

A Vector Approach to Regression Analysis and Its Application to Heavy-Duty Diesel Emissions

H. T. McAdams

AccaMath Services

R. W. Crawford

RWCrawford Energy Systems

G. R. Hadder

Oak Ridge National Laboratory

Copyright © 2000 Society of Automotive Engineers, Inc.

ABSTRACT

An alternative approach is presented for the regression of response data on predictor variables that are not logically or physically separable. The methodology is demonstrated by its application to a data set of heavy-duty diesel emissions. Because of the covariance of fuel properties, it is found advantageous to redefine the predictor variables as *vectors*, in which the original fuel properties are *components*, rather than as *scalars* each involving only a *single* fuel property. The fuel property vectors are defined in such a way that they are mathematically independent and statistically uncorrelated. The available data set is not considered adequate for the development of a full-fledged emission model. Nevertheless, the data clearly show that only a few basic patterns of fuel-property variation affect emissions and that the number of these patterns is considerably less than the number of variables initially thought to be involved.

INTRODUCTION

Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several predictor variables. In spite of its evident success in many applications, however, the regression approach can face serious difficulties when the predictor variables are to any appreciable extent covariant. This point is made quite evident in a recent review by Lee, Pedley, and Hobbs [1], in which efforts to evaluate the separate effects of fuel variables on diesel emissions were often frustrated by the close association of fuel properties.

This paper is an attempt to address these concerns by offering what may be an ameliorative approach to modeling the effects of fuel characteristics on emissions from heavy-duty diesel (HDD) engines. The approach uses an adaptation of Principal Component Regression (PCR) that

is indicated to have certain advantages over stepwise regression, which was widely used in the development of the Complex Model for Reformulated Gasoline [2].

Our approach is only one of many that have been devised to counter regression difficulties, such as ridge regression, all possible regressions, and PCR, as attested to by the extensive literature on these subjects [3-5]. Each has its advantages and disadvantages, and in each a certain degree of art and arbitrariness must be recognized. It is the contention of this paper, however, that PCR, because of its seeming difficulty of interpretation, is used less widely than it might be if better understood. Therefore, we provide in this paper a tutorial demonstration applied to the problem of emissions from HDD engines in an effort to increase awareness of the method.

STATISTICAL PERSPECTIVE

It is easily demonstrated that predictor variables can be naturally associated in a way that defies their separation. For example, it is evident and unarguable that increasing a fuel component such as olefin implies decreasing one or more other components such as aromatics or paraffins. Other examples abound in the refining world, such as the association of distillation characteristics with chemical composition.

These natural associations are not to be confused with apparent associations that arise from inappropriate experiment designs that violate principles enunciated by R.A. Fisher in his pioneering book *Design of Experiments* [6]. Neither are they to be confused with applications involving Principal Component Analysis (PCA) and other factor-analytic methods that are used to understand the interrelation among descriptive variables, as was pioneered by L. L. Thurstone in his book *The Vectors of Mind* [7]. Rather, for variables naturally associated in a physical

system, our approach defines predictor variables in such a way that the new variables are orthogonal. It then uses them as the predictor variables in an ordinary least squares (OLS) regression.

No attempt is made to select a subset of those variables before performing the regression, as is commonly done in many applications of PCR in the literature [8-13]. This usage has been soundly criticized, and rightly so, by Hadi and Ling in their paper "Some Cautionary Notes on the Use of Principal Components Regression" [14], where they point out instances in which the *a priori* choice of variables "fails miserably" in prediction. It should be evident that choosing explanatory variables without regard to the response variable is no more possible by means of PCA than by any other means.

Therefore, it is important to recognize that our approach uses PCA and PCR for two distinct purposes. The first is to resolve the design matrix into eigenvectors that explain, in the most compact way, the variation among the original variables (here, fuel properties). Second, we use those eigenvectors, all of them, as predictors of the response variable (here, emissions). Because of orthogonality, one can partition the model sum of squares (SS) explicitly among eigenvectors and drop from the model those that are deemed unimportant, either because they fail to reach a specified level of significance or because they contribute little to the prediction in terms of magnitude. This approach is similar in many respects to the case studies described by Jeffers [15].

The final step consists of "pruning" the retained eigenvectors of those components (original variables) that contribute little to prediction. This step is possible because the ability to partition the model SS among eigenvectors implies the ability to partition that SS among their components – namely, the original variables. The method by which the partitioning is realized avoids such commonly used procedures as removing variables one at a time to determine how their removal affects the model SS.

It should be evident that partitioning the variance among independent variables (fuels) and partitioning the model SS for the response (emissions) will differ, sometimes radically. It is quite possible that an eigenvector that plays a minor role in explaining the variability among fuels may play a major role in prediction, and vice versa. Where found to exist, such "discrepancies" indicate that variable associations having the greatest effect on response are ones that were varied little in the design matrix. This fact can provide valuable insight for improving the experiment design in follow-on work.

The transition from scalar to vector predictors brings with it a spate of interpretational and inferential issues. Tests of significance, for example, may need to be viewed in a different light, and it may be appropriate to put more emphasis on the magnitude of an effect rather than on its probability of occurring by chance. So firmly embedded in our research culture is the statistical paradigm that most

investigators are disinclined to acknowledge the existence of other criteria for judging the worth of a scientific finding. Moreover, the 0.05 level of significance is a fixed icon and tends to be routinely applied, even though the power of the test is strongly dependent on sample size. Further, one tends to accept, without question, that a variable either is or is not "significant," *in toto*.

In the present circumstance, however, one may have to accept the fact that a variable's significance may depend on its associations. This is because rejecting an eigenvector implies only partial rejection of the original variables comprising it, because the same variables occur in those eigenvectors that are retained. In this paper we accept the notion of partial significance, believing that a variable can be *significant* when found in respectable company and *not significant* in less respectable company.

We also prefer to evaluate effects on the basis of their magnitude in addition to their probability of occurring by chance. Sample size still plays a role, of course, but in a way that is the dual of its role in a conventional test of significance. For a fixed significance level, such as the classic 0.05, the magnitude of the "least detectable" effect is variable and depends on sample size. When a fixed magnitude is used as the threshold for acceptance or rejection of an effect, it is the statistical significance level that varies, again depending on sample size. We believe it is essential to balance the two considerations. Even though an effect may be statistically significant, because of large sample size, there is no reason for it to be retained if it makes only a minuscule difference in predictions.

Many other sensitive philosophical considerations are posed by the methodology demonstrated in this paper, and it is beyond our scope to attempt to resolve them all. Our work is continuing, however, and will be reported in a forthcoming publication by Oak Ridge National Laboratory [16] and in other suitable forums.

The vector approach developed below provides a generally applicable method for identifying an efficient set of vector variables to describe a collection of data. When applied as predictor variables in regression analysis, the vector variables are found to have many advantages, including:

- Economy of representation, because a small number of vector variables may effectively replace a larger number of the original variables.
- Simplification of regression analysis, because properly constructed vector variables (e.g. eigenvectors of the problem) will be mathematically independent and eliminate the complications introduced by multi-collinearity.
- Potentially greater understanding of the patterns of variation present in the data and how these are related to the dependent variable under consideration.

DIESEL EMISSION CONTROL

The background and status of diesel emission research are well documented by the previously cited literature review by Lee, Pedley and Hobbs. Up to this point in time it appears that engine design factors tend to eclipse fuel effects so far as efforts to reduce diesel emissions are concerned. Moreover, as pointed out by the authors, a number of prototype engine technologies are under consideration in order to meet future proposed emission limits (US 2004 and EU 2000/2005).

How fuel properties may influence diesel emissions in the future is particularly problematical, especially in those instances where engine or fuel properties play an *enabling* role for the other. Inseparable effects are not necessarily limited to fuels; they may pertain to engine design features as well, and may even pervade the engine/fuel interface. Evaluation of such coupled effects may resist conventional statistical means and present yet another opportunity for exploitation of the vector-variable approach.

THE VECTOR METHODOLOGY

DESCRIPTION OF DATA BASE

A database representing 280 individual emission tests of HDD engines was compiled from nine publications [18-26] where the following criteria were met:

- The EPA transient test cycle was used and either the composite or hot start result was reported. The hot start portion has a 6/7th weight in the composite result.
- At least NO_x and PM emissions were measured, which are the pollutants examined in this study. Eight of the sources measured all four pollutants (HC, CO, NO_x, and PM), and one source measured all except CO.
- Emissions testing could be matched to fuels for which the following 12 properties were known: natural cetane, cetane number improvement (resulting from additives), density, viscosity, sulfur content, mono-aromatic content, poly-aromatic content, and five points on the distillation curve.

Table 1 lists the variables contained in the database; the field names are used to refer to these variables in exhibits to this paper. Overall, the data represent 11 different engines tested a total of 280 times on 85 different diesel fuels.

Twenty-seven publications were examined in the process of compiling this database [17-43]. Eight publications using the EPA transient test cycle were excluded because one or more of the fuel properties was not reported, most commonly the poly-aromatic content. Ten publications were excluded for reasons related to the emissions data. In one instance, only PM was measured, while European or Japanese test cycles were used in nine other instances.

Table 1: The HDD Emissions Database

Field Name	Units	Description
Engine ID	text	text description of engine
Fuel ID	text	text identifier of fuel
Test	number	sequential number
Source	text	SAE paper number
Cycle	number	0=EPA Composite; 1=EPA Hot Start
Engine	number	unique engine identifier
NRepl	number	number of test replications
HC	gm/bhp-hr	hydrocarbon emissions
CO	gm/bhp-hr	carbon monoxide emissions
NOx	gm/bhp-hr	nitrogen oxide emissions
PM	gm/bhp-hr	particulate emissions
NatCetane	number	natural cetane
CetImprv	number	cetane improvement
Density	gm/cm ³	
Viscosity	mm ² /sec	at/near 40 degrees C
Sulfur	ppm	sulfur content
MonoArom	percent	mono-aromatics content
PolyArom	percent	poly-aromatics content
IBP	Celsius	Initial boiling point
T10	Celsius	10% evaporation point
T50	Celsius	50% evaporation point
T90	Celsius	90% evaporation point
FBP	Celsius	Final boiling point

The 12 fuel properties examined here are a super-set of the fuel properties that have been considered with respect to HDD emissions. They include ones such as aromatics content, cetane rating, and sulfur content that a consensus of investigators believes relevant to emissions performance. Additional properties (IBP, T10 and others) are included that could be independent predictors, could be correlated with the consensus variables, or could prove unrelated to emissions. This purposefully casts the net as wide as possible, leaving the identification of the proper subset of predictor variables to the later analysis. However, there is no intent to imply that only these properties could affect engine emissions. For example, fuel oxygen content is likely to affect CO emissions, and perhaps other pollutants, but is not among the selected properties. While some sources tested oxygenated diesel fuels, these were judged to be too few in number to permit including oxygen content in the list.

The eleven HDD engines represented in the database constitute a very small sample of the engine types present in the on-road vehicle fleet. Nevertheless, they include engines made by the three major manufacturers (Cummins, DDC, and Navistar) and cover a range in model years and horsepower ratings. The engines are generally similar in design, although they are built to varying emissions standards. None are equipped with EGR

systems or with catalysts as tested. They can be taken as reflective of the HDD engines currently on the road, even if the sample is too limited to be considered truly representative.

The large majority of the database represents individual engine tests, but 41 entries from three sources record the mean values of replicated tests. For this paper, we have duplicated the mean values so that each is represented as many times as the test replications. This approach, which gives a total of 280 emission tests, maintains an equal weighting among individual tests and improves the estimation of the total variance. However, the total variance is understated by an unknown extent because the data omit the variation of the (unknown) individual test results around the (reported) mean values.

COMPUTER SOFTWARE

The analysis was conducted using MatLab, a commercially available software package designed for matrix processing [44]. MatLab offers a built-in function *svd*, which extracts the eigenvalues and eigenvectors of a matrix using the singular value decomposition procedure. Other statistical procedures such as computation of correlation matrices and multivariate regression analysis are available as built-in functions or can be easily written using matrix notation. The methodology demonstrated here can be implemented in any computational environment that provides for the calculation of matrix eigenvalues and eigenvectors.

ANALYTICAL RESULTS

We first demonstrate how PCA can be used to resolve the matrix of fuels into a representation based on eigenvectors.

different fuels and replicated a varying number of times corresponding to the number of emissions tests in which each was used. This data set can be viewed as a design matrix *X* of dimension 280 rows (test entries) by 12 columns (fuel properties) that contains all of the fuels-related information available as predictors for emissions.

The total fuels-related variance is defined as the total variance within the columns of this matrix (each representing one fuel property). Each fuel property is standardized to a mean of zero and a variance of one to place the contributions on a common, unit-less basis, so that the total variance of the standardized *X* matrix is 12. All subsequent analysis will be based on the standardized variables, which measure fuel properties in terms of standard deviations from the mean properties of the fuels. The standardized variables can be translated back into the original form whenever required.

Table 2 presents the correlation matrix for the test fuels data set. Correlations greater than 0.50 in absolute magnitude (an arbitrary threshold) have been highlighted to emphasize the inter-relatedness of the physical properties of real diesel fuels. For example, and not surprisingly, the five points on the distillation curve are highly correlated with each other and with viscosity. Other correlations reflect known relationships encountered in fuel blending. Increased natural cetane is correlated with reduced density and aromatic content as would be expected in a fuel blend where the proportion of high-aromatic cracked stocks, which have high density and low cetane, has been reduced. That fuel properties are covariant, sometimes to a large degree, is an unavoidable reality and implies that there are fewer *independent* variables than the number of physical properties measured.

Table 2: Correlation Matrix for the Test Fuel Properties

	1	2	3	4	5	6	7	8	9	10	11	12	
NatCetane	1	1.000											
CetImprv	2	-0.233	1.000										
Density	3	-0.613	0.105	1.000									
Viscosity	4	0.219	0.051	0.460	1.000								
Sulfur	5	-0.022	-0.220	0.202	-0.054	1.000							
MonoArom	6	-0.633	0.247	0.667	0.121	-0.030	1.000						
PolyArom	7	-0.393	-0.040	0.523	-0.084	0.511	0.298	1.000					
IBP	8	0.097	-0.058	0.292	0.514	-0.063	0.074	-0.029	1.000				
T10	9	0.225	-0.098	0.444	0.900	0.016	0.071	0.006	0.622	1.000			
T50	10	0.275	-0.002	0.497	0.889	0.014	0.144	0.097	0.382	0.792	1.000		
T90	11	0.295	0.114	0.307	0.692	-0.038	0.226	0.144	0.237	0.523	0.775	1.000	
FBP	12	0.211	0.168	0.318	0.607	-0.096	0.224	0.118	0.278	0.445	0.633	0.897	1.000

Then, we will demonstrate the use of eigenvectors in regression analysis.

Vector Approach to Representing Fuels

Consider the subset of data describing the fuels used in emissions testing. This test fuels data set consists of the 12 properties selected for study as measured for the 85

To demonstrate this, a singular value decomposition analysis was performed to extract the 12 eigenvalues and eigenvectors from the correlation matrix. The eigenvectors are defined in the computational procedure in a manner that partitions the total variance into orthogonal components, each eigenvalue being a measure of the variance associated with the corresponding eigenvector.

In this context, orthogonality means that the eigenvectors are linearly independent of each other and, as a result of their definition, the correlation between any two eigenvectors over the data set is exactly zero.

Table 3 presents the twelve eigenvalues and eigenvectors of the test fuels data set. Each eigenvector is a linear combination of the original 12 fuel properties. For example, the first eigenvector is described by the coefficients or weights (0.061, 0.034, 0.285, ..., 0.365) applied to the fuel properties (natural cetane, cetane improvement, density, ..., FBP). The largest coefficients have been highlighted to emphasize the most important fuel property components.

The variance among fuels, indicated by the eigenvalues, is highly concentrated in the first few eigenvectors. The first

The following discussion interprets the first four eigenvectors in terms of the associations among fuel properties. Where possible, we have suggested identifications of the eigenvectors with known refinery or blending processes. These largest eigenvectors, as identified by the proportion of the total variance shown in parentheses, are likely to represent generalized characteristics of fuels and therefore to be most amenable to variation in reformulating diesel fuels.

Primary viscosity/density characteristic (38%). A direct relationship among viscosity, distillation temperatures, and to a lesser extent density. This is associated with the largest eigenvalue, meaning that the test fuels vary most among themselves with respect to this characteristic. More viscous compounds found in diesel fuels have higher boiling points, and predictive equations show that viscosity

Table 3: Eigenvectors of the Test Fuels Data Set

	1	2	3	4	5	6	7	8	9	10	11	12
NatCetane	0.061	-0.556	0.163	-0.220	0.071	-0.068	0.138	-0.458	0.104	-0.456	0.045	0.391
CetImprv	0.034	0.143	-0.549	-0.212	0.782	0.061	-0.024	-0.106	0.001	-0.054	-0.048	0.004
Density	0.285	0.449	0.049	0.168	-0.047	0.163	-0.109	0.237	0.343	-0.418	-0.272	0.473
Viscosity	0.432	-0.120	-0.017	0.142	0.084	0.308	0.004	0.150	-0.156	0.292	0.638	0.371
Sulfur	0.002	0.180	0.636	-0.202	0.393	0.118	0.556	0.175	0.013	0.058	-0.024	-0.100
MonoArom	0.146	0.464	-0.256	0.040	-0.305	-0.028	0.548	-0.500	-0.134	-0.078	0.164	-0.037
PolyArom	0.076	0.418	0.385	-0.272	0.092	-0.289	-0.552	-0.343	-0.171	0.030	0.225	0.049
IBP	0.262	-0.072	0.052	0.537	0.248	-0.697	0.118	-0.033	0.234	0.128	0.015	-0.024
T10	0.399	-0.113	0.128	0.316	0.109	0.192	-0.093	-0.123	-0.642	-0.186	-0.392	-0.198
T50	0.431	-0.083	0.064	-0.053	-0.028	0.319	-0.149	-0.240	0.555	0.010	0.065	-0.552
T90	0.392	-0.074	-0.075	-0.428	-0.158	-0.134	0.071	-0.037	0.021	0.551	-0.486	0.252
FBP	0.365	-0.048	-0.147	-0.409	-0.141	-0.359	0.069	0.476	-0.157	-0.406	0.205	-0.254
Eigenvalues	4.531	2.591	1.549	1.181	0.681	0.564	0.390	0.230	0.140	0.083	0.036	0.026
Pct Variance	37.75	21.59	12.90	9.83	5.67	4.69	3.24	1.92	1.17	0.68	0.29	0.21
Cumulative Pct	37.75	59.34	72.25	82.09	87.76	92.46	95.70	97.62	98.79	99.48	99.78	100.0

eigenvector accounts for nearly 40 percent of the total variation among the fuels, the first six together account for more than 90 percent, and the first nine for essentially all (nearly 99 percent). Thus, while the data set contains 12 distinct variables, its total variance is concentrated in a much smaller number of orthogonal patterns as described by the eigenvectors with the largest eigenvalues.

It is often desirable to develop a conceptual interpretation of the eigenvectors to aid the analyst's understanding, although this may not be completely possible in complex systems. Physical systems (of any kind) are normally created from more basic building blocks according to a set of rules that reflect a natural structure. If these building blocks are fully described by the chosen set of variables, one hopes to find an expression of this structure in the eigenvectors. In the context of diesel fuels, the underlying structure (and therefore the eigenvectors) should reflect the properties of the refinery processes and blending stocks used to create these fuels.

is directly related to the square root of density [45]. Diesel blend stocks exhibit a similar relationship among viscosity, distillation temperatures, and density as demonstrated independently by correlation analysis using the database of blend stocks in the Refinery Yield Model (RYM) maintained by Oak Ridge National Laboratory.

Primary aromatics characteristic (22%). An increase in aromatics content (both mono- and poly-aromatic) is associated with higher density and a decrease in natural cetane. This reflects a known property of the high-aromatic cracked stocks that are used in blending diesel fuels; these stocks have higher densities and their aromatics content is known to delay ignition and therefore decrease cetane rating.

Primary sulfur/quality characteristic (13%). This appears to represent sulfur content and its related impact on the boost from cetane improvers, which declines as the quality of diesel fuel declines. Information from the Ethyl Corporation [46] shows that cetane boost is reduced with

lower clear cetane, as for fuels with higher sulfur and polyaromatics.

Primary blend balancing characteristic (10%). Fuels with increased temperatures at the low end of the curve (IBP and T10) tend to be associated with decreased temperatures at the upper end (T90 and FBP). This slope characteristic for the distillation curve may be related to meeting blending specifications. For example, flash point might be satisfied by using heavier blend stocks at the low end of the distillation temperatures, while lighter blend stocks are used on the high end to meet the pour point requirement.

When the variance associated with individual eigenvectors falls to relatively small percentages, the eigenvectors may begin to reflect factors specific to the blending of test fuels in individual sources, rather than characteristics found in a range of fuels. For example, the smallest eigenvectors could reflect specific blend stocks used in one or more sources to vary fuel properties for test purposes or to control one or more fuel properties to fixed values, once another property had been varied for experimental purposes. For this reason, we do not attempt to offer physical interpretations for the smaller eigenvectors. Overall, more work is needed to understand the eigenvector characteristics of diesel fuels, particularly as those characteristics may differ between commercially available fuels and test fuels created for use in the laboratory.

Because the eigenvectors form an orthogonal basis, they can be used to re-express the original matrix in orthogonal terms. This process is closely analogous to Fourier transform analysis, in which a time-varying signal is decomposed into individual frequencies and then re-

expressed as a weighted sum over frequencies. In Fourier analysis, the continuum of harmonic frequencies from $\bar{\omega} = 0$ to $\bar{\omega} \rightarrow \infty$ forms an orthogonal basis from which any time-varying signal can be constructed. In the vector approach defined here, the basis vectors are developed from the experimental data at hand in a manner expressly defined to be orthogonal. Perhaps more familiar to data analysts, the eigenvector approach is also similar to orthogonal polynomials, of which it is a direct generalization.

An experimental design matrix $X_{(m \times n)}$ of m rows and n variables can be represented in eigenvector terms as the linear combination $A_{(m \times k)} * V'_{(k \times n)}$, where $A_{(m \times k)}$ is a matrix of coefficients for the k eigenvectors and $V_{(n \times k)}$ is a matrix in which the eigenvectors, composed of n components each, form the columns. The coefficients $A_{(m \times k)}$ are calculated from the relationship $A_{(m \times k)} = X_{(m \times n)} * V_{(n \times k)}$. In algebraic form, any row m of the X matrix can be expressed as a linear combination of coefficients $a_m(k)$ and eigenvectors $v_j(k)$:

$$X_m(j) = a_m(1)*v_j(1) + \dots + a_m(12)*v_j(12) \quad (1)$$

where $X_m(j)$ = value of the j^{th} variable (fuel property) for the m^{th} fuel; $a_m(k)$ = coefficient of eigenvector k in the m^{th} fuel ; and $v_j(k)$ = component weight for the j^{th} variable in eigenvector k .

The example in Table 4 may help to make these relationships more understandable. Here, the observed

Table 4: Eigenfuel Representation for a Selected Fuel

Fuel Property	1	2	3	4	5	6	7	8	9	10	11	12	
Observed Values	-0.305	-0.535	0.375	0.025	-0.341	0.042	0.005	-0.083	-0.123	0.331	0.154	-0.268	
Calculated Values	-0.305	-0.535	0.375	0.025	-0.341	0.042	0.005	-0.083	-0.123	0.331	0.154	-0.268	
k	Coefficient	Eigenvector Components											
1	0.121	0.061	0.034	0.285	0.432	0.002	0.146	0.076	0.262	0.399	0.431	0.392	0.365
2	0.213	-0.556	0.143	0.449	-0.120	0.180	0.464	0.418	-0.072	-0.113	-0.083	-0.074	-0.048
3	0.066	0.163	-0.549	0.049	-0.017	0.636	-0.256	0.385	0.052	0.128	0.064	-0.075	-0.147
4	0.259	-0.220	-0.212	0.168	0.142	-0.202	0.040	-0.272	0.537	0.316	-0.053	-0.428	-0.409
5	-0.632	0.071	0.782	-0.047	0.084	0.393	-0.305	0.092	0.248	0.109	-0.028	-0.158	-0.141
6	0.229	-0.068	0.061	0.163	0.308	0.118	-0.028	-0.289	-0.697	0.192	0.319	-0.134	-0.359
7	-0.294	0.138	-0.024	-0.109	0.004	0.556	0.548	-0.552	0.118	-0.093	-0.149	0.071	0.069
8	0.011	-0.458	-0.106	0.237	0.150	0.175	-0.500	-0.343	-0.033	-0.123	-0.240	-0.037	0.476
9	0.371	0.104	0.001	0.343	-0.156	0.013	-0.134	-0.171	0.234	-0.642	0.555	0.021	-0.157
10	0.205	-0.456	-0.054	-0.418	0.292	0.058	-0.078	0.030	0.128	-0.186	0.010	0.551	-0.406
11	-0.119	0.045	-0.048	-0.272	0.638	-0.024	0.164	0.225	0.015	-0.392	0.065	-0.486	0.205
12	0.048	0.391	0.004	0.473	0.371	-0.100	-0.037	0.049	-0.024	-0.198	-0.552	0.252	-0.254

values¹ have been taken from a selected observation in the X matrix. Below this, the calculation given by Equ. 1 is shown to exactly reproduce the original observation. The values are calculated, for any fuel property (column) j, as the product of the coefficient a_k times the coefficient for the jth component of eigenvector k, summed over all eigenvectors $k = 1, 2, \dots, 12$. The eigenvectors k are placed in row form in this table, while the fuel properties j form the columns.

Thus, we can express any fuel as the vector of coefficients $a_m(k) = (a_1, a_2, \dots, a_{12})$ corresponding to the 12 eigenvectors instead of describing the fuel by its physical properties. Because this relationship looks much like a blending equation, and the eigenvectors have been shown to have physical interpretations, we adopt the terminology *eigenfuel* in place of *eigenvector*. We then treat fuels as *mathematical blends* of eigenfuels, each of which represents a distinct, mathematically independent characteristic. The coefficients $a_m(k)$ become measures of how fuel m is composed of the eigenfuels k.

Once a data set is translated into this representation, the eigenfuel coefficients are distributed with mean zero and variance equal to the corresponding eigenvalue. Figure 1 shows histograms of the coefficients $a_m(k)$ for the data set used here. The coefficient distributions are broad (have large variance) for the first several eigenfuels, consistent with their large eigenvalues. The distributions narrow as one moves through the series, until they approach a peak clustered about zero by the end. Thus, the fuels vary most widely with respect to the characteristics expressed by the first several eigenfuels and differ only to a very minor

extent with respect to those represented by the later eigenfuels.

A second fundamental property is that the eigenfuels form an orthogonal set and their coefficients are uncorrelated -- i.e., the off-diagonal elements of the correlation matrix for the coefficients $A_{(m \times k)}$ are zero. That the eigenfuel coefficients are mathematically independent and uncorrelated will prove very important when they are used as predictor variables.

Use of Eigenfuels in Regression Analysis

Having reviewed the properties of eigenfuels, we now turn to their use as predictor variables in regression analysis. The empirical relationship between engine emissions and fuel properties is usually determined through regressing an emissions variable Y_e against one or more fuel property variables X_i in a form similar to:

$$Y_e = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n \quad (2)$$

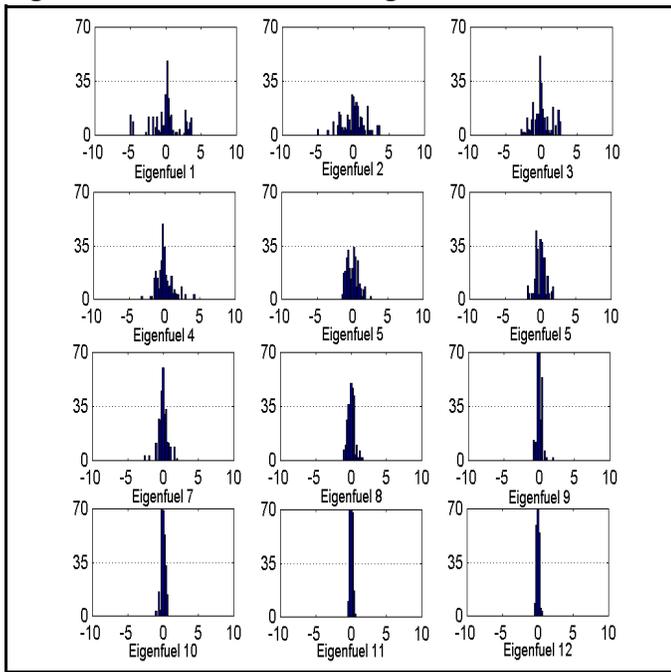
where the coefficients b_i are determined by the regression, and we consider the variables X_i to be scalar quantities.

In the vector approach developed here, the regression model will be of comparable form:

$$Y_e = b_0 + b_1 A_1 + b_2 A_2 + \dots + b_n A_n \quad (3)$$

where the new variables A_i are the coefficients of eigenfuel i in the vector representation. Consistent with other work, the dependent variable Y_e is taken to be the natural logarithm of emissions.

Figure 1. Distribution of the Eigenfuel Coefficients



¹ These are the physical properties in standardized form where mean = 0 and variance = 1.

When used as predictor variables in regression analysis, eigenfuels have two important properties that result from their mathematical independence, as demonstrated in Table 5. Here, NO_x emissions have been regressed against each eigenfuel k individually using equations of the form:

$$\ln(NO_x) = a_0 + a_1 A_k \text{ for } k = 1, 2, \dots, 12 \quad (4)$$

The regression sum of squares and coefficient values are then tabulated against the results of a regression in which all 12 eigenfuels are present simultaneously (see the rightmost column of the table). The regression sum of squares summed across the twelve individual regressions equals the sum of squares in the regression containing all 12 eigenfuels, and the intercept and eigenfuel coefficients for the individual regressions are identical to those estimated in the regression containing all eigenfuels. Thus, the regression sums of squares are *additive* and the coefficient values are *invariant* with respect to the selection of eigenfuels for inclusion in the regression. There is, in fact, a unique partitioning of the variance in the dependent variable into the components identified with the eigenfuels. It can be shown that the contribution of eigenfuel k to the regression sum of squares and R^2 statistic is proportional

Table 5: NO_x Regressions using Eigenfuels as Explanatory Variables

Eigenfuels included in Regression													
	1	2	3	4	5	6	7	8	9	10	11	12	ALL
Regression Sum of Squares	.0236	.7805	.0010	.0506	.1034	.0007	.0148	.0000	.0547	.1229	.0113	.0000	1.1633
Cumulative SS	.0236	.8041	.8051	.8557	.9591	.9597	.9745	.9745	1.0292	1.1521	1.1633	1.1633	
Intercept	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372	1.5372
Eigenfuel 1	.0043												.0043
Eigenfuel 2		.0329											.0329
Eigenfuel 3			.0016										.0016
Eigenfuel 4				.0124									.0124
Eigenfuel 5					-.0233								-.0233
Eigenfuel 6						.0021							.0021
Eigenfuel 7							.0117						.0117
Eigenfuel 8								.0002					.0002
Eigenfuel 9									-.0374				-.0374
Eigenfuel 10										-.0730			-.0730
Eigenfuel 11											-.0335		-.0335
Eigenfuel 12												.0012	.0012

to the product of the eigenvalue λ_k and the square of the regression coefficient b_k .

This outcome contrasts with the usual result in regression analysis when the predictor variables are correlated with each other. The multi-collinearity existing in such circumstances means that parameter estimates change when predictors are added to or removed from the regression. In addition, combining individual variables to create a pooled regression does not increase the regression sum of squares and R^2 statistic to the extent that might be expected. Working through the “fog” created by multi-variables is part of the art of regression analysis and is one of the reasons why independent analysts can reach differing conclusions from the same data. The use of the linearly independent, vector variables eliminates this fog.

The table also begins to indicate insights that will be developed in the next section. Not surprisingly, there is a difference between the *importance* of an eigenfuel in describing the variation among fuels and the *strength* of its relationship to NO_x emissions or another independent variable. Eigenfuels 1 through 12 are defined in decreasing order of variance among the fuels, so that eigenfuel 1 accounts for 38% of the fuel variance, followed by eigenfuel 2 at 22%, and eigenfuel 10 at only 0.7%. However, the regression results indicate that eigenfuel 10 has the strongest relationship to NO_x emissions, as measured by its coefficient, followed by eigenfuels 11 and 2. We also see that some of the eigenfuels (numbers 1, 3, 6, 8, and 12) have very weak relationships to NO_x and are likely candidates to drop from the analysis. However, we must first consider and control for other factors that contribute to the variance in emissions before attempting to draw conclusions from such results.

Application to Diesel Emissions

In this section we show the application of the vector approach to diesel engine emissions and obtain a first look at its implications. There are many factors beyond fuel composition that contribute to the variance in engine emissions, including differences among engines, test cycles, and the sources from which the data are drawn. The intent is to extract these fixed effects and then recompute the regression equation involving the

eigenfuels. This will be done for both NO_x and PM emissions.

Figure 2 suggests the sources of variation likely to be found in the database of diesel engine emissions data. The tested engines are taken to be generally reflective of the population of HDD engines currently on the road. Nine different publications reported tests for 11 individual engines, representing 11 different engine designs, on 85 different fuels using one of two different EPA test cycles. Most, but not all, engine tests were replicated at least once.

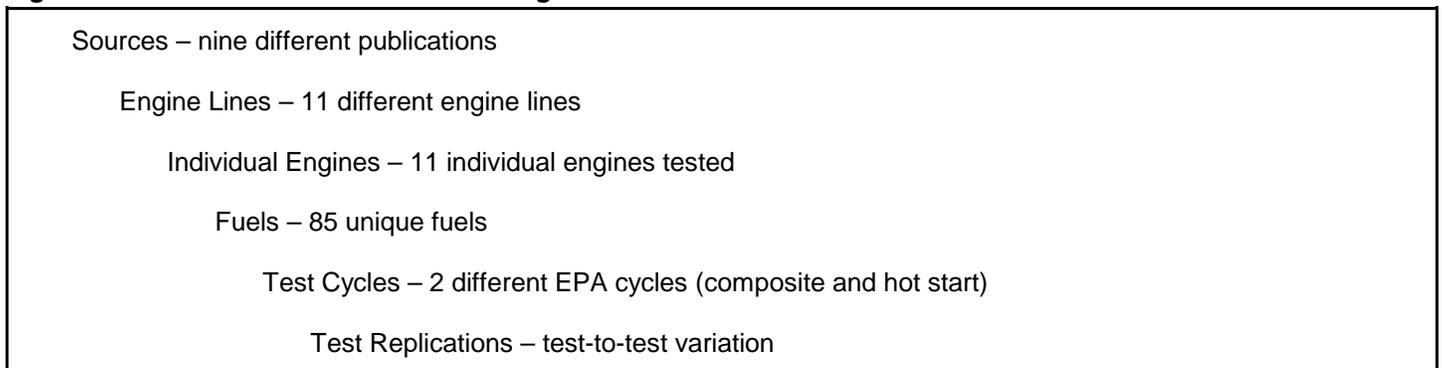
We can hypothesize a series of terms in the overall emissions model that represent, for example, the average emissions level E_0 of the existing fleet, the variation of the average emissions E_i for engine design i around E_0 , and the variation of the average emissions E_{ij} for individual engine j around its engine design average E_i . Other terms in the model would include an effect S_k for differences among the sources and an effect T_l for the different average emissions levels of the two EPA test cycles. This series of terms would be in addition to the effect of fuels on emissions, which is of primary interest. We are unable, however, to estimate an overall emissions model at present because of size and coverage limitations of the database that make it impossible to separate the effects of engine designs, individual engines, test cycles, and sources. Each engine design is represented by only a single specimen and each individual engine has been tested using only one of the two test cycles.

For purposes of this exploratory study, we have incorporated a single fixed effect for individual engines in the regression models. This engine effect represents an undifferentiated, composite effect due to engine designs, individual engines, sources, and test cycles. Thus, we use regression equations of the form:

$$\ln(E_{i,j}) = b_0 + \sum b_i * \delta_i + \sum b_{j,k} * A_{j,k} \tag{5}$$

where the dependent variable is the natural logarithm of emissions for engine i tested on fuel j , $\sum b_i * \delta_i$ represents a dummy variable formulation for the variation in mean emission levels among individual engines $i = 2, \dots, 11$ and $\sum b_{j,k} * A_{j,k}$ represents the emission effects of fuel j expressed in terms of the 12 eigenfuel coefficients.

Figure2: Sources of Variance in Diesel Engine Emissions Data



While the eigenfuels are defined to be mathematically independent of each other, correlations exist between the eigenfuels and the engine dummy variables. As a result, there is no unique partitioning of the variance between fuel and engine effects. Fuel effects should be computed within engines and the separate estimates pooled. Otherwise, differences between emission levels for the various engines can affect the fuel estimates. This separation of fuel and engine effects is achieved by computing the model sum of squares with both engine and fuel variables included and then with only engine variables included. The total model sum of squares with fuel and engine variables, less the model sum of squares when the model is constrained to engine effects only, is the sum of squares attributed to fuels. This assigns to fuels only the sum of squares reduction that can be uniquely associated with fuels and is referred to as “fuels adjusted for engine effects.”

As shown in Table 6, the combination of engine effects (representing the composite of engine designs, individual engines, sources, and test cycles) and the fuel effects explain 91.1 and 98.6 percent of the sum of squares for NO_x and PM, respectively. This suggests that the variability of test-to-test replication (for a given engine and fuel) is relatively small compared to the differences among engines and fuels within this database. The engine effects explain 45.5 percent of the sum of squares for NO_x and

95.4 percent for PM, while the fuel effects represented by the eigenfuel terms explain 45.6 percent and 3.2 percent respectively. The importance of engine effects for PM emissions, while real, is greatly increased by one, older engine whose PM emissions are much above the others.

It is well known that engine design factors have important effects on emissions and it is all the more to be expected when, as in present circumstances, the vehicles were designed to varying certifications standards. Further, engine and fuel effects are correlated in this data set – to a substantial extent for PM – so that we are not able to clearly separate their contributions at present. We take these preliminary results to suggest that fuels may have substantial effects on engine emissions, but further work with additional data is clearly needed to resolve the competing importance of engines and fuels. This paper focuses primarily on the relative contributions made by the eigenfuels to the portion of the total emissions variation that can be attributed to fuels.

Table 7 summarizes the NO_x and PM regressions. Inspection of the table reveals that all of the engine effects are statistically significant at the 0.05 level (t value exceeding 1.96). Among the fuel effects, all but those for eigenfuels 6, 8, and 9 for NO_x and for eigenfuels 4, 6, 7, 8, 11, and 12 are significant at the 0.05 level. Thus, many fuel effects might be retained in the model if the selection

Table 6: Sum of Squares for Fuels Adjusted for Engine Effects

	NO_x EMISSIONS			
Source of Variation	SS	DF	MS	R²
Regression SS	1.6334	22	0.0742	0.911
Engine SS (Unadjusted)	0.8156	10	0.0816	0.455
Fuel SS (Adjusted for Engines)	0.8178	12	0.0818	0.456
Error SS	0.1602	257	0.0006	0.089
Total SS	1.7936	279	0.0064	1.000
	PM EMISSIONS			
Source of Variation	SS	DF	MS	R²
Regression SS	104.2	22	4.736	0.986
Engine SS (Unadjusted)	100.8	10	10.080	0.954
Fuel SS (Adjusted for Engines)	3.4	12	0.283	0.032
Error SS	1.5	257	0.006	0.014
Total SS	105.7	279	0.379	1.000

were based on statistical significance. However, as we argue, it is the predictive capability of an effect that should guide its selection or rejection.

Table 7: Summary of Regression Results for NO_x and PM

Parameter	Ln(NO _x) Emissions		Ln(PM) Emissions	
	Estimate	t ratio	Estimate	t ratio
Engines				
Intercept	1.5229	248.90	-0.7683	40.15
Engine 2	0.0217	3.19	-0.6084	28.62
Engine 3	-0.0308	4.63	-0.5929	28.53
Engine 4	0.0293	3.57	-0.9771	38.03
Engine 5	0.0643	5.84	-0.7530	21.87
Engine 6	0.0339	3.98	-1.7223	64.68
Engine 7	-0.1082	7.82	-0.9844	22.75
Engine 8	0.0670	6.79	-1.4494	47.01
Engine 9	0.0413	4.07	-1.6322	51.49
Engine 10	-0.1323	11.59	-1.6107	45.11
Engine 11	0.0351	3.18	-0.8002	23.17
Fuels				
Eigenfuel 1	0.0043	5.30	0.0233	9.23
Eigenfuel 2	0.0344	30.78	0.0549	15.72
Eigenfuel 3	0.0051	2.46	0.0595	9.24
Eigenfuel 4	0.0120	7.09	-0.0014	0.26
Eigenfuel 5	-0.0149	7.62	0.0150	2.45
Eigenfuel 6	0.0027	0.98	0.0013	0.16
Eigenfuel 7	0.0173	5.87	0.0134	1.45
Eigenfuel 8	-0.0031	0.66	0.0050	0.35
Eigenfuel 9	-0.0057	1.22	-0.0430	2.93
Eigenfuel 10	-0.0156	2.59	-0.0575	3.05
Eigenfuel 11	0.0294	2.85	0.0239	0.74
Eigenfuel 12	0.0325	2.62	-0.0318	0.82

The “predictive capability” statistic, computed by normalizing the quantity $\lambda_k * b_k$ to a value of one, identifies the relative contribution of each eigenfuel to the predictive

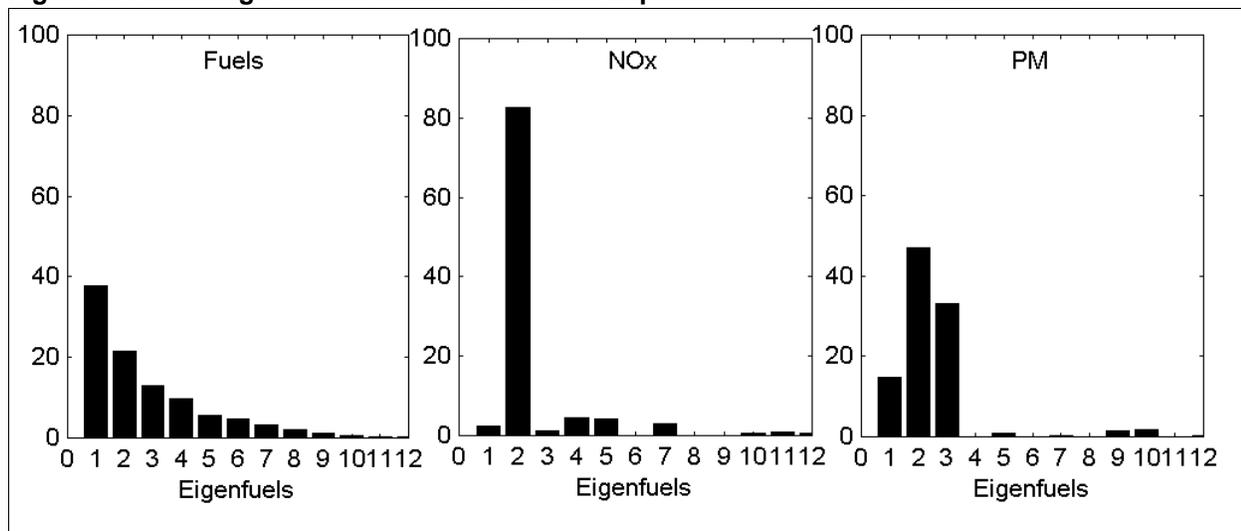
power contributed by all fuels-related information. Using this statistic, Figure 3 shows that only one eigenfuel (number 2) for NO_x and three eigenfuels (numbers 1, 2, and 3) for PM account for nearly all of the predictive power that can be ascribed to fuels. This figure also demonstrates, by comparison to the variance explanation for fuels, why all eigenfuels should initially be retained and considered for deletion only after their relationship to the response variable has been determined. Overall, these results mean that, regardless of the statistical significance of other coefficients, the regression models could be reduced to include only 1 or 3 eigenfuel terms (in addition to the engine effects) without a significant reduction in the ability to capture the impact of fuels.

Let us now briefly examine the substantive meaning of the regression results. For NO_x, only eigenfuel number 2 has substantial predictive power. This eigenfuel was previously described as representing a primary aromatics characteristic, in which an increase in aromatics content (both mono-aromatics and poly-aromatics) was associated with increased density and decreased natural cetane.

From a refinery perspective, this was identified as representing high-aromatic cracked stocks. Exponentiating the individual terms of the regression equation, the results indicate that NO_x emissions are decreased by a factor of $\exp(0.0344) - 1 = 3.5$ percent for each unit reduction in this fuel characteristic. Because the variance associated with the second eigenfuel is 2.591, a unit reduction corresponds to $1.000/\sqrt{2.591} = 0.62$ standard deviations. Therefore, a one standard deviation reduction corresponds to reducing NO_x emissions by 5.6 percent. A reduction by one standard deviation corresponds to approximately one-third of the total change that is possible and is used here as a rule-of-thumb measure of what might be possible to achieve in practice.

This eigenfuel expresses three individual fuel properties that are widely believed to influence NO_x emissions – aromatics content, natural cetane, and density. However, it represents a single mode of variation involving

Figure 3: Percentage Contributions to Variance Explanation for Fuels and Emissions



simultaneous changes in the three variables. Thus, it may be more correct to speak of reducing NO_x emissions by decreasing the content of high-aromatic cracked stocks, rather than by varying any of the three properties independently.

The results are somewhat more complicated for PM, since three of the eigenfuels – numbers 2, 3, and 1 in order – are found to contribute substantially to the fuels-related predictive power. The most important, eigenfuel number 2, was identified with the content of high-aromatic cracked stocks. Using the calculation shown above, a one standard deviation reduction in this eigenfuel corresponds to a $(\exp(0.0549)-1)*\sqrt{2.591} = 9.1$ percent reduction in PM. Eigenvector 3, involving a primary association between sulfur content and cetane boost, corresponds to a 7.6 percent PM reduction for each standard deviation change. Eigenvector 1, involving a primary viscosity and distillation curve characteristic, corresponds to a 5.0 percent reduction per standard deviation. The effects are additive, since the eigenfuels are independent, and would reach a total of 21.7 percent if a one standard deviation reduction were made in all three. All other eigenfuels make negligible contributions, whether they are found to be statistically significant or not.

As indicated in the review by Lee, Pedley, and Hobbs, there is only a weak consensus on how fuel properties affect PM emissions, except that reducing sulfur content is generally accepted to reduce PM emissions. Eigenvector 3 appears to express this consensus relationship. The properties involved in the other eigenfuels (1 and 2) include aromatics content, natural cetane, viscosity, the distillation curve, and to a lesser extent density. Density and poly-aromatics content are thought to have a small effect on PM emissions in some engine groups, while there is no consensus on whether cetane, mono-aromatic content, viscosity, or distillation curve parameters are important. We cannot resolve these points of potential difference based on the current database and analysis. However, the results presented here suggest there is more than one way in which PM emissions can be reduced.

Note also that several of the smaller eigenfuels (numbers 11 and 12 for NO_x and 9 and 10 for PM) have strong relationships to emissions. These smaller eigenfuels point toward the possibility that unexploited modes of reducing NO_x and PM emissions could exist, although the current state of the analysis and the size and scope of the engine emissions database are inadequate to draw conclusions on their meaning or potential emissions importance. These findings could result from correlations to factors that have been inadequately controlled in these regressions. Alternately, even if the relationship to emissions is real, it could prove impossible or impractical to blend fuels that vary significantly in these characteristics because of considerations of cost, safety, driveability, or other reasons.

Model Interpretation and Simplification

Investigators accustomed to scalar predictor variables may perceive vector variables to be needlessly complex and may find it difficult to interpret response in terms of eigenvectors. In the psychological and social sciences, this objection may be particularly *apropos* because of the lack of underlying theory as a logical assist. In the physical sciences, however, interdependent variables may often be readily recognized. Certainly, in the present instance, blending experience was a valuable interpretational aid.

Nevertheless, a model for predicting emissions in terms of fuel characteristics can not be considered complete until all means of simplification have been explored. Therefore, simplification procedures have been an important concern from the beginning and will be a major part of our continuing effort.

Though well aware of rotation procedures (varimax, quartimax and equimax), our experience with these procedures failed to provide any significant insights in the context of our problem. Instead, with regard to emissions response, which is our main concern, we elected to pursue a direction aimed at better understanding the relation between eigenvectors and their fuel-property components.

A very useful outcome is a scheme for re-expressing the SS partitioning among eigenvectors as a partitioning among the underlying fuel variables. The conversion is made possible, of course, by virtue of the fact that each eigenvector can be expressed as a linear combination of the original variables. Since it corresponds directly to the eigenvector partitioning, the derived partitioning among variables is in a sense unique. Certainly it has a claim to distinction from the multiple partitionings that can arise in stepwise regression or when all possible subset models are considered.

The SS partitionings are shown in Table 8 for NO_x and PM, both by eigenvectors and by the original fuel variables. The difference in partitioning for the two pollutants is evident and again emphasizes the futility of attempting to select variables independently of the role they play as predictors. We now have at our command a means for simplifying the emission models. As noted earlier, we recommend a combination of statistical tests of significance and evaluation of the magnitude of effects.

Eigenfuels can be rejected, as is usually done in OLS, by dropping those that fail to meet a specific level of significance. At the 0.05 level, those eigenfuels tagged * in the table would be removed. However, it is to be noted that several other eigenfuels, tagged ** in the table, each contribute less than 1 percent to the model SS. On the assumption that elimination of these eigenfuels would have little influence on the predictive capability of the model, they

Table 8: Statistical Significance and SS Partitioning for NO_x and PM

Eigenvector	log(NO _x) Emissions		log(PM) Emissions	
	t ratio	Model SS	t ratio	Model SS (%)
1	5.30	2.24	9.23	14.8
2	30.78	82.49	15.72	47.1
3	2.46	1.07	9.24	33.1
4	7.09	4.56	0.26*	0.0**
5	7.62	4.08	2.45	0.9**
6	0.98*	0.11**	0.16*	0.0**
7	5.87	3.15	1.45	0.4**
8	0.66*	0.06**	0.35*	0.0**
9	1.22*	0.12**	2.93	1.6
10	2.59	0.54**	3.05	1.7
11	2.85	0.84**	0.74*	0.1**
12	2.62	0.73**	0.82	0.2**
NatCetane		19.28		4.98
CetImprv		1.55		1.55
Density		26.55		17.72
Viscosity		0.24**		0.59**
Sulfur		3.60		31.87
MonoArom		38.14		6.61
PolyArom		8.02		27.94
IBP		0.06**		0.32
T10		0.01**		6.13
T50		0.66**		0.74**
T90		1.72		0.25**
FBP		0.18**		1.30

can be considered for removal, because it is irrelevant whether they are statistically significant or not.

The disposition of eigenfuels on the basis of statistical significance is, of course, dependent on the significance level adopted. Similarly, the cutpoint for practical significance is arbitrary. It is here that art and judgement, tempered by experience, enter just as it does in any other scheme for selecting variables.

It is not the intent of this paper to resolve these issues but rather to illustrate a methodology for their resolution. We suggest, therefore, a possible scenario for model simplification: Reject those eigenvectors that have a t-ratio smaller than 1.96, corresponding to 0.05 significance for large samples. Renormalize the percent SS and transform the eigenvector contributions to the contributions to SS by individual variables. Now, “prune” those components that contribute less than 1 percent to the model SS.

In this way we have retained predictive components that are both statistically and practically significant, given the criteria used for choice. It removes the objection frequently raised in connection with PCR, namely that the model

retains all variables and therefore effects no simplification. The simplified model makes interpretation of the eigenvectors easier while retaining the important parts of the “eigenstructure” that actually drive the system.

In addition, the scenario has isolated those fuel variables that play important roles in prediction. The model can be restated, if desired, in those terms, in which case the relative contributions of the variables can be appraised with the knowledge that the partitioned SS originate in the “clean” eigenvector environment as opposed to the multicollinear environment of stepwise or all-subset regression.

Further refinement can be had by recomputing the eigenvectors in the reduced-variable space, and other

rejection strategies can be envisioned. Our work is continuing in these areas but is beyond the scope of this paper.

PERSPECTIVE AND DISCUSSION

Though demonstrated in the context of emissions from HDD engines, the vector approach to regression analysis is believed to be applicable to a wide range of problems. Indeed, it should be considered in any circumstance where variables are inextricably covariant. It may be the favored approach whenever the covariance of predictor variables cannot reasonably be "broken," but it does not preclude the use of other multivariate methodologies or of scalar predictor variables when the predictor variables can be varied independently of each other in circumstances such as balanced experimental designs.

The development of a vector emissions model in this paper is admittedly simplistic and does not cover many of the difficulties that can be expected to arise in practice. We are well aware that the database used in this demonstration has many shortcomings. It is a pooling of separate studies each of which was performed with a specific objective in mind. Vehicle sampling was inadequate and did not allow estimation of effects attributable to specific engine characteristics. Further, the vector approach presented here does not as yet address a series of methodological refinements that may be needed to meet the challenges of real-world data.

We urge caution in interpreting results, therefore, and stress again that the major purpose of this paper is to demonstrate methodology, rather than to draw firm conclusions regarding the fuel/emissions relationship. In the following sections, we endeavor to identify needed methodological refinements and to set forth some of the steps in developing an emissions model capable of expressing diesel emissions in terms of engine and fuel characteristics.

METHODOLOGICAL REFINEMENTS

Ongoing work is extending this approach in the following areas.

Statistical Inference

The availability of personal computers and "canned" software has provided us with routine methods to interpret the outcome of experiments. The 0.05 level of significance for testing a null hypothesis goes essentially unchallenged in a world where risk, and the consequences of risk, are anything but constant. Still, this icon is comforting because it assures us that we will erroneously reject the null hypothesis (a Type I error) only one time in twenty, and that seems like very good odds. On the other hand, if the effect being rejected actually is real and not an artifact of sampling, we will be wrong 19 times out of 20 (a Type II error), and that seems like very poor odds. There is a whole continuum of tradeoffs that can be made in arriving

at an optimum policy for managing Type I versus Type II risks for the problem under consideration.

But what impact do these various options have on our ability to predict real-world changes in response? Much depends on sample size. If the regression is based on several hundred observations and if the error standard deviation is sufficiently small, it may be possible to declare an effect statistically significant even though its *magnitude*, so far as accuracy of prediction is concerned, may be negligibly small. On the other hand, if sample size is small, we may erroneously accept the null hypothesis unless the effect being tested is, in fact, fairly large. The bottom line is that, as sample size varies from one investigation to another, so will the magnitude of what we can declare to be statistically significant under a fixed probability of Type I error.

An alternative approach is to fix the magnitude of the effect that we would be willing to ignore, rather than fixing the probability of Type I error. Then, as sample size varies from problem to problem, it is the Type I error probability that would vary; if the effect is considered negligible, the associated level of uncertainty would be irrelevant. In the proposed form of the model for diesel emissions, it is a simple matter to determine the total change achievable over the range of an eigenfuel. Thus, we would establish at the outset of an investigation the smallest effect that is meaningful to our purposes and drop any that are found to be smaller than this threshold.

Representation of Non-linear Effects

Work done to date has been based on the assumption that all vector predictor variables exert a linear effect on the response variable. Experience tells us that the effect of a predictor variable may be linear over much of its range but may show curvature for extreme values (e.g., saturation effects). In other instances, theory may suggest that the effect of the predictor may be non-linear over its entire range. The methodology for a vector approach to regression cannot be considered complete unless it can accommodate non-linear effects.

We consider it straightforward to add to the model whatever non-linear effects may be required simply by incorporating basis vectors exhibiting the desired non-linear characteristics. For example, the linear terms x_i constituting the original variables may be augmented by squared terms x_i^2 , interactive terms $x_i x_j$, or more generalized non-linear functional forms $f(x_i)$ in forming the X matrix. Moreover, there appears to be no difficulty in retaining the orthogonality of all vectors in the model. The theory of orthogonal polynomials in a finite domain and their application in regression models is well known and well documented in statistical literature [47].

Generalization and Robustness

Once an orthogonal basis of eigenfuels has been developed, how "robust" is that basis when applied to a

different set of data? The question is not appreciably different from that faced by a conventional regression model when it is applied in a context different from that in which it was developed. Validation sampling, in which the data set is divided into two or more parts and results compared, is an acknowledged approach to this problem [48]. We consider bootstrap sampling [49], in which samples are drawn with replacement from the data set, to be an extension of the validation idea, and one that does not suffer from reduction in sample size.

To some, however, validation sampling may not seem like a test at all, because all samples are drawn from the same source. More appropriate would be a procedure in which samples are drawn from different but similar sources with a view to finding the common aspects of response. This approach sometimes goes under the name “system identification” and has been explored by one of the authors in an earlier paper and in a different context [50]. We expect to build on this experience to whatever extent seems appropriate.

STEPS TO DEVELOP A DIESEL EMISSION MODEL

An improved database is a prime requirement for the future development of a reliable diesel emission model. These are the most important limitations of the database used in this study and our recommendations for future testing:

- The database omits at least one fuel property – oxygen content – that is likely to affect emissions. Few test programs to date have evaluated oxygenated fuels, and more testing of oxygenated fuels will be required before a complete diesel emissions model can be developed.
- The database lacks information on hydrocarbon composition beyond mono- and poly-aromatic content, as do most existing testing programs. It may be important for new testing to report a more complete or detailed hydrocarbon speciation because, when a fuel is changed by the substitution of one constituent for another, it is not possible to attribute an emissions change uniquely to the one constituent (or the other). While such an effort could open a Pandora’s box if carried too far, it could very well be important to know whether it was hydrocarbon species 1 or 2 that was substituted when, for example, aromatics content was reduced.
- The database represents too few engine types and individual engines to be taken as representative of the on-road HDD fleet or to permit the assessment of engine-related effects. Although the total of 280 emissions tests is relatively large, the data are based on only 11 individual engines. An improved database should represent a substantially larger number of engines sampled in a representative manner from the cells created by the intersections of model year, emission certification standard, manufacturer, and engine design. It would be desirable that two or more specimens be included

for each major engine design to permit the estimation of design-specific effects, and that all testing use the one test cycle on which certification decisions for fuels and engines will be based.

The properties of test fuels used in future testing should be varied over the widest practical range and should include any fuel property indicated to affect emissions in a substantive way. Also important is that the fuel properties should be varied in accordance with their *cooperative* effects, as indicated in the eigenfuels. Just how the test fuels should be blended is a topic for discussion.

Conventionally, one method might be to vary a given fuel property over its desired range without regard to other fuel properties. Provided the test fuels were blended from commercially feasible blend stocks, the resulting test fuels and eigenfuels derived from them might be taken as representative of future commercial fuels blended to similar specifications. Other possibilities could include attempts to blend fuels that meet multiple property specifications simultaneously. This latter approach has been the one frequently followed in past testing where, for example, mono- and poly-aromatics content might be varied subject to controlling fuel density and viscosity to predetermined values.

It is here, though, that blending in terms of eigenfuels shows its major advantage. It is necessary only to resolve a desired eigenfuel into a corresponding set of available blend stocks. To match an eigenfuel exactly would require up to 12 blend stocks. Note that in eigenfuel 2, which accounts for 82 percent of the NO_x regression SS, only 4 of its components – natural cetane, density, mono-aromatic and poly-aromatic content – make major contributions. No more than four blend stocks should be sufficient to approximate eigenfuel 2. Though the mechanics of the solution mimic the conventional, it employs a vastly more effective strategy for formulating optimum test fuels.

In varying engine characteristics, the primary requirement is to have a sampling of vehicles that covers all of the design factors that might influence emissions. It may turn out that some of these design factors have relatively little effect on emissions. Here, again, a fixed magnitude of the engine effect should be used as the criterion for including or excluding that factor, rather than an arbitrary test of significance. Since the engine effects are most likely not orthogonal, it might be appropriate to induce orthogonality by a method such as random balance assessment [51].

Engine effects and fuel effects may interact. If so, the computation of “fuel effects adjusted for engine effects” is inadequate, because the adjustment corrects only for the difference in the mean level of emissions among engines, it being assumed that the incremental effect of a fuel change is constant for all engine classes. It is quite possible, however, that the effect of an incremental change in a fuel property is greater for one class of vehicle than another, especially if the fuel change is designed to play an “enabling” role for a vehicle design change. Resolving

such issues would require more extensive testing and a model of greater complexity.

CONCLUSION

In summary, a vector methodology has been demonstrated for regressing a response variable on orthogonal functions developed from predictor variables in circumstances where the predictor variables cannot be varied independently of each other. Though demonstrated on a model for heavy duty diesel emissions, the approach is applicable to a wide variety of problems.

When applied to the fuel/emissions relationship, the preliminary findings illustrate the range of benefits that the vector approach offers:

- Simplification of the regression analysis as a result of the desirable mathematical properties of vector variables – their independence and absence of correlations, and their economy of representation.
- Greater understanding of the patterns of variation that are important to emissions reduction, in this instance, and how these patterns relate to fuel blending and refinery processes.
- Potentially new insight into the optimal formulation of fuels to reduce emissions.
- Improved experiment design for more efficient estimation of fuel effects.

Perceived disadvantages of the methodology are:

- The ineffectiveness of selecting predictor variables by means of PCA as noted in the cautionary notes by Hadi and Ling [8].
- Potential difficulties in interpreting the effects of the predictor vectors on the dependent variable.

We have shown that variable selection can not, and should not, be attempted without regard to the response variable. This fact is made clear by the difference between NO_x and PM response, even though the PCA analysis of the design space is the same.

With regard to interpretation, it may appear that the multicollinearity “fog” disposed of by orthogonalization is merely replaced by a different kind of fog arising from the difficulties of interpreting eigenvectors. We believe that viewing response in terms of eigenvectors is not so much *difficult* as it is *unconventional* and that it affords an improved basis for understanding the factors that actually drive the response. Finally, we offer a means for alternatively viewing response in terms of the original variables, but within the context of an orthogonal, eigenvector solution.

It is, perhaps, most important that the vector approach to analysis does not require or benefit from the attempt to “break” naturally-occurring associations among fuel properties in the blending of test fuels. Without the need to artificially separate these associations, a wider range of

real-world diesel fuels, representing current and future refinery configurations and processes, could be used in engine testing, thereby avoiding possibly unrealistic or unrepresentative emission results. These benefits imply an increased accuracy in assessing emissions benefits and an improved basis for measuring cost effectiveness.

ACKNOWLEDGMENTS

This research, performed under contract with the Energy 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. 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 conceptual guidance and helpful comments provided by Barry McNutt of the U.S. Department of Energy Office of Policy. The authors also wish to acknowledge the assistance of K. G. Duleep of Energy and Environmental Analysis, Inc. (EEA) in providing portions of the data used in this study.

REFERENCES

1. Lee, R., J. Pedley and C. Hobbs, “Fuel Quality Impact on Heavy Duty Diesel Emissions - A Literature Review,” SAE 982649.
2. U.S. Department of Energy. *Estimating the Costs and Effects of Reformulated Gasoline*. DOE/PO-0030. December 1994.
3. Krzanowski, W.J., and F.H.C. Marriott. *Multivariate Analysis*. Kendall’s Library of Statistics. Halstead Press. 1994.
4. Jackson, J.E. *A User’s Guide to Principal Components*. John Wiley & Sons. 1991.
5. Martens, H. and T. Naes. *Multivariate Calibration*. John Wiley & Sons. 1989.
6. Fisher, R. A., *The Design of Experiments*, Edinburgh: Oliver and Boyd, 1935.
7. Thurstone, L. L., *The Vectors of Mind*, Chicago: U. of Chicago Press, 1935.
8. Westerholm, R. and H. Li. *A Multivariate Statistical Analysis of Fuel-Related Polycyclic Aromatic Hydrocarbon Emissions from Heavy-Duty Diesel Vehicles*. Environ. Sci. Technol. 28, pp. 965-972. 1994.
9. Hawkins, D. *On the Investigation of Alternative Regressions by Principal Component Analysis*. Applied Statistics 22, pp. 275-286. 1973.

10. Boneh, S. and G.R. Mendieta. *Variable Selection in Regression Models Using Principal Components*. Communications in Statistics – Theory and Methods 23, pp. 197-213. 1994.
11. Mansfield, E.R., J.T. Webster and R.F. Gunst. *An Analytical Variable Selection Technique for Principal Component Regression*. Applied Statistics 26, pp. 34-40. 1977.
12. Jolliffe, I.T. *Discarding Variables in a Principal Component Analysis. I: Artificial Data*. Applied Statistics 21, pp. 160-173. 1972.
13. Jolliffe, I.T. *Discarding Variables in a Principal Component Analysis. II: Real Data*. Applied Statistics 22, pp. 21-31. 1973.
14. Hadi, A.S. and Ling, R.F. *Some Cautionary Notes on the Use of Principal Components Regression*. The American Statistician, Vol. 52, No. 1. February 1998.
15. Jeffers, J.N. *Two Case Studies in the Application of Principal Component Analysis*. Applied Statistics, Vol 16, pp. 225-236. 1967.
16. McAdams, H.T., R.W. Crawford and G.R. Hadder. *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. 2000.
17. *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, 1999.
18. Ullman, T.L., "Investigation of the Effects of Composition and Injection and Combustion System Type on Heavy-Duty Diesel Exhaust Emissions," SAE 892072.
19. Ullman, T.L., R.L. Mason and D.A. Montalvo, "Effects of Aromatics, Cetane Number, and Cetane Improver on Emissions from a 1991 Prototype Heavy Duty Diesel Engine," SAE 902171.
20. Sienicki, E.J., R.E. Jass and W.J. Slodowske, "Diesel Fuel Aromatic and Cetane Number Effects on Combustion and Emissions from a Prototype 1991 Diesel Engine," SAE 902172.
21. McCarthy, C.I., W.J. Slodowske, E.J. Sienicki and R.E. Jass, "Diesel Fuel Property Effects on Exhaust Emissions from a Heavy Duty Diesel Engine that Meets 1994 Emissions Requirements," SAE 922267.
22. Gonzalez D, M.A., G.R. Rodriguez B., R. Galiasso and E. Rodriguez, "A Low Emission Diesel Fuel: Hydrocracking Production, Characterization, and Engine Evaluations," SAE 932731.
23. Ullman, T.L., K.B. Spreen and R.L. Mason, "Effects of Cetane Number, Cetane Improver, Aromatics, and Oxygenates on 1994 Heavy Duty Engine Emissions," SAE 941020.
24. Spreen, K.B., T.L. Ullman and R.L. Mason, "Effects of Cetane Number, Aromatics and Oxygenates on Emissions from a 1994 Heavy Duty Diesel Engine with Exhaust Catalyst," SAE 950250.
25. Ullman, T.L., K.B. Spreen and R.L. Mason, "Effects of Cetane Number on Emissions from a Prototype 1998 Heavy Duty Diesel Engine," SAE 950251.
26. Schaberg, P.W., I.S. Myburgh, J.J. Botha, P.N. Roets, C.L. Viljoen, L.P. Dancuart and M.E. Starr, "Diesel Exhaust Emissions Using Sasol Slurry Phase Distillate Process Fuel," SAE 972898.
27. Cunningham, L.J., T.J. Henly and A.M. Kulinowski, "The Effects of Diesel Ignition Improvers in Low-Sulfur Fuels on Heavy-Duty Diesel Emissions," SAE 902173.
28. Lange, W.W., "The Effect of Fuel Properties on Particulates Emissions in Heavy-Duty Truck Engines Under Transient Operating Conditions," SAE 912425.
29. Liotta Jr., F.J. and D.M. Montalvo, "The Effect of Oxygenated Fuels on Emissions from a Modern Heavy Duty Diesel Engine," SAE 932734.
30. Liotta Jr., F.J., "A Peroxide Based Cetane Improvement Additive with Favorable Fuel Blending Properties," SAE 932767.
31. Rosenthal, M.L. and T. Bendinsky, "The Effects of Fuel Properties and Chemistry on Emissions and Heat Release of Low Emission Heavy Duty Diesel Engines," SAE 932800.
32. Reynolds, E.G., "The Effect of Fuel Processes on Heavy Duty Automotive Diesel Engine Emissions," SAE 932350.
33. Schmidt, K. and J. Van Gerpen, "The Effect of Biodiesel Fuel Composition on Diesel Combustion and Emissions," SAE 961086.
34. Geiman R.A., P.B. Cullen, P.R. Chant, P.N. Carlson and V. Rao, "Emission Effects of Shell Low NO_x Fuel on a 1990 Model Year Heavy Duty Diesel Engine," SAE 961973.

35. Tanaka, S., M. Morinaga, H. Yoshida, H. Takizawa, K. Sanse and H. Ikebe, "Effects of Fuel Properties on Exhaust Emissions from DI Diesel Engines," SAE 962114.
36. Daniels, T.L., R.L. McCormick, M.S. Graboski, P.N. Carlson, V. Rao and G.W. Rice, "The Effect of Diesel Sulfur Content and Oxidation Catalysts on Transient Emissions at High Altitude from a 1995 Detroit Diesel Series 50 Urban Bus Engine," SAE 961974.
37. Tamanouchi M., H. Morihisa, S. Yamada, J. Iida, T. Sasaki and H. Sue, "Effects of Fuel Properties on Exhaust Emissions for Diesel Engines with and without Oxidation Catalyst and High Pressure Injection," SAE 970758.
38. Lange, W.W., J.A. Cooke, P. Gadd, H.J. Zurner, H. Schogel and K. Richter, "Influence of Fuel Properties on Exhaust Emissions from Advanced Heavy-duty Engines Considering the Effect of Natural and Additive Enhanced Cetane Number," SAE 972894.
39. Starr, M.E., "Influence on Transient Emissions at Various Injection Timings, using Cetane Improvers, Bio-diesel and Low Aromatics Fuels," SAE 972904.
40. Nylund, N-O, P. Aakko, S. Mikkonen and A. Niemi, "Effects of Physical and Chemical Properties of Diesel Fuel on NO_x Emissions of Heavy Duty Diesel Engines," SAE 972997.
41. Akasaka, Y., T. Suzuki and Y. Sakurai, "Exhaust Emissions of a DI Diesel Engine Fueled with Blends of Biodiesel and Low Sulfur Fuel," SAE 972998.
42. Mann, N., F. Kvinge and G. Wilson, "Diesel Fuel Effects on Emissions: Towards a Better Understanding," SAE 982486.
43. Nakakita, K., S. Takasu, H. Ban, T. Ogawa, H. Naruse, Y. Tsukasaki and L. Yeh, "Effect of Hydrocarbon Molecular Structure on Diesel Exhaust Emissions Part 1: Comparison of Combustion and Exhaust Emission Characteristics Among Representative Diesel Fuels," SAE 982494.
44. *MatLab*, The MathWorks, Inc., 24 Prime Park Way, Natick, MA 01760-1500. Website <http://www.mathworks.com>.
45. Thomas, *J. Chem. Soc.*, Part II, pp. 573-579, 1946.
46. Ethyl Corporation. HITEC Performance Chemicals. *A Distillate Fuel Response to "Diesel Ignition Improver."*
47. Fisher, R. A. and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 3rd Edition, Edinburgh, Oliver and Boyd, Ltd., 1948.
48. Wold, S. *Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Models*. *Technometrics* 20, No. 4. November 1978.
49. Efron, B. and R.J. Tibshirani. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall/CRC. 1998.
50. McAdams, H.T., *A Factor Analytic Approach to the Identification of Manufacturing Systems*. Proceedings of CIRP Seminars on Manufacturing Systems, Vol I, No. 2. pp. 79-97. 1972.
51. McAdams, H.T., *A Random Balance Procedure for Simplifying a Complex Model*. American Statistical Association, 1995 Proceedings of the Section on Statistics and the Environment, Alexandria, VA, 1995.

CONTACT

Inquiries regarding this paper may be addressed to
Robert W. Crawford at
R_Crawford@compuserve.com.