "icing on the cake" and not the main point at issue here.

The difference between the z-values and the x-values is the key to understanding the nuances of E-space. In Equation (7), which deals with fuel *properties*, there is only *one* way that we can get a given value of a particular x in the equation. That value is simply the standardized value of the fuel property associated with that particular x. But, in Equation (8), which deals with *eigenvectors*, a fixed value of one of the z-values – call it q – has many ways to come into existence. It can consist of a little bit of standardized density, a smidgeon of standardized aromatics and maybe a whole lot of standardized cetane number; or, maybe a bit more of this, much more of that, and so on. The mix doesn't matter – so long as the final result is the value q that we started out to achieve as a limiting value.

In short, in P-space there is a one-to-one mapping of – say, standardized cetane number – into a value of the corresponding x. But, in E-space, there is a many-to-one mapping of standardized cetane number, density, aromatics, etc. into a given value of z.

Let us now return to Equation (6). How can that equation be used to assure that NO_x emissions will not exceed 5 gm/hpr-hr? Or, equivalently, assure that $\log(NO_x)$ will not exceed log(5)? Clearly, equation (6) can be solved to find a limiting value of z_2, the multiplier, weight or coefficient for Eigenvector 2. That having been done, we can seek, at our discretion and in any way that makes practical sense, a combination of standardized cetane number, density, aromatics and so on that will keep z_2 below the prescribed limiting value.

It turns out, as can be verified by inspection of the eigenvectors, that *most* of the weight comes from: NatCet, Density, Mono and Poly Aromatics. So any combination of the standardized values of these four properties, when multiplied by their associated weighting factors, is a candidate for producing a fuel for which z_2 is below its critical value. If the fuel satisfies that constraint, then as shown above, that fuel should assure that NO_x is below 5 g/hpr-hr. A practical illustration of this principle is shown below.

The Paris data base was searched to find a subset of cases for which NO_x <= 5.0 gm/hp-hr. Then *that* subset was searched to find a subset for which z_2, the weight of Eigenvector 2, does not exceed a certain critical value.

That critical value is determined by solving for z_2 in the basic equation

$$\log(NO_x) = 1.5229 + 0.0344 * z_2$$

It is found that z_2 can not exceed a value of approximately 2.5 if NO_x is to be held below 5 gm/hp-hr.

Recalling the relation between z_2 and its fuel-property components, one can make the approximation that it is only NatCet, SpGrv and Mono and Poly aromatics that have substantial effects on z_2. Accordingly, one can ignore the other components and seek only those combinations of values of NatCet, SpGrv, Mono and Poly whose weighted sum does not exceed 2.5.

Table F.3 displays four cases that satisfy that criterion. They are displayed as a matrix of four rows and five columns, the first four columns representing the fuel-property values for the four cases, displayed as rows. The fifth column displays the corresponding weighting factors for Eigenvector 2.

TABLE F.3. FUELS SATISFYING THE LIMITING CONSTRAINT

NatCet	SpGrv	Mono.	Poly.	Eig.2 Wt.
0.6547	-1.7993	-0.7438	-0.5206	-1.7338
1.0660	-0.9354	-1.2024	-1.0466	-2.0072
-0.9906	0.9891	1.2476	0.9113	**1.9539**
2.7114	-0.5659	-1.2155	-0.0092	-2.3284

Keep in mind, however, that these are only a few of the many combinations of the weights of the four properties that satisfy both conditions. These few, however, illustrate great diversity. Note that all five cases yield products <= 2.5 and that only the highlighted case comes close to the critical value.

To summarize: In an equation of the simple form of Equation (1), a limit rule for keeping NO_x below some specified value k consists simply of keeping z_2 below some corresponding value w. Since z_2 is computed as a weighted combination of the fuel values for NatCet, Dens, Mono and Poly, the rule affords considerable flexibility in satisfying the limit requirement for NO_x.

Compared with the conventional stepwise regression approach to modeling emissions, the eigenvector approach offers a number of advantages. In the stepwise approach, there are many options for arriving at a final regression equation, and it is possible to obtain two or more equations that yield essentially the same predictions even though they may contain different fuel-property variables. The reason for this apparent anomaly is that the variables in the regression equation are "impure" – that is, they are aliased to greater or lesser extent with other predictor variables. The eigenvector form of the regression equation, on the other hand, is based on independent variables sharing nothing in common with each other.

Though a selected stepwise solution might yield predictions comparable to those obtained from the eigenvector equation, the stepwise solution is not robust. Two sets of data have no basis for comparison unless the treatment space was sampled in the same way for both. Otherwise, the aliasing of variables in the two cases will be different, and a fuel-property variable, though carrying the same label in both cases, is not strictly comparable in the two stepwise equations. On the other hand, eigenvector equations derived from the two sets can readily be transformed to a common basis so that predictions can be unambiguously compared. In short, the eigenvector approach is unique and robust.

Application of the eigenvector approach in the refinery also has noted advantages. First, it recognizes and takes into account the covariance of fuel properties. Blending based on eigenvectors, therefore, is more "natural" in that refinery streams or blendstocks are rightly characterized by their *combinations of fuel properties* rather than by isolated fuel properties alone. If a source fuel can be identified with a particular eigenvector, as has been suggested for Eigenvector 2, then the practical way to reduce NO_x emissions is to reduce the content of that source fuel. In doing so, the refiner is simply implementing the reduction of the weight of the corresponding eigenvector.

A blendstock does not *have* to be identifiable with a specific eigenvector in order to apply eigenvector methodology, however. Any source fuel can be *resolved into its eigenvector constituents*, just as were the fuels in the Paris database. That having been done, the methodology shown in Appendix E can be applied to combine blendstocks to satisfy a specified end product.

Sometimes, when an eigenvector *can not* be identified with a source fuel, one might want to attempt improvement by focusing on one or more *fuel properties* known to affect emissions. How can eigenvector information be exploited when our concern is with fuel properties? The answer to this question resides in Section F.2.2, on "participating fuel modifications." It involves transforming the P-Space modification to E-Space and observing how that modification affects the overall eigenstructure of the original fuel. In short, *if* a fuel can be modified by changing the value of a single fuel property, its emission response can be determined by transforming to E-Space and exercising the applicable eigenvector regression equation. In reality, such a change is highly unlikely because of the covariance of fuel properties, the very covariance that forms the eigenvectors of the data set.

F.3. SUMMARY AND CONCLUSIONS

This appendix has reviewed PCA and PCR+ as they pertain to fuel properties and how those properties affect emissions. Examples drawn from the Paris database have been used to illustrate how fuel modifications can be expected to affect emissions. The thrust of this appendix was to provide the mathematical foundation for applying eigenvector theory in refinery processes. The following conclusions are drawn:

1. It is the *weights* of critical eigenvectors that determine whether a fuel modification will increase or decrease emissions. The components of those vectors (the fuel properties) are important only to the extent that they affect the eigenvector weight.
2. There are many ways to realize a specified eigenvector weight; the refiner is not limited by the relations among the fuel properties that constitute the components of the vector except as the fuel properties enter into the computation of the eigenvector weight.
3. Whether a particular perturbation of one or more fuel properties increases or decreases emissions depends on how eigenvector weights are changed by the perturbation.
4. Any perturbation that affects the weights of critical eigenvectors favorably will have a favorable effect on emissions; however, those gains can be diminished or canceled if the perturbation has an unfavorable effect on other critical eigenvectors.

5. There are no numerical constraints on how fuel property values can be marshaled to reduce emissions, other than that the weights of critical eigenvectors be held within prescribed bounds. However, there are likely to be *natural* constraints arising from the covariance of fuel properties.
6. It may be possible to identify refinery streams or blendstocks with specific eigenvectors, so that to increase or decrease the effect of a specified eigenvector one needs only to increase or decrease the corresponding refinery streams or blendstocks.

F.4 REFERENCES

1. McAdams, H. T., R.W. Crawford and G.R.Hadder. 2000. *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*, SAE Technical Paper 2000-01-1961.

2. McAdams, H.T., R.W. Crawford and G.R. Hadder. 2000. *A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions*, ORNL/TM-2000/5, Oak Ridge National Laboratory, Oak Ridge, TN. November.

ADDENDUM

NOMENCLATURE AND NOTATION

The following is a list of fuel property variables. Throughout this report they are referred to by the abbreviation shown in boldface type.

Natural **Cet**ane

Cetane **Impr**over

Specific **Gra**vity

Viscosity

Sulfur

Mono Aromatics

Poly Aromatics

IBP

T10

T50

T90

FBP

INTERNAL DISTRIBUTION

1. S. Das
2. E. C. Fox
3. Ronald L. Graves
4. D.L. Greene
5-26. G.R. Hadder
27. P.S. Hu
28. H.E. Knee
29. R.N. McGill
30. ORNL Patent Office
31. Central Research Library
32. Laboratory Records

EXTERNAL DISTRIBUTION

33. M. Beardsley, U.S. Environmental Protection Agency, 2000 Traverwood Road, Ann Arbor, MI 48105
34. W. Clark, National Renewable Energy Laboratory, 1617 Cole Boulevard, Golden, CO 80401-3393
35-39. R.W. Crawford, RW Crawford Energy Systems, 2853 S. Quail Trail, Tucson, AZ 85730-5627
40-41. L. Cunningham, Technical Business Manager, Refinery Chemicals and Fuel Performance Additives, Ethyl Corporation, 330 South Fourth Street, Richmond, VA 23218-2189
42. R.I. Davidson, Assistant Director, Fuels Research and Development, Ethyl Petroleum Additives, Inc., 500 Spring Street, P.O. Box 2158 Richmond, VA 23218-2158
43. P.R. Devlin, EE-32, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
44. K.G. Duleep, Energy and Environmental Analysis, Inc., 1655 North Fort Meyer Drive, Arlington, VA 22209
45-46. R.G. Dulla, Sierra Research, 1801 J. Street, Sacramento, CA 95814
47. Tom R. Eizember, Americas Regional Planning, ExxonMobil Refining and Supply Company, Room 5B006, Fairfax, VA 22037
48. J.A. Garbak, EE-32, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
49. S.J. Goguen, EE-33, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585
50. Mike Grabowski, Department of Chemical Engineering, Colorado School of Mines, Golden, CO 80401
51. A.M. Hartstein, FE-3, U.S. Department of Energy, Germantown, 19901 Germantown Road, Germantown, MD 20874-1290
52. A.M. Hochhauser, ExxonMobil Research and Engineering Company, 600 Billingsport Road, Paulsboro, NJ 08066
53. David Korotney, U.S. Environmental Protection Agency, 2000 Traverwood Road, Ann Arbor, MI 48105
54. M.E. Leister, Fuels Technology Manager, Marathon Ashland Petroleum LLC, 539 S. Main St., Findlay, OH 45840
55-56. James Lyons, Sierra Research, 1801 J. Street, Sacramento CA 95814
57-58. Robert Mason, Southwest Research Institute, 6220 Culebra Road, San Antonio TX 78228-0510
59-63. H.T. McAdams, AccaMath Services, Carrollton, IL 62016
64. R.L. McCormick, National Renewable Energy Laboratory, 1617 Cole Boulevard, Golden, CO 80401-3393

65-69. B.D. McNutt, PO-62, Room 7H-021, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585

70. W. Stuart Neill, National Research Council Canada, Montreal Road, Building M-9, Ottawa, Ontario, K1A 0R6

71. P.D. Patterson, Jr., EE-30, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585

72. Richard A. Rykowski, Assessments and Standards Division, U.S. Environmental Protection Agency, 2000 Traverwood Drive, Ann Arbor, MI 48105

73. Gurpreet Singh, EE-33, U.S. Department of Energy, Forrestal Building, 1000 Independence Avenue, S.W., Washington, DC 20585

74. Dr. George Sverdrup, National Renewable Energy Laboratory, 1617 Cole Boulevard, Golden, CO 80401-3393

75. M. Singh, Argonne National Laboratory, 955 L'Enfant Plaza, SW, Suite 6000, Washington, DC 20024-2168