SwRI 08.04075

DIESEL FUEL IMPACT MODEL DATA ANALYSIS PLAN REVIEW

Prepared by

Robert L. Mason, Ph.D.
Janet P. Buckingham

DRAFT FINAL REPORT

Prepared for

Environmental Protection Agency
2000 Traverwood
Ann Arbor, MI 48105

July 2001

SOUTHWEST RESEARCH INSTITUTE
6220 Culebra Road P.O. Drawer 28510
San Antonio, Texas 78228-0510

DIESEL FUEL IMPACT MODEL DATA ANALYSIS PLAN REVIEW

Prepared by

Robert L. Mason, Ph.D.
Janet P. Buckingham

DRAFT FINAL REPORT
Work Assignment No. 2-7
EPA Contract 68-C-98-169

Prepared for

Environmental Protection Agency
2000 Traverwood
Ann Arbor, MI 48105

July 2001

Reviewed by:

Robert L. Mason

Robert L. Mason, Staff Analyst

Lawrence R. Smith

Lawrence R, Smith, Manager

Approved:

Charles T. Hare

Charles T. Hare, Director

DEPARTMENT OF EMISSIONS RESEARCH
AUTOMOTIVE PRODUCTS AND EMISSIONS RESEARCH DIVISION

TABLE OF CONTENTS

Page

LIST OF TABLES iii

EXECUTIVE SUMMARY viii

I. INTRODUCTION 1

II. CREATION OF DATABASE 2

A. Studies Included in Database 2
B. FBAT_AD Entity Fuel Property Conversions 7
C. EQUIP_AD Entity Engine Parameter Conversions 8
D. ETEST_AD Entity Emissions Conversions 9

III. Mixed Models for Individual Engine Tech Groups 11

A. Data Selection 11
B. Data Screening 19
C. Standardization of Fuel Properties 23
D. Modeling Issues and Assumption Checks 23
E. Collinearity Checks 28
F. Additional Fuel Terms 30
G. Stepwise Mixed Model Fits 31

IV. EIGENVECTOR MODELS FOR SEPARATE ENGINE TECH GROUPS 35

A. Analysis Steps 35
B. Models for Selected Engine Tech Groups 36

V. MIXED MODELS BASED ON COMBINED ENGINE TECH GROUPS 40

A. LOG(NOx) Analyses 40
B. LOG(PM) Analyses 47
C. LOG(HC) Analyses 54
D. Residual Analyses 59

VI. EIGENVECTOR MODELS BASED ON COMBINED ENGINE TECH GROUPS 60

A. Eigenvector Models 60
B. Methodology Issues 62

TABLE OF CONTENTS (CONT'D)

Page

VII. MODEL PERFORMANCE 65

A. Methodology 65
B. Comparison of Percent Change in Emissions 67

APPENDICES **No of Pages**

A - DESCRIPTION OF WORK STATEMENT 2-7 7
B - DATA ANALYSIS PLAN 3
C - ASSESSMENT OF THE DATA ANALYSIS DATA 6
D - ENTITY NAME/TABLE NAME DEFINITIONS 3
E - REVIEW OF VECTOR APPROACH TO REGRESSION ANALYSIS . . . 4
F - MIXED MODEL METHODOLOGY 2
G - TIME-DRIFT CORRECTION EQUATIONS 2
H - RESIDUAL PLOTS 14
I - SCATTER PLOTS OF PERCENT CHANGE 2

LIST OF TABLES

Table		Page
1	Studies Used in Compiling Diesel Emissions Database	3
2	Test Procedures Used in Emissions Analyses	14
3	Tech Group Definitions by Engine Classifications	15
4	Tech Group Classification by Engine ID and Study ID	16
5	Analysis Identification for Tech Group by Emissions	18
6	Fuel Variables Used in NO_x Analysis	20
7	Fuel Variables Used in PM Analysis	20
8	Fuel Variables Used in HC Analysis	21
9	Test Procedure by Tech Group Frequencies for NO_x	21
10	Test Procedure by Tech Group Frequencies for PM	21
11	Test Procedure by Tech Group Frequencies for HC	22
12	Fuel Property Means and Standard Deviations for NO_x Analysis	24
13	Fuel Property Means and Standard Deviations for PM Analysis	25
14	Fuel Property Means and Standard Deviations for HC Analysis	26
15	Comparison of Fits to Log (NOx) for Tech Groups With and Without Random Engine-By-Fuel Interaction Terms	27
16	Condition Numbers by Tech Groups and Emissions	29
17	Coefficients of Standardized Fuel Properties for Log (NO_x) Using Additional Terms for Natural Cetane for Tech Group T	30
18	Coefficients of Standardized Fuel Properties for Log (NO_x) Using Additional Terms for Natural Cetane for Tech Group F-DD	31
19	Estimated Coefficients for Standardized Fuel Terms in "Best" Step of Stepwise Fit to Log (NO_x)	33

LIST OF TABLES (CONT'D)

Table		Page
20	Estimated Coefficients for Standardized Fuel Terms in "Best" Step of Stepwise Fit to Log (PM)	33
21	Estimated Coefficients for Standardized Fuel Terms in "Best" Step of Stepwise Fit to Log (HC)	34
22	Coefficients of Standardized Fuel Properties for Log (NO_x) After Eigenvector Analysis for Tech Group T	37
23	Coefficients of Standardized Fuel Properties for Log (NO_x) After Eigenvector Analysis for Tech Group Q-OO	37
24	Coefficients of Standardized Fuel Properties for Log (NO_x) After Eigenvector Analysis for Tech Group B	37
25	Coefficients of Standardized Fuel Properties for Log (NO_x) After Eigenvector Analysis for Tech Group H	38
26	Coefficients of Standardized Fuel Properties for Log (PM) After Eigenvector Analysis for Tech Group F	39
27	Coefficients of Standardized Fuel Properties for Log (PM) After Eigenvector Analysis for Tech Group OO	39
28	Coefficients of Standardized Fuel Properties for Log (PM) After Eigenvector Analysis for Tech Group B	39
29	Average-Repeat Data Available for Log (NO_x) After Deletion of Outliers	41
30	Fuel Property Means and Standard Deviations for Log (NO_x) Analysis Using Average-Repeat Data	41
31	Fuel Property Means and Standard Deviations for Log (NO_x) Analysis Using Combined Data	43
32	Estimated Coefficients of Standardized Variables for Log (NO_x) from Mixed-Effects Model No. 1 Analysis Based on EPA Stepwise Approach	44
33	Estimated Coefficients of Standardized Variables for Log (NO_x) from Mixed-Effects Model No. 2 Analysis Based on EPA Stepwise Approach	45

LIST OF TABLES (CONT'D)

Table		Page
34	Estimated Coefficients of Standardized Variables for Log (NO_x) from Mixed-Effects Model No. 3 Analysis Based on EPA Stepwise Approach	46
35	Coefficients of Standardized Variables for Log (NOx) from Mixed-Effects Model No. 4 Analysis Based on EPA Stepwise Approach	47
36	Average-Repeat Data Available for Log (PM) After Deletion of Outliers	48
37	Fuel Property Means and Standard Deviations for Log (PM) Analysis Using Average-Repeat Data	48
38	Fuel Property Means and Standard Deviations for Log (PM) Analysis Using Combined Data	50
39	Estimated Coefficients of Standardized Variables for Log (PM) from Mixed-Effects Model No. 1 Analysis Based on EPA Stepwise Approach	51
40	Estimated Coefficients of Standardized Variables for Log (PM) from Mixed-Effects Model No. 2 Analysis Based on EPA Stepwise Approach	52
41	Estimated Coefficients of Standardized Variables for Log (PM) from Mixed-Effects Model No. 3 Analysis Based on EPA Stepwise Approach	53
42	Coefficients of Standardized Variables for Log (PM) from Mixed-Effects Model No. 4 Analysis Based on EPA Stepwise Approach	54
43	Average-Repeat Data Available for Log (HC) After Deletion of Outliers	55
44	Fuel Property Means and Standard Deviations for Log (HC) Analysis Using Average-Repeat Data	55
45	Fuel Property Means and Standard Deviations for Log (HC) Analysis Using Combined Data	57
46	Estimated Coefficients of Standardized Variables for Log (HC) from Mixed-Effects Model No. 1 Analysis Based on EPA Stepwise Approach	58
47	Estimated Coefficients of Standardized Variables for Log (HC) from Mixed-Effects Model No. 2 Analysis Based on EPA Stepwise Approach	58

LIST OF TABLES (CONT'D)

Table		Page
48	Coefficients of Standardized Variables for Log (HC) from Mixed-Effects Model No. 3 Analysis Based on EPA Stepwise Approach	59
49	Coefficients of Standardized Variables for LOG(NOx) After Eigenvector Analysis	62
50	Coefficients of Standardized Variables for LOG(PM) After Eigenvector Analysis	63
51	Coefficients of Standardized Variables for LOG(HC) After Eigenvector Analysis	64
52	Comparison of Observed and Predicted % CE for NO_x	67
53	Frequency Distribution of the Absolute Difference in % CE for NO_x	68
54	Comparison of Observed and Predicted % CE for PM	68
55	Frequency Distribution of the Absolute Difference in % CE for PM	69
56	Comparison of Observed and Predicted % CE for HC (Without Restrictions)	70
57	Comparison of Observed and Predicted % CE for HC (With Restrictions)	70
58	Frequency Distribution of the Absolute Difference in % CE for HC (Without Restrictions)	71
59	Frequency Distribution of the Absolute Difference in % CE for HC (With Restrictions)	71

EXECUTIVE SUMMARY

This work was motivated by the interests of EPA, states, and other stakeholders in quantifying the effects of selected diesel fuel properties in reducing the emissions emitted by heavy-duty compression-ignition engines. Three separate tasks were conducted.

Task 1 included assessing the adequacy of the proposed data analysis plan, suggesting improvements to it, and providing a review of a methodology based on an eigenvector approach to regression analysis. In the work performed under this task, the data analysis plan proposed by EPA was modified with respect to the choice of the fuel and engine properties to be considered, and with respect to the modeling procedures to be selected. Also, it was determined that the eigenvector approach was not sufficiently developed for usage as the primary modeling procedure.

Task 2 involved creating the database to be used in the study. The source of the data were 39 studies provided by the EPA in the form of SAE papers, study reports, or CD-ROMs with data. The data were entered into spreadsheets for the fuel, engine, and emissions data. Extensive effort was taken to insure that the data was translated correctly into the computer files, and that the units of measurement were the same across the studies. The final data set included 1777 observations on 73 different engines and 300 different fuels.

Task 3 involved using the data in the database to construct models appropriate for assessing the impact of a given fuel on diesel engine emissions changes. The three emissions modeled in this study included NO_x, PM, and HC. The fuel properties evaluated included natural cetane (NATCET), cetane difference (CETDIFF) due to the inclusion of an additive, total aromatics (TAROM), specific gravity (SPGRAV), sulfur(SULFUR), oxygen (OXY), and the 10, 50 , and 90 percent distillation values (T10, T50, and T90). The engine data were categorized into 16 technology groups, and 7 different test procedures were chosen to be used for analysis of the emissions data.

Two different approaches were utilized. One included separately fitting the data from each engine technology group, where sufficient data was available. The other included combining the technology groups and creating a combined set of data for analysis. In both approaches, the models were constructed using regression techniques, mixed-model procedures, and eigenvector analysis methods. EPA chose to estimate model performance (in terms of predicting the percent change in emissions from a baseline fuel) with the mixed models based on the combined data set.

The selected NO_x prediction model included CETDIFF, TAROM, SPGRAV, T50, and various tech-group-by-fuel interaction terms (with the interaction terms involving SULFUR and NATCET, CETDIFF, and T50). The chosen PM prediction model included NATCET, CETDIFF, TAROM, SULFUR, SPGRAV, OXY, and NATCET*CETDIFF, as well as some tech-group-by-fuel interaction terms. The HC prediction model included NATCET, CETDIFF, T10, and T50.

I. INTRODUCTION

Under Work Assignment 2-7 of EPA Contract 86-C-98-169, the U.S. Environmental Protection Agency (EPA) directed Southwest Research Institute (SwRI) to study the effects of diesel fuel properties on heavy-duty compression-ignition engine emissions (see Appendix A for a complete description). The effort was divided into three separate tasks. Task 1 included assessing the data analysis plan, Task 2 involved creating a database, and Task 3 included generating statistical models that would be appropriate in assessing the impact of a given fuel on diesel engine emissions. The time period of the program extended from February 2001 until August 2001. This report describes the obtained results.

The motivation for this work stems from the interests on the part of states and stakeholders in quantifying the effects of various diesel fuel properties in reducing engine emissions. In particular, several estimates of the emissions benefits obtained by controlling such diesel fuel properties as cetane and aromatics have been presented by these various groups. EPA, in turn, is concerned with the accuracy, magnitude, and consistency of these projections. Thus, EPA proposed this program whereby all pertinent data are collected into one database, and different modeling strategies are used to provide an assessment of the impact of a variety of fuel properties on emissions.

The objective of Task 1 was to assess the scientific and statistical validity and robustness of the EPA's sampling strategy and data analysis plan (see Appendix B for a complete description of the Data Analysis Plan). It included three subtasks:

- Assessing the adequacy of the data analysis plan
- Suggesting improvements to the plan
- Providing a concise review and assessment of the methodology based on a vector approach to regression analysis

Because the main emphasis in this final report is on the creation of the database and the building of the prediction models, the results of the Task 1 effort have been placed in the appendices (See Appendices C, D, and E for a complete description).

II. CREATION OF DATABASE

This section describes the creation of the diesel fuel database. The format and the structure of the database were established by EPA. Additional variables were included in the database structure as described in Appendix C. Table C-1 of the Appendix lists the database entity definitions as provided by EPA.

A. Studies Included in Database

EPA identified 39 studies to be used in compiling the diesel emissions database. The studies were provided to SwRI in several different formats, including in the form of an SAE paper, a study report, or a CD-ROM. In some cases, SwRI already had copies of the study report, particularly for those studies performed at SwRI. Each study was reviewed by a group of engineers and analysts at SwRI in order to extract the information to be entered into the fuel, engine, and emissions database files. A list of the 39 studies is provided in Table 1. This table includes the study title, SAE paper number (where applicable), authors, number of valid observations, number of engines, and number of fuels.

Note that four studies listed in Table 1(and labeled No. 3, 7, 14, and 33) were eliminated from the database because they did not meet the criteria established by EPA. These studies were:

- SAE1999-01-1508 – fuel property data only available on one fuel
- SAE972903 – no fuel property data available
- SAE961166 – only one fuel available, and no fuel property data available on biodiesel blends
- HDEWG PHASE III – raw emissions data were not available

The final data set included 1777 observations on 73 different engines and 300 different fuels. The study data were entered into three separate Excel spreadsheets: one each for the fuels, engines, and emissions. Coding information, contained in translation tables provided by EPA, were used to enter categorical variables. All values were entered into the database in the units specified for each field as provided by EPA. EPA entered the data from the VE-10 and HDEWG Phase II studies along with the engine data from the VE-1 Phase I study. SwRI entered all the data from the remaining studies. In order to merge the three Excel files into a single one, SwRI first combined the engine and emissions files by the STUDY_ID and ENG_MS_ID keywords. The resulting file was merged with the fuel file by using the STUDY_ID and FBATCH_ID keywords. No individual modal data were entered in the EMODE_AD database due to time constraints.

Several decisions were made during the review of the studies and the subsequent entry into the database. EPA was consulted at each of these decision points for directions. The guidelines used to enter the data into the database files are listed in the following sections.

TABLE 1. STUDIES USED IN COMPILING DIESEL EMISSIONS DATABASE

Paper No.	Description	Title	Authors	No. of Observations	No. of Engines	No. of Fuels
1	SAE 2000-01-2890	Effects of Fuel Properties and Source on Emissions from Five Different Heavy Duty Diesel Engines	Ken Mitchell	87	4	10
2	SAE 1999-01-3606	Emissions Performance of Oxygenate-in-Diesel Blends and Fischer-Tropsch Diesel in a Compression Ignition Engine	Adelbert S. Cheng, Robert W. Dibble	6	1	2
3	SAE 1999-01-1508	Methylal and Methylal-Diesel Blended Fuels for Use in Compression-Ignition Engines	Keith D. Vertin, James M. Ohi, David W. Naegeli, Kenneth H, Childress, et al	Deleted study – fuel property data available on only one fuel		
4	SAE 1999-01-1478	The Effects of 2-Ethylhexyl Nitrate and Di-Tertiary-Butyl Peroxide on the Exhaust Emissions from a Heavy-Duty Diesel Engine	Scott D. Schwab, Gregory H. Guinther, Timothy J. Henly, Keith T. Miller	126	1	22
5	SAE 1999-01-1117	Transient Emissions Comparisons of Alternative Compression Ignition Fuels	Nigel N. Clark, Christopher M. Atkinson, Gregory J. Thompson, Ralph D. Nine	24	1	7
6	SAE 972904	Influence on Transient Emissions at Various Timings, Using Cetane Improvers, Bio-Diesel, and Low Aromatic Fuels	Michael E. Starr	54	3 (1 engine configured 3 ways)	6
7	SAE 972903	Reduction in Particulate and Black Smoke in Diesel Exhaust Emissions	R.F. Becker, P. Ndiomu, D.H. Hoskin	Deleted study – no fuel property data available		
8	SAE 972898	Diesel Exhaust Emissions Using Sasol Slurry Phase Distillate Process Fuels	Paul W. Schaberg, Ian S. Myburgh, Jacobus J. Botha, Piet N. Roets, Carl L.Vijoen	25	1	7
9	SAE 972894	Influence of Fuel Properties on Exhaust Emissions from Advanced Heavy-Duty Engines Considering the Effect of Natural and Additive Enhanced Cetane Number	W.W. Lange, J.A. Cooke, P. Gadd, H.J. Zurner, H. Schlogl, K. Richter	16	1	5
10	SAE 971635	The Influence of Fuel Properties and Injection Timing on the Exhaust Emissions and Fuel Consumption of an Iveco Heavy-Duty Diesel Engine	Richard Stradling, Paul Gadd, Meinrad Signer, Claudio Operti	15	3 (1 engine configured 3 ways)	9

TABLE 1 (CONT'D). STUDIES USED IN COMPILING DIESEL EMISSIONS DATABASE

Paper No.	Description	Title	Authors	No. of Observations	No. of Engines	No. of Fuels
11	SAE 970758	Effects of Fuel Properties on Exhaust Emissions for Diesel Engines With and Without Oxidation Catalyst and High Pressure Injection	Mitsuo Tamanouchi, Hiroki Morihisa, Shigehisa Yamada, et al	48	4	10
12	SAE 961974	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	Teresa L. Daniels, Robert L. McCormick, Michael S. Graboski, Philip N. Carlson, Venkatesh Rao, Gary W. Rice	42	3 (1 engine configured 3 ways)	6
13	SAE 961973	Emission Effects of Shell LOW NO_X Fuel on a 1990 Model Year Heavy-Duty Diesel Engine	Richard A. Geiman, Patrick B. Cullen, Peter R. Chant, et al	31	1	2
14	SAE 961166	Transient Emissions from a No. 2 Diesel and Biodiesel Blends in a DDC Series 60 Engine	M.S. Graboski, J.D. Ross, R.L. McCormick	Deleted study – only one fuel available; no fuel property data available on biodiesel blends		
15	SAE 942053	Impact of Diesel Fuel Aromatics on Particulate, PAH and Nitro-PAH Emissions	K. Mitchell, D.E. Steere, J.A. Taylor, B. Manicom, et al	72	3 (1 engine configured 2 ways)	4
16	SAE 942019	The Performance of a Peroxide-Based Cetane Improvement Additive in Different Diesel Fuels	Manish K. Nandi, David C. Jacobs, Frank J. Liotta, Jr., H.S. Kesling, Jr.	48	1	12
17	SAE 932800	The Effects of Fuel Properties and Chemistry on the Emissions and Heat Release of Low-Emission Heavy Duty Diesel Engines	M. Lori Rosenthal, Tracy Bendinsky	20	1	5
18	SAE 932767	A Peroxide Based Cetane Improvement Additive with Favorable Fuel Blending Properties	Frank J. Liotta, Jr.	10	1	3
19	SAE 932734	The Effect of Oxygenated Fuels on Emissions from a Modern Heavy-Duty Diesel Engine	Frank J. Liotta, Jr., Daniel M. Montalvo	104	1	14

TABLE 1 (CONT'D). STUDIES USED IN COMPILING DIESEL EMISSIONS DATABASE

Paper No.	Description	Title	Authors	No. of Observations	No. of Engines	No. of Fuels
20	SAE 932731	A Low Emission Diesel Fuel: Hydrocracking Production, Characterization and Engine Evaluations	Manuel A. Gonzalez D., Guillermo Rodriguez B., Roberto Galiasso, Edilberto Rodriguez	6	1	2
21	SAE 932685	The Influence of Fuel Properties on Exhaust Emissions from Advanced Mercedes Benz Diesel Engines	W.W. Lange, A. Schafer, A. Le'Jeune, D. Naber, et al	44	1	12
22	SAE 922267	Diesel Fuel Property Effects on Exhaust Emissions From a Heavy Duty Diesel Engine that Meets 1994 Emissions Requirements	Christopher I. McCarthy, Warren J. Slodowske, Edward J. Sienicki, Richard E. Jass	61	1	12
23	SAE 912425	The Effect of Fuel Properties on Particulate Emissions in Heavy-Duty Truck Engines Under Transient Operating Conditions	W.W. Lange	93	1	7
24	SAE 910735	Fuel and Maladjustment Effects on Emissions from a Diesel Bus Engine	Terry L. Ullman, David M. Human	26	3 (1 engine configured 3 ways)	5
25	SAE 902173	The Effects of Diesel Ignition Improvers In Low-Sulfur Fuels on Heavy-Duty Emissions	Lawrence J. Cunningham, Timothy J. Henly, Alexander M. Kulinowski	61	1	18
26	SAE 902172	Diesel Fuel Aromatic and Cetane Number Effects on Combustion and Emissions From a Prototype 1991 Diesel Engine	Edward J. Sienicki, Richard E. Jass, Warren J. Slodowske, Christopher I. McCarthy, Allen L. Krodel	26	1	11
27	SAE 881173	"Future" Diesel Fuel Compositions – Their Influence on Particulates	Hans Walter Knuth, Hellmut Garthe	6	1	3
28	SAE 852078	Heavy-Duty Diesel Engine/Fuels Combustion Performance and Emissions – A Cooperative Research Program	E.G. Barry, L.J. McCabe, D.H. Gerke, J.M. Perez	12	1	6
29	VE-1, PHASE I CAPE32-80	Investigation of the Effects of Fuel Composition and Injection and Combustion System Type on Heavy-Duty Diesel Exhaust Emissions	Terry L. Ullman	82	3	10

TABLE 1 (CONT'D). STUDIES USED IN COMPILING DIESEL EMISSIONS DATABASE

Paper No.	Description	Title	Authors	No. of Observations	No. of Engines	No. of Fuels
30	VE-1, PHASE II	Study of Fuel Cetane Number and Aromatic Content Effects on Regulated Emissions From a Heavy-Duty Diesel Engine	Terry L. Ullman, Robert L. Mason, Daniel A. Montalvo	41	1	13
31	CRC Contract No. VE-10	Effects of Fuel Oxygenates, Cetane Number, and Aromatic Content on Emissions From 1994 and 1998 Prototype Heavy-Duty Diesel Engines	Kent B. Spreen, Terry L. Ullman, Robert L. Mason	77	5	23
32	HDEWG PHASE II EPA68-C-98-169	Gaseous Emissions From a Caterpillar 3176 (with EGR) Using a Matrix of Diesel Fuels (Phase 2)	Andrew C. Matheaus, Thomas W. Ryan III, Robert Mason, Gary Neely, Rafal Sobotowski	81	4 (1 engine configured 4 ways)	19
33	HDEWG PHASE III		Glen Passavant	Study deleted - raw emissions data not available		
34	CARB TOXICITY	Evaluation of Factors That Affect Diesel Exhaust Toxicity	Timothy J. Truex, Joseph M. Norbeck, Matthew R. Smith	93	1	3
35	CARB LOCOMOTIVE	Diesel Fuel Effects on Locomotive Exhaust Emissions	Steven G. Fritz	12	1	3
36	SAE 961074 EPEFE STUDY	European Programme on Emissions, Fuels and Engine Technologies (EPEFE) – Heavy Duty Diesel Study	M. Signer, P. Heinze, R. Mercogliano, H.J. Stein	275	12 (4 engines configured 3 ways)	11
37	SAE 922214	Effects of Fuel Properties on Diesel Engine Exhaust Emission Characteristics	Yasuo Asaumi, Motohiro Shintani, Yoshito Watanabe	12	2	8
38	SAE 790490	Characterization of Heavy-Duty Diesel Gaseous and Particulate Emissions, and Effects of Fuel Composition	Charles T. Hare, Ronald L. Bradow	20	2	5
39	ACEA REPORT	Influence of Diesel Fuel Quality in Heavy Duty Diesel Engine Emissions	G. Kleinschek, K. Richter, A. Roj, M. Signer, H.J. Stein	21	1	5
			TOTAL	**1777**	**73**	**300**

B. FBAT_AD Entity Fuel Property Conversions

The units used for each fuel property were specified by EPA and are given in Appendix D. If the study provided fuel data in other units, they were converted and then entered into the database in the requested units. The following conversions were used prior to data entry. Some are standard unit conversions while others were provided by EPA.

- Specific gravity=141.5/(API gravity + 131.5)
- Specific gravity = 0.001 x density kg/m^3
- °F = (°C*9/5) + 32
- cetane improver: (vol %) = (ppmv/10,000)
- HCRATIO: (molecular) = 12.0 * (mass %)
- For the VE-10 study data, the cetane improver (vol %) conversions are as follows:
  For DTBP, vol% = wt% x (base fuel specific gravity)/0.794
  For EHN, vol% = wt% x (base fuel specific gravity)/0.964
- Use WSPA equations to convert total aromatics SFC wt% to FIA vol% as follows:
  FIA vol% = (0.916 x SFC wt%) + 1.33
- Aromatics wt% to vol% conversions:
  Total aromatics vol% = wt% x [specific gravity of fuel]/0.94
  Monoaromatics vol% = wt% x [specific gravity of fuel]/0.90
  Polyaromatics vol% = wt% x [specific gravity of fuel]/1.05
- Correlations for total aromatics:
  Vol% FIA = 0.738 x [vol% by HPLC] + 127.6 x [specific gravity] – 100.0
  Vol% FIA = 0.760 x [wt% by HPLC] + 178.0 x [specific gravity] – 144.4
- Correlations for monoaromatics:
  Wt% by SFC = 0.882 x [wt % by mass spec] + 2.37
  Wt% by SFC = 0.885 x [wt% by HPLC] + 0.88
- Correlations for polyaromatics:
  Wt% by SFC = 1.22 x [wt% by mass spec] + 0.33
  Wt% by SFC = 1.27 x [wt% by HPLC] + 0.69

Additional fuel property decisions included the following:

- If a fuel contained oxygenate, the corresponding data was included in the database only if the oxygenate had been blended at 20 vol% or less.
- Viscosity was entered into the FBAT_AD data file for tests at 40°C.
- If no oxygenate was added to the fuel, then OXY_TYP=NONE and OXYGEN=0. OXYGEN was blank only if oxygen was added, but the amount was not provided.
- Cetane index was substituted for cetane number for the following 2 studies:
  SAE790490 - 5 fuels
  SAE922214 - 8 fuels

- It was assumed that no inherent bias existed in one test method relative to another for measured cetane number. Thus, the test method was ignored and only the cetane number provided in the study was entered.
- For aromatics, if the test method was not given, it was assumed that the wt% values were based on SFC, and that the vol% values were based on FIA tests.
- All aromatics values were entered into the database in terms of vol% for total aromatics as measured with a FIA (D 1319) test method, or in terms of wt% for mono or poly aromatics as measured with an SFC (D5186) test method.
- If only one of the three aromatics values was missing, the missing value was estimated using the following relationship: total aromatics = monoaromatics + polyaromatics. The values were converted to the correct units before applying the relationship.
- For the VE-10 study, fuels AA-KK contained total aromatics estimated from only one lab Therefore, their corresponding total aromatics values were not entered. Instead, total aromatics was estimated by adding monoaromatics and polyaromatics.
- For cetane improver additives, the fuel properties of the base fuel (other than cetane) were considered to be equal to the fuel properties of the blend. For all other blends, no matter what the blending volumes, the fuel properties of the base fuel were not considered to be equal to the fuel properties of the blend.
- If there was no cetane improver additive, then CETANE_DIFF=0. If the study indicated that a cetane improver additive was added, but the study provided neither the vol% additive, nor the cetane levels with and without the additive, then CETANE_DIF was left blank.

C. EQUIP_AD Entity Engine Parameter Conversions

The units for each engine parameter were specified by EPA and are given in Appendix D. If the study provided engine property data in other units, they were converted and then entered into the database in the requested units. The following conversions were used prior to data entry:

- in = mm*0.03937
- ft-lb=0.7375*N-m
- hp=1.3405*kw

Additional engine parameters decisions included the following:

- Only engines that ran on more than one fuel in a study were included.
- The model year of the engine was to be its emissions representative model year, not necessarily the actual model year. For example, if a 1990 engine was calibrated to meet 1993 emissions standards, then the engine was identified as a 1993 model year engine.

- Model years were defined as representative years for the following:
  SAE790490 - both engines are pre-1984, entered 1984 in the database
  SAE972904 - engine calibrated for 1991-1993 emissions, entered 1993 in the database
- CARB-TOXIC - calibrated for 1991-93 model years, entered 1993 in the database
- Several studies did not provide complete engine parameter information. For those studies, engine experts at SwRI were consulted, and engine manufacturer handbooks were examined for the missing information. This provided a way to gather information in order to allow the engines to be assigned to tech groups.
- For the SAE971635 study, three engines were grouped according to the injection timing settings of 10, 9.2 and 8.7 BTDC. Injection timings at 10.7 and 9.7 were not included in the database because they were only run with one fuel.
- If a single engine was used in a study, but was modified (retarded timing for example) and then retested, it was given a new engine identification and considered as a different engine.

D. ETEST_AD Entity Emissions Conversions

The units for each emission parameter were specified by EPA and are given in Appendix D. If the study provided emissions in other units, they were converted and then entered into the database in the requested units. The following conversions were used prior to data entry:

- For BSFC: lb/bhp-hr = (0.001645) x g/kw-hr
- For Emissions: g/bhp-hr = (g/kw-hr)/1.3405

Additional emissions decisions included the following:

- For all transient test procedures, set NO_MODES = 0. Otherwise, enter the number of modes for the specific test procedure.
- To enter the appropriate data for the FTP composite tests (UDDS) the following guidelines were used:
  - If hot start and composite tests were available, only composite data were input. Hot start data were not entered.
  - If both individual hot start and cold start data were available, but no composite, a composite was computed as (6/7) × (Hot Start) + (1/7) × (Cold Start) for each individual pair of hot and cold start data.
  - If average hot and cold starts were given, a composite was computed as (6/7) × (Avg Hot Start) + (1/7) × (Avg Cold Start). This computed composite was entered into the database "X" times, where "X" equals the number of tests used to compute the average hot start.
  - If average hot and cold starts were given, a composite was computed as (6/7) × (Avg Hot Start) + (1/7) × (Avg Cold Start). If the number of hot

starts used to compute the average was not known, the computed composite was entered into the database two times.
- If only hot start data were available, the hot starts were input as the UDDSH test procedure.

- Repeat data were entered into the database for study/engine/fuel/test procedure combinations. The following guidelines were used to enter repeat tests:
  - If individual repeat data were provided, all tests were included in the database.
  - For repeat data when only an average emission was provided, the average emission was included in the database the same number of times it was tested in the original study. If the number of tests used to compute the average was unknown, the average emission was entered in the database two times.
- In several studies, the emissions data needed to be adjusted because of drift across the time period of the experiment. Some of these studies did not provide the time adjustment criteria implemented in their report analysis. In these cases, adjusted emissions data were not entered into the database. However, four studies did provide time-adjusted emissions data, or the equations to convert to time-adjusted data. The following time-adjusted data were entered into the database and the corresponding time-drift correction equations are provided in Appendix G.
  - SAE2000-01-2890 - time-adjusted values for NO_x and PM on 95CAT 3404E engine and time-adjusted values for NO_x on 96 Series 50 engine.
  - CAPE32-80, VE-1 (Phase I) - time-adjusted values for NO_x, CO, HC, and PM emissions on all engines
  - CRCVE-1 (Phase II) - time-adjusted values for NO_x, CO, HC, and PM emissions on all engines.
  - CRCVE-10 - time-adjusted values for NO_x, CO, HC and PM emissions on all engines.
- In studies where emissions were adjusted for humidity, the humidity-adjusted data was entered.
- In studies where adjustments were made for fuel sulfur or any other fuel property, no adjusted data was entered.

III. MIXED MODELS FOR INDIVIDUAL ENGINE TECH GROUPS

The major task of this project involved analyzing the applicable data in the database in order to obtain models for assessing the impact of a given fuel on diesel-engine emissions. Four different sets of prediction models were developed in this process. These included:

- mixed models based on data from individual engine tech groups
- eigenvector models based on data from individual engine tech groups
- mixed models based on the data from the combined tech groups
- eigenvector models based on the data from the combined tech groups.

Models were developed for NO_x, PM, and HC. The following discussion details the steps taken to obtain the mixed models for individual engine tech groups. It also provides the rationale for many of the analysis decisions.

A. Data Selection

The database described in Section II contained a variety of problems. Among these were missing and incomplete data, incompatible test cycles, unequal sets of repeat data, varying sample sizes for engine tech groups, time-adjusted and unadjusted emissions data, and limited ranges of the fuel properties. Because of these concerns, a smaller subset of the data was selected that was more homogeneous in its composition and more complete in its observations.

1. Response Variables

Several response variables were selected from the various studies outlined in Section II and entered into the database. These included total HC, CO, NO_x, particulates (PM), BSFC, and total work. EPA chose to consider analyses for HC, CO, NO_x, PM and BSFC. During the project, however, there was only enough time and resources available to analyze NO_x, PM and HC. Since no analyses were performed on CO, BSFC, or total work, the corresponding data was not checked for outliers or coding errors.

2. Fuel Properties

Several fuel variables were entered into the database. These included test fuel cetane number, cetane index, amount of cetane improver, type of cetane improver, sulfur, nitrogen, total aromatics, monoaromatics, polyaromatics, initial boiling point, 10 percent distillation, 50 percent distillation, 90 percent distillation, 95 percent distillation, end point of distillation, specific gravity, viscosity, molecular ratio of hydrogen to carbon, oxygen, type of oxygenate, net heating value of fuel, ash content, and cetane difference. Cetane difference was defined as the difference in the cetane number between the test fuel with additive and the base fuel without the additive. Also, natural cetane was computed by subtracting the cetane difference from the test cetane number.

Many of the above fuel variables contained missing data. Either the studies did not report the fuel properties of fuel blends or the fuel property tests were not performed. In order to obtain the largest amount of data possible from the database, EPA chose the following nine fuel properties to be used in the data analyses: natural cetane (NATCET), cetane difference (CETDIFF), total aromatics (TAROM), specific gravity (SPGRAV), 10 percent distillation (T10), 50 percent distillation (T50), 90 percent distillation (T90), sulfur (SULFUR), and oxygen (OXY). In addition, EPA, based on their review of the original studies to determine which second-order effects were actually investigated, included three squared fuel properties. This brought the total number of fuel variables to twelve. The squared fuel properties were natural cetane (NATCET2), cetane difference (CETDIF2), and total aromatics (TAROM2).

3. Repeat Data

Repeat data were entered into the database as described in Section II. In some cases, the same test fuel was run many times (>4). Given this situation, EPA decided to limit the number of repeats for a given engine and fuel combination so as not to over-weight those combinations in the analysis. Therefore, each study-by-engine-by-fuel combination was limited to 4 repeat tests. The following criteria were used in selecting the repeat tests to include in the database.

a. For repeat tests where an average emission was computed from more than 4 observations in the original study, and the individual observations were not available, the average was included 4 times in the database.
b. For repeat tests where individual emissions data from more than 4 tests were available in the study, 4 of the observations were randomly selected and these values were retained in the database.
c. For the SAE961974 study, the following special procedure was used. The average emissions were computed from the two sets of reference fuel runs (3 runs were made in each set). Both averages were input twice to create the set of four observations for this fuel.
d. For the SAE932734 study, 52 tests were conducted that represented 13 “sets” of 4 runs made on the reference fuel. Therefore, 4 of the 13 averages were randomly selected to retain in the database.
e. For choosing the repeat data in the CARB TOXICITY study, composite emissions were computed as follows and then the repeat data were selected:
    - For the LOW AROMATICS fuel, enter the following:
      Day 1: 4 hot starts – randomly choose one hot start for Day 1 and enter into the database
      Day 2: 1 cold, 4 hot starts – average the 4 hot starts then combine this average with the single cold start in the 6/7 and 1/7 weighting to produce a single composite for Day 2
      Day 3: 1 cold, 6 hot starts – average the 6 hot starts into a single value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 3.
      Day 4: 1 cold, 6 hot starts – average the 6 hot starts into a single

value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 4.

- For the REF BLEND fuel, enter the following:
  Day 1: 1 cold, 4 hot starts - average the 4 hot starts then combine this average with the single cold start in the 6/7 and 1/7 weighting to produce a single composite for Day 1
  Day 2: 1 cold, 4 hot starts - average the 4 hot starts then combine this average with the single cold start in the 6/7 and 1/7 weighting to produce a single composite for Day 2
  Day 3: 1 cold, 4 hot starts - average the 4 hot starts then combine this average with the single cold start in the 6/7 and 1/7 weighting to produce a single composite for Day 3
  Day 4: 1 cold, 4 hot starts - average the 4 hot starts then combine this average with the single cold start in the 6/7 and 1/7 weighting to produce a single composite for Day 4
- For the PRE 1993 fuel, enter the following:
  Day 1: 1 cold, 6 hot starts - average the 6 hot starts into a single value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 1.
  Day 2: 1 cold, 7 hot starts - average the 7 hot starts into a single value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 2.
  Day 3: 1 cold, 7 hot starts - average the 7 hot starts into a single value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 3.
  Day 4: 1 cold, 7 hot starts - average the 7 hot starts into a single value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 4.
  Day 5: 1 cold, 7 hot starts - average the 7 hot starts into a single value, then combine this average with the single cold start in a 6/7 and 1/7 weighting to produce a single composite for Day 5. Randomly choose 4 of the 5 composite test results for entry into the database.

4. Test Procedures

Nine different test procedures were identified in the various studies used in the database formulation. EPA chose the test procedures to be used for analysis of each emissions and BSFC. These test procedures included:

- 8MAVL - AVL 8-mode engine test
- 9MODE - 9-mode steady-state engine test
- EPA13 - EPA 13-mode steady-state engine test
- JAP13 - Japanese 13-mode engine test
- R49 - European 13-mode engine test
- UDDS - EPA test schedule for heavy-duty diesel engines, composite of hot- and cold-start
- UDDSH - EPA test schedule for heavy-duty diesel engines, hot-start test

Whenever UDDS or UDDSH data were available, emissions measurements for other test cycles were excluded. The Japanese 13-mode results also were excluded due to the overall low engine load for the cycle. Table 2 outlines the test procedures chosen by EPA for the various emissions analysis.

TABLE 2. TEST PROCEDURES USED IN EMISSIONS ANALYSES

Test Procedure	Emissions				
	NO_x	PM	HC	CO	BSFC
8MAVL	Yes	No	Yes	No	Yes
9MODE	No	No	No	No	No
EPA13	No	No	No	No	No
JAP13	No	No	No	No	No
R49	Yes	No	Yes	No	Yes
UDDS	Yes	Yes	Yes	Yes	Yes
UDDSH	Yes	Yes	Yes	Yes	Yes

5. Tech Groups

EPA decided to separately analyze the diesel emissions data by engines with similar classifications of technology. The criteria outlined in Table 3 were established by EPA and used to classify each engine in the database according to a technology group (hereafter designated as 'tech group'). Table 4 identifies the non-overlapping tech group designations used for the engine-by-study combinations in the database.

6. Tech Groups Selected for Analysis

Table 5 lists the analysis decisions made by EPA for each tech group and emission. Note that for NO_x, some tech groups have been combined for analysis: F-DD includes F and DD, P-NN includes P and NN, and Q-OO includes Q and OO. In each case, the only difference between the two tech groups was the presence or absence of an oxidation catalyst, which was assumed to have no impact on NO_x. The "Model" designation indicates that an analysis was requested, "Set Aside" indicates that the data for that particular tech group was to be estimated using models developed from other tech groups, "N/A" indicates that the tech group classification was not used for that emission, and "No Data" indicates there were no data available in the database for that tech group and emission.

TABLE 3. TECH GROUP DEFINITIONS BY ENGINE CLASSIFICATIONS

Tech Group Category	Governed Speed (rpm)	Injector Type	Aspiration	HP	Displacement (L)	Oxy Catalyst	Injection Control	Injection Type	Cycle	Other
B	Any	Any	Turbo	Any	Any	No	Mechanical	Direct	2-stroke	
F	≤3000	Inline or rotary	Turbo	Any	≤9.4	No	Mechanical	Direct	4-stroke	
G	2100-2400 incl.	Unit	Turbo	Any	>9.4	No	Mechanical	Direct	4-stroke	
H	2100-2400 incl.	Inline or rotary	Turbo	Any	>9.4	No	Mechanical	Direct	4-stroke	
I	≤2000	Unit	Turbo	<500	>9.4	No	Mechanical	Direct	4-stroke	
L	Any	Any	Turbo	Any	Any	No	Electronic	Direct	Any	EGR
P	≤3000	Unit	Turbo	Any	≤9.4	No	Electronic	Direct	4-stroke	
Q	≤3000	Inline or rotary	Turbo	Any	>9.4	No	Electronic	Direct	4-stroke	
R	2100-2500 incl.	Unit	Turbo	Any	>9.4	No	Electronic	Direct	4-stroke	
T	≤2000	Unit	Turbo	<500	>9.4	No	Electronic	Direct	4-stroke	
V	≤2000	Inline or rotary	Turbo	Any	>9.4	No	Electronic	Direct	4-stroke	
X	Any	Any	Any	Any	Any	No	Mechanical	Indirect	4-stroke	
DD	≤3000	Inline or rotary	Turbo	Any	≤9.4	Yes	Mechanical	Direct	4-stroke	
NN	≤3000	Unit	Turbo	Any	≤9.4	Yes	Electronic	Direct	4-stroke	
OO	≤3000	Inline or rotary	Turbo	Any	≤9.4	Yes	Electronic	Direct	4-stroke	
ZZ	≤3000	Inline or rotary	Natural	Any	Any	No	Mechanical	Direct	4-stroke	

TABLE 4. TECH GROUP CLASSIFICATION BY ENGINE ID AND STUDY ID

Tech Group	Engine ID	Study ID
T	DDC-SWRI	ACEA
T	06RE001123	CARB-LOCO
I	34705128	CARB-TOXIC
F	V_STD	EPEFE
F	X_STD	EPEFE
V	Y_STD	EPEFE
Q	Z_STD	EPEFE
F	V_+2	EPEFE
F	X_+2	EPEFE
V	Y_+2	EPEFE
Q	Z_+2	EPEFE
F	V_-2	EPEFE
F	X_-2	EPEFE
V	Y_-2	EPEFE
Q	Z_-2	EPEFE
L	HDEWG EGR	HDEWG II
L	HDEWG EGR T2	HDEWG II
L	HDEWG EGR T3	HDEWG II
T	HDEWG No EGR	HDEWG II
P	T444E	SAE1999-01-1117
T	1999-01-1478-1	SAE1999-01-1478
F	3606-1	SAE1999-01-3606
L	04 SWRI/CAT 10.3	SAE2000-01-2890
T	95 CAT 3406E	SAE2000-01-2890
T	95 CUMMINS N14	SAE2000-01-2890
P	96 SERIES 50	SAE2000-01-2890
ZZ	790490-1	SAE790490
F	790490-2	SAE790490
T	852078-1	SAE852078
ZZ	881173-1	SAE881173
F	DTA466 PROTO	SAE902172
T	902173-1	SAE902173
B	AIR RESTRICTION	SAE910735
B	BASELINE	SAE910735
B	THROTTLE DELAY	SAE910735
R	912425-1	SAE912425

TABLE 4 (CONT'D). TECH GROUP CLASSIFICATION BY ENGINE ID AND STUDY ID

Tech Group	Engine ID	Study ID
F	2	SAE922214
ZZ	3	SAE922214
F	922267-1	SAE922267
F	932685-1	SAE932685
R	S60 PROTO	SAE932731
T	932734-1	SAE932734
T	932767-1	SAE932767
T	932800-N14	SAE932800
T	S60PROTO	SAE942019
Q	466216-1	SAE942053
OO	466216-2	SAE942053
T	SN6R6344	SAE942053
G	961973-1	SAE961973
P	L15220	SAE961974
NN	L15220-HIPT	SAE961974
NN	L15220-LOW	SAE961974
R	A	SAE970758
DD	B	SAE970758
F	C	SAE970758
ZZ	D	SAE970758
H	8460.41-10	SAE971635
H	8460.41-8.7	SAE971635
H	8460.41-9.2	SAE971635
F	972894-1	SAE972894
T	S60-0/98	SAE972898
T	S60-0	SAE972904
T	S60-3	SAE972904
T	S60-5	SAE972904
T	VE_10_1	VE 10
T	VE_10_2	VE 10
OO	VE_10_3	VE 10
OO	VE_10_4	VE 10
T	VE_10_5	VE 10
G	NTCC 400	VE-1_PHASE I
T	DDC 60	VE-1_PHASE I
X	NIC 7.3	VE-1_PHASE I
T	6R-510/6067G740	VE-1_PHASE II

TABLE 5. ANALYSIS IDENTIFICATION FOR TECH GROUP BY EMISSIONS

Tech Group	Emissions				
	NO_x	PM	HC	CO	BSFC
B	Model	Model	Model	Model	Model
F	N/A	Model	Model	Model	N/A
DD	N/A	Set Aside	Set Aside	Set Aside	N/A
F-DD	Model	N/A	N/A	N/A	N/A
G	Set Aside	Set Aside	Set Aside	Set Aside	No Data
H	Model	No Data	Model	No Data	Model
I	Set Aside	Set Aside	Set Aside	Set Aside	Set Aside
L	Model	No Data	Model	No Data	Model
P	N/A	Model	Model	Model	N/A
NN	N/A	No Data	No Data	No Data	N/A
P-NN	Model	N/A	N/A	N/A	No Data
Q-OO	Model	N/A	N/A	N/A	Model
Q	N/A	No Data	Model	No Data	N/A
OO	N/A	Model	Model	Model	N/A
R	Set Aside	Set Aside	Set Aside	Set Aside	Set Aside
T	Model	Model	Model	Model	Model
V	Set Aside	Set Aside	Set Aside	Set Aside	Set Aside
X	Set Aside	Set Aside	Set Aside	Set Aside	No Data
ZZ	Set Aside	Set Aside	Set Aside	Set Aside	Set Aside

7. Tech Groups by Fuel Variables

After the tech group definitions were established, frequency tables were used to identify the number of tests available in the database for analysis by tech group and emission. As stated above, EPA defined a set of 12 fuel variables to use in the analyses. For tech groups with small sample sizes, EPA designated a subset of the original 12 fuel variables to use in the analyses. This was based on a review of the original studies to determine which fuel properties were actually investigated. Tables 6, 7, and 8 list the variables used by tech group in the analyses for NO_x, PM, and HC, respectively. A "Y" indicates the fuel variable was utilized in the selected analysis.

8. Tech Groups by Test Procedures

As a final division of the database, EPA grouped the test procedures in order to combine the emissions data into similar testing situations that would be appropriate for modeling. Tables 9, 10, and 11 list the test procedures by tech groups used in the database for analyzing NO_x, PM, and HC, respectively. The total number of observations is listed on the last row of each table.

B. Data Screening

The data set described above was initially screened using a variety of graphs and descriptive statistics. Analyses were done separately for the data subsets consisting of a tech group and emissions combination. Scatter plots of each emission versus each linear fuel property were constructed to determine if any transformations of the response variable or the fuel properties were needed, as well as to identify any extremely aberrant data points. Histograms of each emissions variable also were constructed and analyzed. In addition, a variety of descriptive statistics were examined on each emissions variable including means, standard deviations, maximums and minimums.

In the above screening, several data errors were detected and corrected. These mainly consisted of situations where data was entered incorrectly due to errors in translating the data to the correct units. After the corrections were made, a revised data set was generated and used in the remainder of the analysis.

The various plots and statistics indicated that a log transformation of the emissions would help reduce the variation in the emissions variables. Given this result, EPA decided to express NO_x, PM, and HC in natural logarithm units for the remainder of the analyses. Unless otherwise indicated, any reference to emissions in model fitting will indicate the natural logarithm of the emissions was being fit. The natural logarithm will be designated in this report using the label LOG.

There also were isolated instances where the pattern in the scatter plots indicated a nonlinearity in a fuel property. Since these cases occurred for those fuel properties where EPA had already decided to include a squared term of the corresponding fuel property, EPA decided to make no further changes to the selected set of fuel properties.

TABLE 6. FUEL VARIABLES USED IN NO_x ANALYSES

Fuel Variable	Tech Group						
	B	F-DD	H	L	P-NN	Q-OO	T
NATCET	Y	Y	Y	Y	Y	Y	Y
NATCET2		Y		Y		Y	Y
CETDIFF	Y	Y	Y	Y	Y	Y	Y
CETDIF2		Y	Y	Y		Y	Y
TAROM	Y	Y	Y	Y	Y	Y	Y
TAROM2		Y	Y	Y	Y	Y	Y
SPGRAV	Y	Y	Y	Y	Y	Y	Y
T10		Y		Y		Y	Y
T50		Y		Y		Y	Y
T90		Y	Y	Y	Y	Y	Y
SULFUR		Y	Y	Y	Y	Y	Y
OXY						Y	Y
NO. ENGINES	3	11	3	4	1	5	20
NO. OBS	26	254	15	56	27	89	482

TABLE 7. FUEL VARIABLES USED IN PM ANALYSES

Fuel Variable	Tech Group				
	B	F	P	OO	T
NATCET	Y	Y	Y	Y	Y
NATCET2		Y			Y
CETDIFF	Y	Y	Y	Y	Y
CETDIF2		Y		Y	Y
TAROM	Y	Y	Y	Y	Y
TAROM2		Y	Y		Y
SPGRAV	Y	Y	Y	Y	Y
T10		Y		Y	Y
T50		Y		Y	Y
T90		Y	Y	Y	Y
SULFUR		Y	Y	Y	Y
OXY				Y	Y
NO. ENGINES	3	2	1	2	19
NO. OBS	26	56	27	28	465

TABLE 8. FUEL VARIABLES USED IN HC ANALYSES

Fuel Variable	Tech Group							
	B	F	H	L	OO	P	Q	T
NATCET	Y	Y	Y	Y	Y	Y	Y	Y
NATCET2		Y		Y			Y	Y
CETDIFF	Y	Y	Y	Y	Y	Y	Y	Y
CETDIF2		Y	Y	Y	Y		Y	Y
TAROM	Y	Y	Y	Y	Y	Y	Y	Y
TAROM2		Y	Y	Y		Y	Y	Y
SPGRAV	Y	Y	Y	Y	Y	Y	Y	Y
T10		Y		Y	Y			Y
T50		Y		Y	Y		Y	Y
T90		Y	Y	Y	Y	Y	Y	Y
SULFUR		Y	Y	Y	Y	Y	Y	Y
OXY					Y			Y
NO. ENGINES	3	9	3	4	2	1	3	20
NO. OBS	26	207	15	56	28	27	61	482

TABLE 9. TEST PROCEDURE BY TECH GROUP FREQUENCIES FOR NO_x

Test Procedure	Tech Group							Total
	B	F-DD	H	L	P-NN	Q-OO	T	
8MAVL	0	0	0	56	0	0	16	72
R49	0	177	15	0	0	61	0	253
UDDS	26	28	0	0	0	28	104	186
UDDSH	0	49	0	0	27	0	362	438
TOTAL	26	254	15	56	27	89	482	949

TABLE 10. TEST PROCEDURE BY TECH GROUP FREQUENCIES FOR PM

Test Procedure	Tech Group					Total
	B	F	OO	P	T	
UDDS	26	20	28	0	102	176
UDDSH	0	36	0	27	363	426
TOTAL	26	56	28	27	465	602

TABLE 11. TEST PROCEDURE BY TECH GROUP FREQUENCIES FOR HC

Test Procedure	Tech Group								Total
	B	F	H	L	OO	P	Q	T	
8MAVL	0	0	0	56	0	0	0	16	72
R49	0	140	15	0	0	0	61	0	216
UDDS	26	20	0	0	28	0	0	103	177
UDDSH	0	48	0	0	0	27	0	363	438
TOTAL	26	208	15	56	28	27	61	482	903

The following corrections/deletions were made to the database described in Section II before analyses were performed.

1. SAE972898 study with TEST_ID=19, FBATCH=B1, and ENGMSID=S60-0/98 had a very low NO_x value of 3.848. Two other repeats with the same fuel had higher and similar values. Thus, the NO_x value of 3.848 was set to missing.

2. EPA compared the fuel property data to survey data from the Alliance of Automobile Manufacturers (AAM) to determine if fuels in the database were representative. If two or more fuel properties were outside the boundaries of the AAM data and were more than 4 standard deviations from the AAM mean, then that fuel was deleted as it was considered to be unrepresentative of in-use fuels. However, fuels with high cetane number, or low aromatics or low specific gravities, were retained due to the fact that fuels with such properties are being considered as potential low emission fuels of the future. After all analyses, the following six fuels were deleted from the database:
   - SAE932800, fuels 2 and 7
   - SAE881173, fuels 4, 5, and 6
   - SAE1999-01-1117, fuel F-T

3. EPA deleted two engines because they were not representative of in-use engines. These included:
   - SAE910735, ENGMSID=RETARDED TIMING
   - SAE922214, ENGMSID=1

4. The HDEWG study contained two tests that were not included in the original data analysis in the HDEWG report because the data were questionable. Thus, TESTID=H8-5 and TESTID=H8-5N were deleted from the database.

C. Standardization of Fuel Properties

All the analyzed fuel properties were standardized prior to the data analysis efforts. This was done to facilitate the comparisons of the estimated coefficients of the fuel properties as well as to reduce the potential correlation between the linear and squared fuel terms. The standardization for each linear fuel term involved subtracting its mean and dividing the result by the corresponding standard deviation. The unstandardization involved multiplying the standardized fuel term by its standard deviation and then adding its mean to the result. For the squared and interactive fuel properties, standardization was first applied to the corresponding linear terms, and then the result was squared or multiplied. To unstandardize a squared fuel term, each component was unstandardized, and the results were multiplied together. Tables 12, 13, and 14 contain the means and standard deviations of the fuel properties used in the model-building effort for NO_x, PM and HC, respectively.

D. Modeling Issues and Assumption Checks

A mixed model (see Appendix F for details on approach) was chosen to model each of the three emissions variables: LOG(NO_x), LOG(PM), and LOG(HC). In making this decision, EPA decided initially to fit a separate model to the data from each tech group. Later an analysis of the combined database was performed (as described in Section VI). When the data in the tech group was too small to be fit with a mixed model, a regression model based on only fixed effects was utilized. In the initial fits for the mixed models, the fixed effects included the 12 fuel properties listed in Section III.A, and the random effects included the engine terms. Similarly, in the regression fits, the regression models contained the same 12 fuel properties; however, the engine terms were treated as fixed and defined using categorical variables (i.e., the categorical variable was set equal to 1 if an engine was present, and set equal to 0 if an engine was absent).

Random engine-by-fuel interaction terms were initially included in the above mixed models. As an example, the model for LOG(NO_x) using tech group T data and random interaction terms is summarized in Table 15 along with the corresponding model without random interaction terms. However, after observing the nonsignificance of many of the covariance components associated with the random terms, a decision was made by EPA to delete these terms from the models. Thus, none of these random terms were included in the subsequent analyses.

Various checks were made to support the modeling effort. This was done to insure that the selected model would be stable and would provide useful predictions for emissions. The analyses were used to check for outliers, and to check on the validity of the assumption of normality. The outlier checks included examining a plot of the observed versus the predicted emissions values, a plot of the residuals versus the predicted emissions, and a normal probability plot of the residuals. The checks for normality included examining a histogram of the residuals, a normal probability plot of the residuals, and various statistical tests for the residuals, including the Shapiro-Wilks and Kolmogorov-Smirnov tests for normality.

TABLE 12. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR NO_X ANALYSES

Tech Group	Stats	Fuel Property								
		NATCET	CETDIFF	SPGRAV	T10	T50	T90	TAROM	SULFUR	OXY
B (n=26)	Mean	45.384615	1.200000	0.834692	412.907692	502.769231	582.938462	25.169231	723.076923	0
	Stdev	3.213433	2.869983	0.013596	23.874688	40.968034	43.964013	9.214891	819.027566	0
F-DD (n=254)	Mean	50.117323	2.326772	0.841400	430.269291	511.583465	618.016535	22.929809	410.248031	0
	Stdev	4.946857	3.783393	0.017045	38.383432	32.674751	34.823222	8.778672	310.688407	0
H (n=15)	Mean	51.226667	1.086667	0.834333	402.320000	500.840000	634.880000	21.176667	224.466667	0
	Stdev	3.663423	2.375730	0.008068	22.008284	21.329014	29.783389	8.173855	114.400841	0
L (n=56)	Mean	43.387500	3.937500	0.841291	429.200000	491.671429	577.508929	22.558143	209.821429	0
	Stdev	2.588089	3.888845	0.013930	24.759564	15.203429	10.464403	7.934837	162.930654	0
P-NN (n=27)	Mean	42.566667	0.500000	0.828433	372.400000	441.400000	565.400000	21.462222	117.500000	0
	Stdev	2.656921	0.977438	0.007071	17.346602	19.103846	39.785231	6.348429	147.251590	0
Q-OO (n=89)	Mean	49.834831	2.974157	0.842381	430.382022	515.157303	615.820225	24.182921	411.404494	0.293258
	Stdev	3.546256	3.971476	0.011787	45.032692	28.324780	29.111682	5.500256	43.296735	0.869913
T (n=480)	Mean	45.634333	3.956667	0.842201	412.982083	497.179583	593.366250	28.143917	387.483333	0.058396
	Stdev	6.412144	4.990810	0.017158	37.118927	30.891603	28.646254	10.635963	536.061118	0.390912

TABLE 13. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR PM ANALYSES

Tech Group	Stats	Fuel Property								
		NATCET	CETDIFF	SPGRAV	T10	T50	T90	TAROM	SULFUR	OXY
B (n=26)	Mean	45.384615	1.200000	0.834692	412.907692	502.769231	582.938462	25.169231	723.076923	0
	Stdev	3.213433	2.869983	0.013596	23.874688	40.968034	43.964013	9.214891	819.027566	0
F (n=56)	Mean	45.064286	3.344643	0.848689	421.103571	501.382143	605.278571	26.249821	368.035714	0
	Stdev	5.298659	5.444828	0.022377	34.752632	30.912058	29.817571	12.871943	470.963962	0
OO (n=28)	Mean	47.182143	3.839286	0.841607	420.178571	516.857143	605.500000	25.500000	367.750000	0.932143
	Stdev	4.711928	5.005800	0.003604	46.268182	13.713238	14.738461	6.412488	50.435603	1.359726
P (n=27)	Mean	42.566667	0.500000	0.828433	372.400000	441.400000	565.400000	21.462222	117.500000	0
	Stdev	2.656921	0.977438	0.007071	17.346602	19.103846	39.785231	6.348429	147.251590	0
T (n=461)	Mean	45.719046	4.077657	0.842340	412.892408	497.419523	593.847722	28.269995	395.010846	0.060803
	Stdev	6.471472	5.018470	0.017339	37.794104	31.501131	29.294098	10.698129	545.079699	0.398719

TABLE 14. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR HC ANALYSES

Tech Group	Stats	Fuel Property								
		NATCET	CETDIFF	SPGRAV	T10	T50	T90	TAROM	SULFUR	OXY
B (n=26)	Mean	45.384615	1.200000	0.834692	412.907692	502.769231	582.938462	25.169231	723.076923	0
	Stdev	3.213433	2.869983	0.013596	23.874688	40.968034	43.964013	9.214891	819.027566	0
F (n=207)	Mean	49.425604	2.806763	0.843927	428.960386	510.230918	618.508213	24.025699	397.478261	0
	Stdev	4.537949	3.981663	0.016669	40.114760	31.711635	34.047601	8.204711	252.003503	0
H (n=15)	Mean	51.226667	1.086667	0.834333	402.320000	500.840000	634.880000	21.176667	224.466667	0
	Stdev	3.663423	2.375730	0.008068	22.008284	21.329014	29.783389	8.173855	114.400841	0
L (n=56)	Mean	43.387500	3.937500	0.841291	429.200000	491.671429	577.508929	22.558143	209.821429	0
	Stdev	2.588089	3.888845	0.013930	24.759564	15.203429	10.464403	7.934837	162.930654	0
OO (n=28)	Mean	47.182143	3.839286	0.841607	420.178571	516.857143	605.500000	25.500000	367.750000	0.932143
	Stdev	4.711928	5.005800	0.003604	46.268182	13.713238	14.738461	6.412488	50.435603	1.359726
P (n=27)	Mean	42.566667	0.500000	0.828433	372.400000	441.400000	565.400000	21.462222	117.500000	0
	Stdev	2.656921	0.977438	0.007071	17.346602	19.103846	39.785231	6.348429	147.251590	0
Q (n=61)	Mean	51.052459	2.577049	0.842736	435.065574	514.377049	620.557377	23.578361	431.442623	0
	Stdev	1.913775	3.368600	0.014054	44.046517	33.016745	32.752366	4.968470	17.526670	0
T (n=482)	Mean	45.599959	3.961618	0.842301	413.022407	497.165975	593.285892	28.143362	380.153527	0.058154
	Stdev	6.337323	4.983253	0.017170	36.995688	30.755575	28.772793	10.618101	523.049828	0.390116

TABLE 15. COMPARISON OF FITS TO LOG(NO_x) FOR TECH GROUP T WITH AND WITHOUT RANDOM ENGINE-BY-FUEL INTERACTION TERMS

Fuel Property	Coefficients	
	With Random Interactions	Without Random Interactions
INTERCEPT	***1.4784***	***1.4754***
NATCET	-0.00650	***-0.00829***
NATCET2	0.006851	-0.00130
CETDIFF	***-0.02203***	***-0.01927***
CETDIFF2	***0.003788***	***0.004034***
TAROM	***0.03365***	***0.02737***
TAROM2	-0.00018	-0.00077
SULFUR	***0.003383***	***0.004790***
SPDRAV	***0.02456***	***0.02212***
OXY	0.001725	0.001206
T10	0.005924	-0.00384
T50	-0.01356	-0.00418
T90	-0.00569	***-0.00455***

The residuals used in the above analyses were obtained by fitting a mixed model to the emissions variable for each tech group. Although studentized residuals were not available in the mixed model analysis, approximate standardized residuals were computed by dividing each residual by the square root of the residual variance estimate. An observation with an approximate standardized residual that exceeded 4.0 in absolute value was then declared to be an outlier. EPA made the decision to delete any outlying observations identified in these initial fits to the data. However, EPA chose to neither identify nor interactively delete any subsequent outliers occurring in the modeling effort.

If a mixed model could not be fit to a set of data (due to the small sample sizes in some tech groups), the residuals were obtained from fitting a regression model to each emission. In these situations, an observation was declared to be an outlier if its studentized residual exceeded 4.0 in absolute value. Again, EPA chose to delete only outlying observations identified in the initial fit to the data.

After conducting the various model fits, seven observations were identified as outliers. EPA chose to eliminate all seven from the database. The deleted observations are described below:

- For LOG(NO_x) and tech group T, two observations were deleted from the SAE902173 study run with fuel A3 because they had low NO_x values.
- For LOG(PM) and tech group T, four observations were deleted from the SAE942019 study run with fuel C because they had high PM values.
- For LOG(HC) and tech group H, one observation was deleted from the SAE922267 study run with fuel G, TESTID=66. This was a high HC value.

The checks for normality generally supported a normal distribution for the residuals obtained from the model fits. There were five exceptions.

- For LOG(NO_x) and LOG(PM) with tech group T data, the statistical tests indicated rejection of normality for the residuals. However, the distribution of the residuals was very symmetrical, and the normal probability plot appeared to follow a straight line. Since the statistical test may have been falsely influenced by the large sample size, it was decided that the normality assumption was valid.
- For LOG(NO_x) with tech group P-NN, some of the statistical tests supported normality and some rejected it. Since the normal probability plot was reasonably linear, it was decided that the normality test was valid.
- For LOG(HC) with tech groups T and P, the statistical tests indicated rejection of normality for the residuals. In addition, the frequency distributions of the residuals were slightly skewed to the right, and the normal probability plots showed slight nonlinearity in the upper tail of the plot. Attempts were made to remove this skewness by using other transformations, such as inverse HC and square root of HC. Neither of these improved the distribution. Since the skewness was not pronounced, it was decided to retain the assumption of normality.

E. Collinearity Checks

In the context of the modeling described in Section III, a collinearity is a linear combination of the p fuel properties. It has the form

$$a_1F_1 + a_2F_2 + \ldots + a_pF_p = c,$$

where the F_i are the fuel properties, the a_i are constant terms (at least two of which are nonzero), and c is a fixed constant. When a collinearity is an exact relationship (i.e., exactly equaling c) or is approximately exact (i.e., almost equal to c), the estimation procedure in the model can be severely affected.

Severe collinearities can be detected in several different ways. One useful approach is to examine the condition indices of the correlation matrix of the fuel properties. A condition index is the square root of the ratio of the largest eigenvalue of the correlation matrix relative to any other eigenvalue. Weak collinearities are often associated with condition indices around 5-10 while moderate-to-severe collinearities are often associated with values of 30-100.

The largest condition index is labeled the condition number. The condition numbers for the fuel correlation matrices for each tech group and each emission were computed and are listed in Table 16. For the tech groups with limited observations, the condition numbers were infinite as exact collinearities were detected. In these instances a smaller set of fuel properties was selected in order to fit a regression model, and the condition number was recomputed using this smaller set. These smaller sets are defined below:

- For NO_x, PM, and HC for tech group B: only NATCET, CETDIFF, TAROM, and SPGRAV were included.
- For NO_x and HC for tech group H: only NATCET, CETDIFF, squared CETDIFF, TAROM, squared TAROM, SPGRAV, T90, and SULFUR were included.
- For NO_x, PM and HC for tech group P: only NATCET, CETDIFF, TAROM, squared TAROM, SPGRAV, T90, and SULFUR were included.
- For PM and HC for tech group OO: squared TAROM and squared NATCET were excluded.
- For HC for tech group Q: T10 was excluded.
- For NO_x, PM, and HC for all tech groups except T and OO: OXY was deleted as it was a constant.

TABLE 16. CONDITION NUMBERS BY TECH GROUPS AND EMISSIONS

Tech Group	Emissions		
	NO_x	PM	HC
B	163.09	163.09	163.09
F	8.46 [a]	34.24	10.65
H	41.70	NA	41.70
L	9.76	NA	9.76
P	11.24 [a]	NA	9.76
Q	10.88 [a]	NA	62.40
T	4.56	4.51	6.25

[a] For NO_x, tech group F includes DD, tech group P includes NN, and tech group Q includes OO.

The results in Table 16 indicate that severe collinearities exist among the fuel properties for tech group B data for all emissions, and moderate-to-severe collinearities exist for tech group H for NO_x, tech group F for PM, and tech groups H and Q for HC. As described more fully in Section VI, an eigenvector analysis of some of the tech groups was performed when collinearities appeared to be severe.

F. Additional Fuel Terms

The only fuel terms included in the initial modeling effort were the 12 terms listed in Section III. However, late in the program, EPA decided it might be beneficial to make some additional mixed model runs (following the mixed-model procedures described in Section III.D) using two alternative forms of natural cetane. These included the following:

- MODEL 1: Use of total cetane (CETNUM) and squared total cetane (CETNUM2) in place of NATCET, squared NATCET, CETDIFF, and squared CETDIFF(CETDIF2).
- MODEL 2: The addition of NATCET*CETDIFF (NATDIFF) interaction to the previous 12 fuel-term set.

These runs were restricted to LOG(NO_x) mixed-models for the data from tech groups T and F-DD. The results are contained in Tables 17 and 18. Significant coefficients at the α=0.05 level are noted in bold italics. In both tech groups, CETNUM (and NATCET) was significant and CETNUM2 (and NATCET2) was not significant. The interaction term between NATCET and CETDIFF was significant in tech group F-DD, but not significant in tech group T.

TABLE 17. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(NO_x) USING ADDITIONAL TERMS FOR NATURAL CETANE FOR TECH GROUP T

Model 1		Model 2	
Fuel Property	**Standardized Coefficient**	**Fuel Property**	**Standardized Coefficient**
INTERCEPT	***1.4793***	INTERCEPT	***1.4752***
TAROM	***0.02338***	TAROM	***0.02741***
TAROM2	-0.00082	TAROM2	-0.00061
SULFUR	***0.005102***	SULFUR	***0.004767***
SPGRAV	***0.01880***	SPGRAV	***0.02192***
OXY	0.000648	OXY	0.001227
T10	-0.00380	T10	-0.00380
T50	-0.00020	T50	-0.00438
T90	-0.00285	T90	***-0.00434***
CETNUM	***-0.01818***	NATCET	***-0.00785***
CETNUM2	-0.00046	NATCET2	-0.00109
		CETDIFF	***-0.01895***
		CETDIF2	***0.004272***
		NATDIFF	0.001702

TABLE 18. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(NO_x) USING ADDITIONAL TERMS FOR NATURAL CETANE FOR TECH GROUP F-DD

Model 1		Model 2	
Fuel Property	**Standardized Coefficient**	**Fuel Property**	**Standardized Coefficient**
INTERCEPT	***1.5933***	INTERCEPT	***1.5931***
TAROM	***0.02716***	TAROM	***0.02868***
TAROM2	-0.00011	TAROM2	-0.00022
SULFUR	-0.00077	SULFUR	-0.00116
SPGRAV	0.006332	SPGRAV	***0.008390***
T10	***0.006345***	T10	0.006413
T50	0.000400	T50	-0.00264
T90	-0.00086	T90	***-0.00043***
CETNUM	***-0.01453***	NATCET	***-0.01027***
CETNUM2	0.0004686	NATCET2	***0.004309***
		CETDIFF	***-0.01102***
		CETDIF2	***0.002242***
		NATDIFF	***0.006189***

G. Stepwise Mixed Model Fits

An additional modeling application was conducted using a stepwise approach to the mixed model building effort described in Section III.D. Each emission (LOG(NO_x), LOG(PM), and LOG(HC)) was modeled separately by tech group. The mixed model contained the tech group engine terms as random effects and fuel properties as fixed effects. However, the fuel properties were added to the model in a stepwise procedure and the resultant models were compared in order to choose the fuel property that was the most "significant" at that step in the model-building process. The advantages of using a mixed-model approach included (a) better modeling for the engine effects, (b) additional estimates of the components of variance due to the various engine terms, and (c) improved estimation of the coefficients of the fuel properties.

The stepwise selection procedure followed similar guidelines to those used in a standard stepwise regression analysis. This approach was used primarily because software for performing variable selection was not available. Thus, any stepwise method had to be done one-step-at-a-time. Since starting with the simplest model at the initial steps was expected to greatly reduce the run time, this methodology was chosen over other approaches.

The stepwise model-building procedure followed these general guidelines:

- Initially a subset of the data was selected based on the chosen emissions and the given tech group of interest. The data had been cleaned of outliers, and contained no missing data. Also identified at this time was the candidate list of fuel properties to consider in the stepwise fits. This list consisted of all or some subset of the 12 fuel properties listed in Section III.A.
- The candidate fuel properties were standardized by subtracting the mean and dividing by the standard deviation of the fuel property for each observation in the chosen data set.
- The PROC MIXED procedure in SAS was then used to model the emissions using the engine categorical variables as random effects and the standardized fuel terms as fixed effects.
- Each "step" of the process consisted of individually fitting a series of mixed models in which each of the fuel properties in the candidate set were added one-at-a-time to the model containing the chosen fuel properties. For example, when the candidate list included the entire 12 fuel properties of interest, the first "step" included 12 individual mixed model runs. Each run contained the random engine terms and one individual fuel term.
- After each candidate fuel term was separately added to the model in the first step, the fuel term with the smallest p-value below 0.05 (i.e., 5% significance level) was chosen to be included in the model for the next step. If none of the mixed models produced a significant fuel term (p-value <0.05), then the stepwise process was terminated.
- The stepwise process was repeated again using the engine categorical variables as random effects and the fuel term chosen in Step No. 1 as the fixed effect. The remaining fuel terms in the candidate set again were added one-at-a-time, and the one with the smallest p-value below 0.05 was chosen for inclusion in the model. At that point, any terms that became nonsignificant (i.e., p-value $\geq$0.05) were removed from the model and added back into the candidate set.
- The stepwise process continued until no fuel terms produced a p-value below 0.05.
- If at any step a squared term was chosen as the significant term to add to the model, and its corresponding linear term had not been included in the model in prior steps, both the quadratic and the linear terms were forced into the model. This was done in order to maintain hierarchical model-building principles.
- For each model generated, several measures of the adequacy of the fit of the model were computed and compared. These included Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (BIC). These statistics are defined in Appendix F.
- If there was insufficient data in a tech group to generate a mixed model, a standard regression stepwise fit was conducted using categorical variables to represent the engine effects. The criterion for entry and removal of the fuel terms remained at a 0.05 significance level.

The results of the stepwise fits by emissions and tech group are summarized in Tables 19, 20, and 21. Included are the fuel terms in the model with the lowest AIC, as well as the estimated coefficients for the standardized fuel properties.

TABLE 19. ESTIMATED COEFFICIENTS FOR STANDARDIZED FUEL TERMS IN "BEST" STEP OF STEPWISE FIT TO LOG(NO_X)

Fuel Property	Tech Group						
	B	F-DD	H	L	P-NN	Q-OO	T
Intercept	***2.2679***	***1.5982***	***1.5572***	***0.9101***	***1.85015***	***1.5574***	***1.4769***
TAROM		***0.02762***	***0.01721***	***0.02310***	***0.02486***	***0.01205***	***0.02712***
TAROM2					***0.00940***		
SPGRAV	***0.01860***		***0.02315***	***0.02365***			***0.01945***
CETDIFF	***0.01166***	***-.01071***					***-0.01455***
NATCET		***-0.01615***					***-0.01276***
T10		***0.01114***					
T90							***-0.00707***
AIC	-128.5	-1260.1	-60.0	-294.7	R^2=0.762	-350.5	-2102.0

TABLE 20. ESTIMATED COEFFICIENTS FOR STANDARDIZED FUEL TERMS IN "BEST" STEP OF STEPWISE FIT TO LOG(PM)

Fuel Property	Tech Group				
	B	F	OO	P	T
Intercept	***-1.5078***	***-2.0228***	***-2.3585***	***-2.22371***	***-1.9430***
TAROM		***0.08588***	***0.06839***	***0.01662***	***0.02201***
SPGRAV	***0.09147***				
CETDIFF	***-0.04596***		***-0.03314***		***-0.02440***
NATCET				***0.03721***	***-0.04427***
T50		***0.05110***			***0.03287***
OXY			***-0.07850***		***-0.03104***
SULFUR					***0.04578***
CETDIF2					***0.01203***
AIC	-59.1	-149.9	-55.3	R^2=0.833	-1114.0

TABLE 21. ESTIMATED COEFFICIENTS FOR STANDARDIZED FUEL TERMS IN "BEST" STEP OF STEPWISE FIT TO LOG(HC)

Fuel Property	Tech Groups							
	B	F	H	L	OO	P	Q	T
Intercept	**-0.7539**	**-1.7909**	**-2.87314**	**-2.0209**	**-1.4530**	**-1.29275**	**-1.6363**	**-2.0107**
TAROM			**0.07367**	**0.04953**		**-0.11548**		
TAROM2						**-0.10366**		
SPGRAV	**-0.1060**						**-0.06753**	**0.08413**
CETDIFF	**-0.01877**	**-0.04688**		**-0.07579**	**-0.2600**			**-0.2910**
CETDIF2				**0.02751**				**0.06816**
NATCET	**-0.1134**	**-0.07625**		**-0.1188**	**-0.3465**			**-0.2289**
NATCET2		**0.02740**		**0.01789**				**0.04611**
T10								**-0.08350**
T50		**-0.09012**						**-0.1233**
T90						**-0.11545**		
AIC	-81.1	-255.0	R^2=0.672	-136.9	21.5	R^2=0.837	-91.0	148.8

IV. EIGENVECTOR MODELS FOR SEPARATE ENGINE TECH GROUPS

The eigenvector approach discussed in Appendix E was partially implemented in this project to provide a comparison to the results of the stepwise fits given in Section III.G.

A. Analysis Steps

The steps utilized in this process are listed below.

- Collect a set of emissions data as a function of engine and fuel data
- Assure that the assumptions of a correct model and a normal distribution are valid
- If any assumptions are invalid, consider appropriate data transformations or additional terms in the model (such as nonlinear or interactive terms) and apply as necessary
- Compute the correlation matrix of the fuel properties
- Determine the eigenvectors (i.e., eigenfuels) of the fuel properties
- Regress the emissions variable on the eigenfuels, and, if appropriate, include engine variables in the model
- Delete “inappropriate” eigenfuels from the analysis
- Regress the emissions variable on the remaining eigenfuels and on the set of engine variables
- Delete non-contributing fuel properties and re-regress to obtain a new model.

Each tech group within each emissions data set was modeled separately. Some of the steps involved running SAS procedures, while others necessitated the construction of SAS code as no software was readily available for use in the analysis. The steps involved in the eigenvector analysis are listed below.

- STEP 1: Run a traditional least squares regression model (PROC REG) on log(emissions) using categorical variables to describe the engine effects. The engine variables were coded 0 if an engine was not used and 1 if an engine was used in obtaining a given emissions observation. Obtain the resultant residuals from this model. These residuals represent the engine-adjusted emissions and are used in the subsequent analyses. This approach was taken in order to avoid having to adjust for the engine effects in subsequent runs, and to simplify the analysis.
- STEP 2: Run PROC PRINCOMP on the residuals from Step 1 to obtain the eigenvectors of the correlation matrix describing the fuel effects. All linear fuel properties are standardized prior to analysis. However, to simplify the computations, the squared fuel properties are post-standardized. This is accomplished by first squaring the original linear unstandardized fuel values and then standardizing them.
- STEP 3: Run PROC REG on the residuals from Step 1 using the eigenvectors from Step 2 as the independent variables. Identify the t-ratios and sums of squares (SS) associated with each eigenvector. Initially reject

any eigenvectors with t-ratios smaller than 1.96 in absolute value, or that contribute less than 1 percent to the model SS.

- STEP 4: Refit the residuals from Step 1 using PROC REG with the retained eigenvectors from Step 3. Transform the eigenvectors from this fit back into the original fuel variables and calculate the sum of squares for the fuel variables using SAS code developed for this step.
- STEP 5: Eliminate those fuel variables that individually contribute less than 1% to the model SS.
- STEP 6: Run the chosen set of fuel properties in Step 5 in a mixed-effects model using the log(emissions) as the response variable. In fitting this model, use the pre-standardization method for the fuel terms; this is done to keep everything compatible. Also include the engine terms as random effects and the fuel terms as fixed effects in the mixed model. The result of the Step 6 analysis is considered the final model.

B. Models for Selected Engine Tech Groups

Due to time constraints in the project, only a few sets of data were analyzed using the above methodology. This included the data from tech groups T, Q-OO, B and H for LOG(NO_x) analyses, and from tech groups F, OO, and B for LOG(PM) analyses. Except for tech group T (which comprised about half of the data in the database), these groups exhibited the most severe collinearities. The results of the eigenvector analysis for each of these groups are given below. Significant coefficients at the α=0.05 level are noted in bold italics.

- LOG(NO_x) for tech group T: The twelve fuel properties outlined in Table 6 were included in the analysis for tech group T. After Step 3, six eigenvectors were retained (Nos. 1, 3, 5, 9, 10, and 11). After Step 5, fuel variables T10 and OXY were eliminated. Table 22 lists the results of the final mixed-effects model from Step 6, with bold italics designating the coefficients that are significant at the 0.05 significance level.

- LOG(NO_x) for tech group Q-OO: The twelve fuel properties outlined in Table 6 were included in the analysis for tech group Q-OO. After Step 3, five eigenvectors were retained (Nos. 1, 2, 4, 8, and 12). After Step 5, no fuel variables were eliminated. Table 23 lists the results of the final mixed-effects model from Step 6.

- LOG(NO_x) for tech group B: The four fuel properties outlined in Table 6 were included in the analysis for tech group B. After Step 3, two eigenvectors were retained (Nos. 1 and 2). After Step 5, no fuel variables were eliminated. Table 24 lists the results of the final mixed-effects model from Step 6.

TABLE 22. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(NO_X) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP T

Fuel Property	Coefficient
INTERCEPT	***1.4755***
NATCET	***-0.00857***
NATCET2	-0.00131
CETDIFF	***-0.01886***
CETDIF2	***0.003787***
TAROM	***0.02756***
TAROM2	-0.00102
SULFUR	***0.004659***
SPGRAV	***0.02126***
T50	***-0.00720***
T90	***-0.00371***

TABLE 23. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(NO_X) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP Q-OO

Fuel Property	Coefficient
INTERCEPT	***1.5161***
NATCET	0.006060
NATCET2	***-0.02142***
CETDIFF	***-0.01409***
CETDIF2	0.002217
TAROM	***0.05794***
TAROM2	***0.05869***
SPGRAV	***-0.05890***
OXY	***0.03873***
T10	0.01441
T50	***0.03775***
T90	0.004395
SULFUR	0.01118

TABLE 24. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(NO_X) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP B

Fuel Property	Coefficient
INTERCEPT	***2.2680***
NATCET	-0.02699
CETDIFF	***0.01105***
TAROM	-0.04784
SPGRAV	0.06087

- LOG(NO_x) for tech group H: The eight fuel properties outlined in Table 6 were included in the analysis for tech group H. After Step 3, two eigenvectors were retained (Nos. 1, and 2). After Step 5, no fuel variables were eliminated. Table 25 lists the results of the final mixed-effects model from Step 6.

TABLE 25. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(NO_x) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP H

Fuel Property	Coefficient
INTERCEPT	***1.5313***
NATCET	0.002445
CETDIFF	-0.01422
CETDIF2	0.003374
TAROM	***0.03925***
TAROM2	0.02412
SPGRAV	***0.05548***
T90	-0.00425
SULFUR	-0.01177

- LOG(PM) for tech group F: The 11 fuel properties outlined in Table 7 were included in the analysis for tech group F. After Step 3, four eigenvectors were retained (Nos. 1, 2, 6 and 9). After Step 5, no fuel variables were eliminated. Table 26 lists the results of the final mixed-effects model from Step 6.

- LOG(PM) for tech group OO: The ten fuel properties outlined in Table 7 were included in the analysis for tech group OO. After Step 3, four eigenvectors were retained (Nos. 2, 3, 5 and 8). After Step 5, no fuel variables were eliminated. Table 27 lists the results of the final mixed-effects model from Step 6.

- LOG(PM) for tech group B: The four fuel properties outlined in Table 8 were included in the analysis for tech group B. After Step 3, all four eigenvectors were retained. After Step 5, no fuel variables were eliminated. Table 28 lists the results of the final mixed-effects model from Step 6.

TABLE 26. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(PM) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP F

Fuel Property	Coefficient
INTERCEPT	***-2.0069***
NATCET	***-0.1467***
NATCET2	0.009533
CETDIFF	-0.01694
CETDIF2	0.01161
TAROM	-0.01737
TAROM2	***-0.04259***
SPGRAV	-0.07697
T10	***-0.08997***
T50	***0.2636***
T90	0.02033
SULFUR	0.01328

TABLE 27. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(PM) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP OO

Fuel Property	Coefficient
INTERCEPT	***-2.3791***
NATCET	-0.1551
CETDIFF	-0.06822
CETDIF2	0.02243
TAROM	-0.06043
SPGRAV	0.03981
OXY	-0.03057
T10	-0.00446
T50	0.02021
T90	0.04980
SULFUR	0.09754

TABLE 28. COEFFICIENTS OF STANDARDIZED FUEL PROPERTIES FOR LOG(PM) AFTER EIGENVECTOR ANALYSIS FOR TECH GROUP B

Fuel Property	Coefficient
INTERCEPT	***-1.5060***
NATCET	***1.1894***
CETDIFF	***-0.02985***
TAROM	***2.4211***
SPGRAV	***-2.0368***

V. MIXED MODELS BASED ON COMBINED ENGINE TECH GROUPS

An additional analysis was performed that was based on modeling a composite of the entire database. This was done in order to determine if the data from the various tech groups could be combined together rather than analyzed separately. The data from all available tech groups were included in the analysis, but because of the time constraints of the project, only LOG(NO_x), LOG(PM), and LOG(HC) were modeled. The database was expressed in two different forms. In the first grouping, the data from the various tech groups were combined, but the repeat data were averaged. This meant that the emissions values from each study-by-engine-by-fuel combination were averaged over all repeat tests to obtain a single "average" emissions value. In the second grouping, all the data were again combined, but the repeat data were not averaged.

The reason for dividing the data into the two groups was to simplify the data analysis. When the average-repeat data were analyzed, it was not possible to estimate the engine-by-fuel interaction terms. Thus, the analysis could be performed using standard fixed-effects models. It also was easier to identify any significant tech-group-by-fuel interactions that might add complexity to the model. When the repeat data was not averaged, the engine-by-fuel interaction terms could be estimated. Thus, mixed model procedures could be applied, and a final model could be obtained for prediction purposes.

A. LOG(NO_x) Analyses

1. Stepwise Regression Fits With Average-Repeat Data

Initially, using only the average-repeat data, a stepwise regression model was fit to LOG(NO_x). The candidate variables for the model included:

- 55 engine categorical variables to represent the 56 engines in the study,
- 9 linear fuel terms, including NATCET, CETDIFF, TAROM, SULFUR, SPGRAV, T10, T50, T90, and OXY,
- the corresponding 9 squared fuel terms,
- 28 fuel-by-fuel interaction terms, excluding all interactions with SULFUR (based on an EPA decision),
- 117 tech-group-by-fuel interactions (based on 13 tech groups and 9 fuel terms), and
- 117 tech-group-by-squared fuel interactions.

Previous screening analyses identified two average-repeat values that were considered outliers since their standardized residuals were greater than 4.0 in absolute value. These two outliers were from the SAE902173 study with fuel A3 and the SAE972898 study with fuel 2D. The data remaining after removing these outliers is described in Table 29.

TABLE 29. AVERAGE-REPEAT DATA AVAILABLE FOR LOG(NO_x) AFTER DELETION OF OUTLIERS

Tech Group	No. of Observations	No. of Engines
B	13	3
F-DD	82	11
G	9	1
H	15	3
I	3	1
L	31	4
P-NN	9	1
Q-OO	35	5
R	10	2
T	166	20
V	17	3
X	10	1
ZZ	4	1
TOTAL	404	56

All fuel properties included in the models based on the average-repeat data were standardized prior to the modeling effort, but after the removal of any outliers. The means and standard deviations used in the standardization are given in Table 30.

TABLE 30. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR LOG(NO_x) ANALYSIS USING AVERAGE-REPEAT DATA

Fuel Property	Sample Size	Mean	Standard Deviation
NATCET	404	47.312030	6.038833
CETDIFF	404	3.055198	4.372982
SPGRAV	404	0.841357	0.016072
T10	404	419.738614	39.302783
T50	404	502.880941	32.667426
T90	404	602.655693	33.394534
TAROM	404	25.529828	9.883846
SULFUR	404	478.297277	642.394091
OXY	404	0.116807	0.592569

EPA performed a stepwise regression analysis on the average-repeat data. All runs were made using the SAS procedure PROC REG with the MODEL=STEPWISE option. Also, a significance level of 0.05 was used for both entry and removal of candidate terms for the model. The steps followed in the analysis included the following:

1. Force the 55 categorical engine variables into the model. Limit the candidate variables for entry into the model to the 9 linear fuel terms. Identify and retain the significant candidate fuel terms.
2. Force into the model the 55 categorical engine variables and the significant linear fuel terms identified in Step 1. Limit the candidate list of entering variables to the 9 squared fuel terms. Identify and retain the significant candidate terms.
3. Force into the model the 55 categorical engine variables and the significant terms identified in Steps 1 and 2. Limit the candidate list of entering variables to the 28 fuel-by-fuel interaction terms. Identify and retain the significant candidate terms.
4. Force into the model the 55 categorical engine variables and the significant terms identified in Steps 1-3. Limit the candidate list of entering variables to the 117 tech-group-by-fuel interaction terms. Identify and retain the significant candidate terms.
5. Force into the model the 55 categorical engine variables and the significant terms identified in Steps 1-4. Limit the candidate list of entering variables to the 117 tech-group-by-squared- fuel interaction terms. Identify and retain the significant candidate terms.

The results of the stepwise procedure were as follows:

- Six linear fuel terms entered the model in Step 1. These included NATCET, CETDIFF, TAROM, SULFUR, SPGRAV, and T50.
- No significant squared fuel terms were identified in Step 2.
- No significant fuel-by-fuel interactions were identified in Step 3.
- Eleven tech-group-by-fuel interactions were identified in Step 4.
- Since the F-DD*T10 term was significant, the T10 linear term was added to the model to retain hierarchy.
- Six tech-group-by-squared-fuel terms were identified in Step 5. However, only one term, $X*T50^2$, was included in the final model. This choice was made based on use of the C_p criterion in assessing the regression fit.
- Since the $X*T50^2$ term was included in the model, the X*T50 term was also added to retain hierarchy in the model-building process.
- The ZZ*SULFUR term was significant, but it had a variance inflation factor greater than 800. Thus, EPA decided to delete it from the model.
- The X*SULFUR and X*SPGRAV tech-group-by-fuel interaction terms were nonsignificant after the C_p criterion was applied. Thus, EPA decided to delete them from the model.

2. Mixed Models With Combined Data

A mixed-effects model was run on the terms identified in the stepwise process described above. In addition to the terms listed above, 7 tech-group categorical variables were included which represented the 7 tech groups identified in the tech-group-by-fuel interactions. These included categorical variables for tech groups B, F-DD, G, H, L, R, and X. All observations in the combined data set were used, including all repeats. In fitting this model, the standardization method described in Section III.C was used for the fuel terms. All of the fuel terms, fuel interactions, and tech-group categorical variables identified above were included as fixed effects, while the engine terms and the 7 engine-by-fuel interactions were treated as random effects.

The mixed-effects model identified one additional outlier; it corresponded to an observation taken from the EPEFE study run on Fuel EPD6 and engine Z_+2. Thus, the fuel property means and standard deviations used in the standardization for this final model-building effort excluded this observation in addition to the two previously identified outliers. The corresponding means and standard deviations are given in Table 31.

TABLE 31. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR LOG(NO_x) ANALYSIS USING COMBINED DATA

Fuel Property	Sample Size	Mean	Standard Deviation
NATCET	1345	47.160506	5.801869
CETDIFF	1345	2.533606	4.105230
SPGRAV	1345	0.842843	0.015609
T10	1345	421.158810	37.446975
T50	1345	504.488104	31.992201
T90	1345	602.984610	31.866925
TAROM	1345	26.204843	9.624028
SULFUR	1345	446.640892	600.533619
OXY	1345	0.040439	0.330662

The results of the mixed-effects model analysis, after deleting the 3 outliers, are summarized in Table 32. This model will be denoted as Model No. 1. Significant coefficients at the 5 percent significance level are designated using bold italics in this and the coefficient tables that follow it.

TABLE 32. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(NO_x) FROM MIXED-EFFECTS MODEL NO. 1 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***1.5312***
NATCET	-0.00033
CETDIFF	***-0.01187***
TAROM	***0.02679***
SULFUR	0.000644
SPGRAV	***0.02375***
T10	0.003553
T50	***-0.01459***
B*NATCET	***0.03407***
B*CETDIFF	***0.03175***
F-DD*T10	-0.00269
G*TAROM	-0.01145
H*T50	***0.02861***
L*CETDIFF	***0.01758***
L*SPGRAV	0.001954
R*SULFUR	***0.06367***
X*T50	0.01433
$X*T50^2$	**0.02347**
TECH GROUP B	**0.7676**
TECH GROUP F-DD	0.07774
TECH GROUP G	-0.03266
TECH GROUP H	0.05482
TECH GROUP L	**-0.6118**
TECH GROUP R	0.01110
TECH GROUP X	-0.06653

Subsequently, a series of mixed-effects models were run in which the nonsignificant tech-group-by-fuel interaction terms identified in Table 32 were sequentially eliminated from the model. The second model, denoted as Model No. 2, included the same fuel terms as Model No. 1 except the three nonsignificant interactions, F-DD*T10, G*TAROM, and L*SPGRAV, were deleted. The X*T50 interaction was also nonsignificant, but it was retained because of its hierarchy with $X*T50^2$. The estimated coefficients for Model No. 2 are listed in Table 33.

TABLE 33. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(NO_X) FROM MIXED-EFFECTS MODEL NO. 2 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***1.5313***
NATCET	-0.00007
CETDIFF	***-0.01191***
TAROM	***0.02644***
SULFUR	0.000493
SPGRAV	***0.02411***
T10	0.002672
T50	***-0.01461***
B*NATCET	***0.03420***
B*CETDIFF	***0.03159***
H*T50	***0.02871***
L*CETDIFF	***0.01769***
R*SULFUR	***0.06346***
X*T50	0.01557
$X*T50^2$	***0.02399***
TECH GROUP B	***0.7679***
TECH GROUP F-DD	0.07726
TECH GROUP G	-0.03584
TECH GROUP H	0.05417
TECH GROUP L	***-0.6117***
TECH GROUP R	0.01144
TECH GROUP X	-0.06683

The next model, designated as Model No. 3, was based on deleting the nonsignificant tech group categorical variables associated with the tech-group-by-fuel interactions deleted in Model No. 2. Thus, the categorical variables for tech groups F-DD and G were deleted. The categorical variable for tech group L was also nonsignificant, but it was retained because of its hierarchy with the significant interaction fuel term between tech group L and cetane difference. The results are listed in Table 34.

TABLE 34. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(NO_X) FROM MIXED-EFFECTS MODEL NO. 3 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***1.5502***
NATCET	-0.00001
CETDIFF	***-0.01190***
TAROM	***0.02639***
SULFUR	0.000481
SPGRAV	***0.02417***
T10	0.002684
T50	***-0.01463***
B*NATCET	***0.03410***
B*CETDIFF	***0.03157***
H*T50	***0.02870***
L*CETDIFF	***0.01767***
R*SULFUR	***0.06350***
X*T50	0.01565
$X*T50^2$	***0.02401***
TECH GROUP B	***0.7490***
TECH GROUP H	0.03520
TECH GROUP L	***-0.6307***
TECH GROUP R	-0.00752
TECH GROUP X	-0.08577

The final model, designated as Model No. 4, was based on deleting the nonsignificant linear fuel terms identified in Model No. 3. Thus, NATCET, SULFUR, and T10 were deleted from Model No. 3, in sequence. The results for this model are listed in Table 35. The terms that are retained in the model either have significant coefficients, or have nonsignificant coefficients but are included to maintain model hierarchy.

TABLE 35. COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(NO_X) FROM MIXED-EFFECTS MODEL NO. 4 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***1.5500***
CETDIFF	***-0.01141***
TAROM	***0.02812***
SPGRAV	***0.02180***
T50	***-0.01287***
B*NATCET	***0.03222***
B*CETDIFF	***0.03029***
H*T50	***0.02820***
L*CETDIFF	***0.01622***
R*SULFUR	***0.06114***
X*T50	0.005562
$X*T50^2$	***0.02189***
TECH GROUP B	***0.7460***
TECH GROUP H	0.03268
TECH GROUP L	***-0.6301***
TECH GROUP R	-0.01241
TECH GROUP X	-0.07986

B. LOG(PM) Analyses

1. Stepwise Regression Fits With Average-Repeat Data

Initially, using only the average-repeat data, a stepwise regression model was fit to LOG(PM). The candidate variables for the model included:

- 34 engine categorical variables to represent the 35 engines in the study,
- 9 linear fuel terms, including NATCET, CETDIFF, TAROM, SULFUR, SPGRAV, T10, T50, T90, and OXY,
- the corresponding 9 squared fuel terms,
- 36 fuel-by-fuel interaction terms,
- 99 tech-group-by-fuel interactions (based on 11 tech groups and 9 fuel terms), and
- 99 tech-group-by-squared fuel interactions.

Previous screening analyses identified two average-repeat values that were considered outliers since their standardized residuals were greater than 4.0 in absolute value. These two outliers were from the SAE942019 study with fuel C and engine S60PROTO and the SAE922214 study with fuel K and engine 3. The data remaining after removing these outliers is described in Table 36.

TABLE 36. AVERAGE-REPEAT DATA AVAILABLE FOR LOG(PM) AFTER DELETION OF OUTLIERS

Tech Group	No. of Observations	No. of Engines
B	13	3
DD	8	1
F	24	2
G	9	1
I	3	1
OO	18	2
P	9	1
R	17	3
T	160	19
X	10	1
ZZ	3	1
TOTAL	274	35

All fuel properties included in the models based on the average-repeat data were standardized prior to the modeling effort, but after the removal of any outliers. The means and standard deviations used in the standardization are given in Table 37.

TABLE 37. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR LOG(PM) ANALYSIS USING AVERAGE-REPEAT DATA

Fuel Property	Sample Size	Mean	Standard Deviation
NATCET	274	46.446934	6.449916
CETDIFF	274	3.257299	4.757146
SPGRAV	274	0.842215	0.017194
T10	274	415.610949	39.698196
T50	274	501.836861	34.033487
T90	274	599.510219	31.633623
TAROM	274	27.407949	10.656241
SULFUR	274	563.471168	774.10221
OXY	274	0.172226	0.713279

A stepwise regression analysis was performed on the average-repeat data using the guidelines established for the LOG(NO_x) model. All runs were made using the SAS procedure PROC REG with the MODEL=STEPWISE option. Also, a significance level of 0.05 was used for both entry and removal of candidate terms for the model. The steps followed in the analysis were as follows:

1. Force the 34 categorical engine variables into the model. Limit the candidate variables for entry into the model to the 9 linear fuel terms. Identify and retain the significant candidate fuel terms.
2. Force into the model the 34 categorical engine variables and the significant linear fuel terms identified in Step 1. Limit the candidate list of entering variables to the 9 squared fuel terms. Identify and retain the significant candidate terms.
3. Force into the model the 34 categorical engine variables and the significant terms identified in Steps 1 and 2. Limit the candidate list of entering variables to the 36 fuel-by-fuel interaction terms. Identify and retain the significant candidate terms.
4. Force into the model the 34 categorical engine variables and the significant terms identified in Steps 1-3. Limit the candidate list of entering variables to the 99 tech-group-by-fuel interaction terms. Identify and retain the significant candidate terms.
5. Force into the model the 34 categorical engine variables and the significant terms identified in Steps 1-4. Limit the candidate list of entering variables to the 99 tech-group-by-squared- fuel interaction terms. Identify and retain the significant candidate terms.

The results of the stepwise procedure were as follows:

- Four linear fuel terms entered the model in Step 1. These included TAROM, OXY, SULFUR, and SPGRAV.
- No significant squared fuel terms were identified in Step 2.
- Three significant fuel-by-fuel interactions entered the model in Step 3. These included NATCET*CETDIFF, NATCET*SPGRAV, and SULFUR*NATCET.
- Since the NATCET*CETDIFF interaction was significant, the nonsignificant NATCET and CETDIFF linear fuel terms were added to the model to retain hierarchy.
- Seven tech-group-by-fuel interactions were identified in Step 4.
- Two tech-group-by-squared-fuel terms were identified in Step 5. However, only one term, X*$NATCET^2$, was included in the final model. This choice was made based on use of the C_p criterion in assessing the regression fit.
- Since the X*$NATCET^2$term was included in the model, the X*NATCET term was also added to retain hierarchy in the model-building process.
- The SULFUR*NATCET interaction, F*T50 and P*NATCET tech-group-by-fuel interaction terms were nonsignificant at the last modeling step. Thus, EPA decided to delete them from the model.

2. Mixed Models With Combined Data

A mixed-effects model was run on the terms identified in the stepwise process described above. In addition to the terms listed above, 4 tech-group categorical variables were included which represented the 4 tech groups identified in the tech-group-by-fuel interactions. These included categorical variables for tech groups DD, X, ZZ, and OO. All observations in the combined data set were used, including all repeats. In fitting this model, the pre-standardization method was used for the fuel terms. All of the fuel terms, fuel interactions, and tech-group categorical variables identified above were included as fixed effects, while the engine terms and the 6 engine-by-fuel interactions were treated as random effects.

The mixed-effects model identified two sets of outliers in the combined database with all the repeats. Four outliers were from the SAE942019 study with fuel C and engine S60PROTO and one outlier was from the CARB-TOXIC study with fuel PRE1993, engine 34705128, and testid 9H5. The corresponding means and standard deviations after removing these five outliers are given in Table 38.

TABLE 38. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR LOG(PM) ANALYSIS USING COMBINED DATA

Fuel Property	Sample Size	Mean	Standard Deviation
NATCET	996	46.453193	6.257064
CETDIFF	996	2.440361	4.339194
SPGRAV	996	0.843873	0.016041
T10	996	416.939960	34.959236
T50	996	504.930823	32.362881
T90	996	603.181124	32.526123
TAROM	996	28.094638	10.240353
SULFUR	996	572.947088	794.567454
OXY	996	0.054608	0.383293

The results of the mixed-effects model analysis, after deleting the 5 outliers, are summarized in Table 39. This model will be denoted as Model No. 1. Significant coefficients at the 5 percent significance level are designated using bold italics in this and the coefficient tables that follow it.

TABLE 39. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(PM) FROM MIXED-EFFECTS MODEL NO. 1 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.8253***
NATCET	-0.01297
CETDIFF	-0.00590
TAROM	***0.02009***
SULFUR	***0.05579***
SPGRAV	***0.03727***
OXY	***-0.03061***
NATCET*CETDIFF	***0.02726***
NATCET*SPGRAV	0.000329
DD*T50	0.09079
X*TAROM	0.03534
X*SULFUR	0.03458
ZZ*T90	***0.2509***
OO*T10	-0.02018
X*NATCET	-0.05888
$X*NATCET^2$	***0.04618***
TECH GROUP DD	-0.6951
TECH GROUP X	0.3081
TECH GROUP ZZ	0.7162
TECH GROUP OO	-0.4188

Subsequently, a series of mixed-effects models were run in which the nonsignificant tech-group-by-fuel interaction terms identified in Table 41 were sequentially eliminated from the model. The second model, denoted as Model No. 2, included the same terms as Model No. 1 except the nonsignificant interaction terms, DD*T50, X*AROM, X*SULFUR, and OO*T10, were deleted. The estimated coefficients for Model No. 2 are listed in Table 40.

TABLE 40. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(PM) FROM MIXED-EFFECTS MODEL NO. 2 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.8207***
NATCET	-0.01308
CETDIFF	-0.00593
TAROM	***0.02200***
SULFUR	***0.06733***
SPGRAV	***0.03801***
OXY	***-0.02752***
NATCET*CETDIFF	***0.02744***
NATCET*SPGRAV	-0.00027
ZZ*T90	***0.2432***
X*NATCET	-0.06581
$X*NATCET^2$	***0.04710***
TECH GROUP DD	-0.6018
TECH GROUP X	0.3038
TECH GROUP ZZ	0.6861
TECH GROUP OO	-0.4377

The next model, designated as Model No. 3, was based on deleting the tech group categorical variables associated with the tech-group-by-fuel interactions deleted in Models 2 and 3. Thus, the categorical variable for tech groups DD and OO were deleted. The categorical variables for tech groups X and ZZ were also nonsignificant, but they were retained because of their hierarchy with the significant interaction fuel terms between tech group X and SULFUR, and between tech group ZZ and T90. The results for this model are listed in Table 41.

TABLE 41. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(PM) FROM MIXED-EFFECTS MODEL NO. 3 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.8445***
NATCET	-0.01171
CETDIFF	-0.00590
TAROM	***0.02118***
SULFUR	***0.05606***
SPGRAV	***0.04028***
OXY	***-0.03069***
NATCET*CETDIFF	***0.02729***
NATCET*SPGRAV	0.000221
ZZ*T90	***0.2473***
X*NATCET	-0.06888
$X*NATCET^2$	***0.04734***
TECH GROUP X	0.3253
TECH GROUP ZZ	0.7353

The final model, designated as Model No. 4, was based on deleting the nonsignificant fuel interaction terms identified in Model No. 3. Thus, NATCET*SPGRAV was deleted from Model No. 3. The results for this model are listed in Table 42. The terms that are retained in the model either have significant coefficients, or have nonsignificant coefficients but are included to maintain model hierarchy.

TABLE 42. COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(PM) FROM MIXED-EFFECTS MODEL NO. 4 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.8658***
NATCET	-0.01288
CETDIFF	-0.00594
TAROM	***0.02209***
SULFUR	***0.06663***
SPGRAV	***0.03803***
OXY	***-0.02757***
NATCET*CETDIFF	***0.02740***
ZZ*T90	***0.2433***
X*NATCET	-0.06613
X*NATCET^2	***0.04720***
TECH GROUP X	0.3492
TECH GROUP ZZ	0.7326

C. LOG(HC) Analyses

1. Stepwise Regression Fits With Average-Repeat Data

Initially, using only the average-repeat data, a stepwise regression model was fit to LOG(HC). The candidate variables for the model included:

- 55 engine categorical variables to represent the 56 engines in the study,
- 9 linear fuel terms, including NATCET, CETDIFF, TAROM, SULFUR, SPGRAV, T10, T50, T90, and OXY,
- the corresponding 9 squared fuel terms,
- 36 fuel-by-fuel interaction terms,
- 135 tech-group-by-fuel interactions (based on 15 tech groups and 9 fuel terms), and
- 135 tech-group-by-squared fuel interactions.

Previous screening analyses identified three average-repeat values that were considered outliers since their standardized residuals were greater than 4.0 in absolute value. These three outliers were from the SAE942019 study with fuel C and engine S60PROTO, the SAE902173 study with fuel B1 and engine 902173-1, and the SAE972904 study with fuel A and engine S60-5. The data remaining after removing these outliers is described in Table 43.

TABLE 43. AVERAGE-REPEAT DATA AVAILABLE FOR LOG(HC) AFTER DELETION OF OUTLIERS

Tech Group	No. of Observations	No. of Engines
B	13	3
DD	8	1
F	63	9
G	9	1
H	15	3
I	3	1
L	31	4
OO	18	2
P	9	1
Q	17	3
R	10	2
T	165	20
V	17	3
X	10	1
ZZ	7	2
TOTAL	395	56

All fuel properties included in the models based on the average-repeat data were standardized prior to the modeling effort, but after the removal of any outliers. The means and standard deviations used in the standardization are given in Table 44.

TABLE 44. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR LOG(HC) ANALYSIS USING AVERAGE-REPEAT DATA

Fuel Property	Sample Size	Mean	Standard Deviation
NATCET	395	47.135595	5.978797
CETDIFF	395	3.139241	4.396994
SPGRAV	395	0.841749	0.016213
T10	395	419.241013	39.642029
T50	395	502.587848	32.649437
T90	395	602.632658	33.239478
TAROM	395	25.851120	9.900138
SULFUR	395	492.210380	667.296516
OXY	395	0.119468	0.599033

A stepwise regression analysis was performed on the average-repeat data using the guidelines established for the LOG(NO_x) model. All runs were made using the SAS procedure PROC REG with the MODEL=STEPWISE option. Also, a significance level of 0.05 was used for both entry and removal of candidate terms for the model. The steps followed in the analysis were as follows:

1. Force the 55 categorical engine variables into the model. Limit the candidate variables for entry into the model to the 9 linear fuel terms. Identify and retain the significant candidate fuel terms.
2. Force into the model the 55 categorical engine variables and the significant linear fuel terms identified in Step 1. Limit the candidate list of entering variables to the 9 squared fuel terms. Identify and retain the significant candidate terms.
3. Force into the model the 55 categorical engine variables and the significant terms identified in Steps 1 and 2. Limit the candidate list of entering variables to the 36 fuel-by-fuel interaction terms. Identify and retain the significant candidate terms.
4. Force into the model the 55 categorical engine variables and the significant terms identified in Steps 1-3. Limit the candidate list of entering variables to the 135 tech-group-by-fuel interaction terms. Identify and retain the significant candidate terms.
5. Force into the model the 55 categorical engine variables and the significant terms identified in Steps 1-4. Limit the candidate list of entering variables to the 135 tech-group-by-squared- fuel interaction terms. Identify and retain the significant candidate terms.

The results of the stepwise procedure were as follows:

- Four linear fuel terms entered the model in Step 1. These included NATCET, CETDIFF, T50, and T10.
- One significant squared fuel term ($NATCET^2$) was identified in Step 2.
- Two significant fuel-by-fuel interactions entered the model in Step 3. These included NATCET*CETDIFF, and CETDIFF*T90.
- Since the CETDIFF*T90 interaction was significant, the nonsignificant T90 linear term was added to the model to retain hierarchy.
- Seven tech-group-by-fuel interactions were identified in Step 4.
- Three tech-group-by-squared-fuel terms were identified in Step 5. These included $T*CETDIFF^2$, $T*T10^2$, and $F*SPGRAV^2$.
- The C_p criterion was used to determine which of the above terms should be retained in the final model. In assessing the regression fit it was decided to exclude the tech-group-by-fuel interactions and the tech-group-by-squared-fuel interactions. Thus, the following terms were included in the model: NATCET, CETDIFF, T10, T50, $NATCET^2$, NATCET*CETDIFF, and CETDIFF*T90.

2. Mixed Models With Combined Data

A mixed-effects model was run on the terms identified in the stepwise process described above. Since there were no tech-group-by-fuel interaction terms in the model, no tech-group categorical variables were included. All observations in the combined data set were used, including all repeats. In fitting this model, the pre-standardization method was used for the fuel terms. All of the fuel terms and fuel interactions identified above were included as fixed effects, while the engine terms and the 5 engine-by-fuel interactions were treated as random effects.

The mixed-effects model identified two outliers in the combined database with all the repeats. One outlier was from the VE-1_PHASE I study with fuel 0/686, engine DDC 60, and testid 2, while the other outlier was from the CARB-LOCO study with fuel ON HIGHWAY, engine 06RE001123, and testid ON-HWAY H2. The corresponding means and standard deviations after removing these two outliers are given in Table 45.

TABLE 45. FUEL PROPERTY MEANS AND STANDARD DEVIATIONS FOR LOG(HC) ANALYSIS USING COMBINED DATA

Fuel Property	Sample Size	Mean	Standard Deviation
NATCET	1320	46.931045	5.711013
CETDIFF	1320	2.589394	4.129932
SPGRAV	1320	0.843401	0.015576
T10	1320	420.891667	37.55828
T50	1320	504.266894	31.695372
T90	1320	602.845076	31.563171
TAROM	1320	26.552389	9.534263
SULFUR	1320	449.582576	608.302413
OXY	1320	0.041205	0.333734

The results of the mixed-effects model analysis, after deleting the two outliers, are summarized in Table 46. This model will be denoted as Model No. 1. Significant coefficients at the 5 percent significance level are designated using bold italics in this and the coefficient tables that follow it.

TABLE 46. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(HC) FROM MIXED-EFFECTS MODEL NO. 1 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.7296***
NATCET	***-0.1723***
CETDIFF	***-0.09192***
T10	***-0.04079***
T50	***-0.05792***
T90	-0.01252
$NATCET^2$	***0.05088***
NATCET*CETDIFF	***0.08391***
CETDIFF*T90	0.01063

Subsequently, a series of mixed-effects models were run in which the insignificant terms, identified in Table 52, were eliminated from the model. The second model, denoted as Model No. 2, included the same terms as Model No. 1 except the one nonsignificant fuel-by-fuel interaction, CETDIFF*T90, was deleted. The estimated coefficients for Model No. 2 are listed in Table 47.

TABLE 47. ESTIMATED COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(HC) FROM MIXED-EFFECTS MODEL NO. 2 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.7309***
NATCET	***-0.1729***
CETDIFF	***-0.08894***
T10	***-0.04065***
T50	***-0.05555***
T90	-0.01602
$NATCET^2$	***0.05106***
NATCET*CETDIFF	***0.08387***

The final model, designated as Model No. 3, included all the terms in Model No. 2 except the nonsignificant T90 linear term. The results for this model are listed in Table 48. The terms that are retained in the model all have significant coefficients.

TABLE 48. COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(HC) FROM MIXED-EFFECTS MODEL NO. 3 ANALYSIS BASED ON EPA STEPWISE APPROACH

Variable	Coefficient
INTERCEPT	***-1.7241***
NATCET	***-0.1764***
CETDIFF	***-0.09659***
T10	***-0.03684***
T50	***-0.07760***
$NATCET^2$	***0.05125***
NATCET*CETDIFF	***0.08272***

D. Residual Analyses

The residuals from the (fixed+random) portion of the mixed model fits were analyzed to check on the validity of the model assumptions, and to check for the presence of any outliers in the data. These analyses were done for the final mixed models for LOG(NO_x), LOG(PM), and LOG(HC) based on using the combined data without averaging the repeats. Prior to analysis, the residuals were standardized by dividing them by the square root of the Mean Square Error, which corresponded to the estimate of the error covariance component. The output included the following graphs (and these are presented in Appendix H):

- plots of the standardized residuals versus the predicted emissions
- plots of the standardized residuals versus each of the fuel properties in the model
- normal probability plot of the standardized residuals
- histogram of the standardized residuals.

1. LOG(NO_x) Analyses

Only a random pattern was present in the plot of the standardized mixed-model residuals versus the predicted LOG(NO_x). This indicated that no additional transformations (besides the natural logarithm) of the emissions appeared to be needed. The plots of the standardized mixed-model residuals against each of the fuel properties in the model also had random patterns, though there appeared to be a slight wedge shape in the SPGRAV and possibly TAROM plots. No transformation of the fuel properties appeared to be necessary.

In the final mixed model fit for LOG(NO_x), there were 3 large standardized residuals (with values of -5.7, -5.2, and +4.8) out of the 1345 total residuals. Based on the EPA rules for this analysis, which were to identify and delete outliers only a single time (and not iteratively) during the mixed models runs, the observations corresponding to these residuals were not deleted.

The histogram of the standardized mixed-model residuals for LOG(NO_x) was very symmetrical and very bell-shaped (except for the 3 outliers). The corresponding normal probability plot was fairly linear with slight deviations in the tails (partly due to the 3 outliers). However, all of the normality tests were significant, and indicated a rejection of normality for the distribution. Since it was believed that this last result might have been affected by the large sample size, the shape of the histogram was considered to be a better indicator of the distribution. The shape was very symmetrical and the plot supported normality.

2. LOG(PM) Analyses

Only a random pattern was present in the plot of the standardized mixed-model residuals versus the predicted LOG(PM). This indicated that no additional transformation (besides the natural logarithm) of the emissions appeared to be needed. The plots of the standardized mixed-model residuals against each of the fuel properties in the model also had random patterns, though there appeared to be a slight wedge shape in the SPGRAV and OXY plots. No transformations of the fuel properties appeared to be necessary.

In the final mixed model fit for LOG(PM), there were 5 large standardized residuals (with values of -5.3, -4.6, -4.3, -4.0, and +4.5) out of 996 total residuals. Based on the EPA rules for this analysis, which were to identify and delete outliers only a single time (and not iteratively) during the mixed models runs, the observations corresponding to these residuals were not deleted.

The histogram of the standardized mixed-model residuals for LOG(PM) was fairly symmetrical and fairly bell-shaped. The corresponding normal probability plot was fairly linear with slight deviations in the tails. However, all of the normality tests were significant and indicated a rejection of normality for the distribution. Since it was believed that this last result might have been affected by the large sample size, the shape of the histogram was considered to be a better indicator of the distribution. The shape was very symmetrical and the plot supported normality.

3. LOG(HC) Analyses

Only a random pattern was present in the plot of the standardized mixed-model residuals versus the predicted LOG(HC). This indicated that no additional transformation (besides the natural logarithm) of the emissions appeared to be needed. The plots of the standardized mixed-model residuals against each of the fuel properties in the model seemed to have random patterns. No transformations of the fuel properties appeared to be necessary.

In the mixed model fit for LOG(HC), there was 1 large standardized residual (with a value of +5.1) out of 1320 total residuals. Based on the EPA rules for this analysis, which were to identify and delete outliers only a single time (and not iteratively) during the mixed models runs, the observation corresponding to this residual was not deleted.

The histogram of the standardized mixed-model residuals for LOG(HC) was very symmetrical and very bell-shaped (except for the one outlier). The corresponding normal probability plot was fairly linear with slight deviations in the tails (especially for the one outlier). However, all of the normality tests were significant and indicated a rejection of normality for the distribution. Since it was believed that this last result might have been affected by the large sample size, the shape of the histogram was considered to be a better indicator of the distribution. The shape was very symmetrical and the plot supported normality.

VI. EIGENVECTOR MODELS BASED ON COMBINED ENGINE TECH GROUPS

The eigenvector approach outlined in Section IV.A was applied to the combined database in order to separately analyze LOG(NO_x), LOG(PM), and LOG(HC). The database included all the repeat measurements, without limiting the number of repeats and without averaging them. The results of the eigenvector analysis for each of these three responses are given below.

A. Eigenvector Models

The following nine linear fuel properties initially were included in each of the emissions analysis: NATCET, CETDIFF, T10, T50, T90, SPGRAV, TAROM, SULFUR, and OXY. The fuel terms retained in the final mixed models are listed in the accompanying tables along with their estimated coefficients. In order to adjust for the engine effects, the procedures described in Section IV.A were followed. These included fitting engine-adjusted emissions when applying the eigenvector methodology.

1. LOG(NOx) Analyses

After Step 3 of the eigenvector process for LOG(NOx), seven eigenvectors were retained (Nos. 1, 2, 3, 5, 7, 8, and 9). However, after Step 5, no fuel variables were eliminated. Thus, all nine fuel terms were retained. Table 49 lists the results for the mixed-effects model fit to these terms.

TABLE 49. COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(NO_x) AFTER EIGENVECTOR ANALYSIS

Variable	Coefficient
INTERCEPT	***1.5406***
NATCET	***-0.00417***
CETDIFF	***-0.01039***
T10	0.000378
T50	***-0.00569***
T90	***-0.00233***
SPGRAV	***0.01525***
TAROM	***0.02622***
SULFUR	0.002013
OXY	0.000456

2. LOG(PM) Analyses

After Step 3 of the eigenvector process for LOG(PM), six eigenvectors were retained (Nos. 1, 2, 3, 5, 6, and 8). However, after Step 5, no fuel variables were eliminated. Thus, all nine fuel terms were retained. Table 50 lists the results for the mixed-effects model fit to these terms.

TABLE 50. COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(PM) AFTER EIGENVECTOR ANALYSIS

Variable	Coefficient
INTERCEPT	***1.8437***
NATCET	***-0.02801***
CETDIFF	***-0.01135***
T10	0.005301
T50	0.009835
T90	0.007685
SPGRAV	***0.02842***
TAROM	***0.01938***
SULFUR	***0.06140***
OXY	***-0.02666***

3. LOG(HC) Analyses

After Step 3 of the eigenvector process for LOG(HC), five eigenvectors were retained (Nos. 1, 2, 3, 5, and 7). However, after Step 5, no fuel variables were eliminated. Thus, all nine fuel terms were retained. Table 51 lists the results for the mixed-effects model fit to these terms.

B. Methodology Issues

The methodology used in selecting the fuel terms to retain in the eigenvector models described in this section and in Section IV.A was based on formulas given in the tutorial listed in the McAdams, Crawford, and Hadder report (i.e., see pp 92-94) described in Appendix E. In comparing the results given in Table 47 for LOG(NOx) with those obtained by Crawford and McAdams in an analysis of the same data set, an error was found in their report tutorial when using the formulas for partitioning the model sums of squares for a model with multiple fuel terms. Proper partitioning of the model sums of squares is a critical step in determining the terms to delete from the fitted models.

TABLE 51. COEFFICIENTS OF STANDARDIZED VARIABLES FOR LOG(HC) AFTER EIGENVECTOR ANALYSIS

Variable	Coefficient
INTERCEPT	***-1.6927***
NATCET	***-0.1408***
CETDIFF	***-0.1273***
T10	***-0.04138***
T50	***-0.09748***
T90	***-0.02565***
SPGRAV	***0.04785***
TAROM	-0.01019
SULFUR	-0.00885
OXY	0.004819

The correct formula for the partitioning is actually contained in a listing of some Mathlab program code contained on p.96 of the McAdams report. Further, all other steps of the process are correctly listed elsewhere in the report, and, when followed, yield the correct eigenvectors to delete prior to the deletion of the fuel terms. However, with this erroneous step, the fuel terms to delete cannot be properly determined. The consequences of this error are that all the eigenvector models presented in this report probably include too many fuel terms.

Since the error in the description of the eigenvector methodology was not discovered until late in this program, there was not sufficient time or resources to refit the models. However, in re-analyzing the LOG(NO_x) data using the correct procedure, it was found in the first pass through the data that the correct methodology would have led to the deletion of SULFUR, OXY, and T50, instead of their retention. No subsequent fits were made to this data nor to any other emissions data.

VII. MODEL PERFORMANCE

Several different models were examined in this project. Some were generated using mixed models, some were generated using standard linear regression models, and some were generated using eigenvector models. There was not sufficient time or resources available to compare the performance of all of these models. Thus, attention was focused on assessing the emissions prediction performance of the mixed models obtained when analyzing the combined tech-group data set (without averaging the repeat data). This section discusses the results of this evaluation.

A. Methodology

Mixed model techniques were used in this study because, along with other reasons, they can provide good predictors of the aggregate emissions from an overall group of engines. One means of checking the prediction performance of the constructed models is to compare the percent change in emissions observed in the database with the percent change predicted by the corresponding mixed model. The percent change in emissions, denoted % CE, is defined as follows:

$$\% \text{ CE} = 100\% \times (E_{\text{CANDIDATE FUEL}} - E_{\text{BASELINE FUEL}})/(E_{\text{BASELINE FUEL}})$$

where

$E_{\text{CANDIDATE FUEL}}$ = emissions for the candidate fuel
$E_{\text{BASELINE FUEL}}$ = emissions for the baseline fuel.

In order to apply this approach it was necessary to select an overall baseline fuel and to define its properties. In computing the Predicted % CE for a given observation, EPA selected to use a National Average Baseline Fuel as the overall baseline fuel. Its properties were defined as follows:

NATCET = 44.1
CETDIFF= 0.8
TAROM=34.4
SPGRAV=0.85
SULFUR=333
OXY=0
T10=422
T50=505
T90=603

Computing the Observed % CE for a given observation required a different methodology for finding a baseline fuel. Since emissions measurements were not actually observed for the above National Average Baseline Fuel, a different baseline fuel definition was required. In solving this problem, EPA used the following procedure for identifying a baseline fuel:

1. For a given study-engine combination, all the available fuels were considered as candidates for the baseline fuel.

2. The fuel properties for each of these candidate fuels were standardized following the procedures described in Section V.

3. The fuel properties for the National Average Baseline Fuel described above were similarly standardized.

4. The significant linear fuel terms in the final set of mixed models in Section V were identified. These included the following:
   NO_x: NATCET, CETDIFF, TAROM, SPGRAV, T50
   PM: NATCET, CETDIFF, TAROM, SPGRAV, OXY, T90
   HC: NATCET, CETDIFF, T10, T50

5. For each fuel in a given study, and for only the fuel properties listed in Step 4, the absolute difference between each standardized fuel property in the observed fuel and the corresponding standardized fuel property in the National Average Baseline Fuel was computed.

6. For each emissions, the applicable absolute fuel differences in Step 5 were summed, and the fuel with the smallest sum was chosen as the observed Baseline Fuel for the given engine-study combination.

7. The average of the applicable emissions values for all observations of the selected Baseline Fuel was then computed, and this average was used as the observed average baseline emissions.

In the ensuing computations, the observed emissions for a given fuel was based on the average of the repeat emissions observations available for that fuel. This was done in order to obtain a better measure of the difference in the observed emissions between the given fuel and the chosen baseline fuel. Similarly, the predicted emission values were obtained using the appropriate mixed-model equation given in Section V. These predicted values estimate the average emissions for the given fuel. In obtaining the predicted emissions, the predicted values obtained from the mixed models were in terms of natural logarithms. These values were converted into original units of grams/hp-hr by taking anti-logs prior to computing the % CE. Thus, the percent change was in terms of the original emissions units.

As an example, suppose Fuel A was tested in a given engine-study combination, and suppose there was interest in computing the % CE for the average NO_x value associated with this fuel. Using the above methodology, a Baseline Fuel would be found as well as its associated average NO_x value. The Observed % CE would be given by:

$$\text{Observed \% CE} = 100\% \times [(\text{Fuel A})_{\text{AvgObs}} - (\text{Baseline Fuel})_{\text{AvgObs}}] / [(\text{Baseline Fuel})_{\text{AvgObs}}]$$

where

$(\text{Fuel A})_{\text{Obs}}$ = average NO_x value associated with observations on Fuel A
$(\text{Baseline Fuel})_{\text{AvgObs}}$ = average NO_x value associated with observations on Baseline Fuel

Similarly, the Predicted % CE would be given by

$$\text{Predicted \% CE} = 100\% \times [(\text{Fuel A})_{\text{Pred}} - (\text{Avg Baseline Fuel})_{\text{Pred}}] / [(\text{Avg Baseline Fuel})_{\text{Pred}}]$$

where

$(\text{Fuel A})_{\text{Pred}}$ = predicted NO_x value associated with Fuel A
$(\text{Avg Baseline Fuel})_{\text{Pred}}$ = predicted NO_x value associated with National Average Baseline Fuel.

As a final step, the above Observed and Predicted percentages would be compared by taking their difference. This process would be repeated for all the fuels in the given engine-study combination, except for the observed baseline fuels.

B. Comparison of Percent Change in Emissions

1. Percent Change in NO_x Values

The NO_x data contained 56 different engines and 347 averaged observations after deleting the various baseline fuels. Summary statistics for the Observed and Predicted % CE, as well as for their absolute difference, are listed in Table 52. Similarly, a scatter plot of the Observed % CE versus the Predicted % CE is contained in Appendix I. For both the Observed and Predicted % CE, approximately 95 percent of the values are between ±10 percent.

TABLE 52. COMPARISON OF OBSERVED AND PREDICTED % CE FOR NO_x

Statistic	Observed % CE	Predicted % CE	Abs. Value of (Obs. - Pred.) %
Median	-2.82	-2.72	1.61
Mean	-2.63	-2.56	2.08
Std. Dev.	5.17	4.29	1.89
Max Value	13.71	10.34	11.91
Min Value	-19.65	-20.54	0.006
Sample Size	347	347	347

For the absolute values of the differences between the Observed and Predicted % CE for NO_x, the frequency distribution is given in Table 53. Notice that the difference in the % CE values is less than 6 percent for over 95 percent of the data, and less than 2 percent for over 60 percent of the data.

TABLE 53. FREQUENCY DISTRIBUTION OF THE ABSOLUTE DIFFERENCE IN % CE FOR NO_x

Difference Interval	Frequency	Percent	Cumulative %
< 2%	218	62.8	62.8
2% to 4%	81	23.3	86.2
4% to 6%	31	8.9	95.1
6% to 8%	11	3.2	98.3
8% to 10%	4	1.2	99.4
10% to 12%	2	0.6	100.0

2. Percent Change in PM Values

The PM data contained 35 different engines and 239 averaged observations after deleting the various baseline fuels. Summary statistics for the Observed and Predicted % CE, as well as their absolute difference, are listed in Table 54. Similarly, a scatter plot of the Observed % CE versus the Predicted % CE is contained in Appendix I. For both the Observed and Predicted % CE, approximately 95 percent of the values are between ± 30 percent. The largest Observed % CE of 110 percent is due to data taken from the study based on the paper SAE9922214. The second largest value reduces to 50 percent.

TABLE 54. COMPARISON OF OBSERVED AND PREDICTED % CE FOR PM

Statistic	Observed % CE	Predicted % CE	Abs. Value of (Obs. - Pred.) %
Median	-5.68	-5.55	4.91
Mean	-4.82	-5.87	6.32
Std. Dev.	15.53	11.92	5.52
Max Value	110.06	43.48	66.58
Min Value	-41.34	-34.99	0.01
Sample Size	239	239	239

For the absolute values of the differences between the Observed and Predicted % CE for PM, the frequency distribution is given in Table 55. Notice that the difference in the % CE values is less than 8 percent for over 70 percent of the data, and less than 4% for over 40 percent of the data.

TABLE 55. FREQUENCY DISTRIBUTION OF THE ABSOLUTE DIFFERENCE IN % CE FOR PM

Difference Interval	Frequency	Percent	Cumulative %
< 2%	56	23.4	23.4
2% to 4%	41	17.2	40.6
4% to 6%	47	19.7	60.3
6% to 8%	25	10.5	70.7
8% to 10%	24	10.0	80.8
10% to 12%	17	7.1	87.9
12% to 14%	10	4.2	92.1
14% to 16%	4	1.7	93.7
16% to 18%	6	2.5	96.2
18% to 20%	4	1.7	97.9
≥20%	5	2.1	100.0

3. Percent Change in HC Values

The HC data contained 56 different engines and 341 averaged observations after deleting the various baseline fuels. However, because of the wide range of the observed HC emission changes (essentially from -80% to +290%), the data were further reduced to include only HC results for fuel pairs with an Observed % CE between -20 percent and +20 percent. This reduced the data set from 341 to 198 observations. Summary statistics for the Observed and Predicted % CE for both the complete and reduced data sets, as well as for their absolute difference, are listed in Tables 56 and 57. Similarly, scatter plots of the Observed % CE versus the Predicted % CE are contained in Appendix I.

TABLE 56. COMPARISON OF OBSERVED AND PREDICTED % CE FOR HC (WITHOUT RESTRICTIONS)

Statistic	Observed % CE	Predicted % CE	Abs. Value of (Obs. - Pred.) %
Median	-6.67	-13.82	13.80
Mean	-5.62	-7.14	18.76
Std. Dev.	39.99	28.86	21.07
Max Value	287.01	119.75	183.22
Min Value	-81.36	-34.99	0.002
Sample Size	341	341	341

TABLE 57. COMPARISON OF OBSERVED AND PREDICTED % CE FOR HC (WITH RESTRICTIONS [a])

Statistic	Observed % CE	Predicted % CE	Abs. Value of (Obs. - Pred.) %
Median	-4.47	-9.36	10.55
Mean	-3.14	-7.38	14.40
Std. Dev.	9.33	22.26	12.61
Max Value	17.39	65.20	71.87
Min Value	-19.90	-53.74	0.016
Sample Size	198	198	198
[a] $-20\% \leq$ Observed % CE $\leq +20\%$			

Without restrictions on the Observed % CE, approximately 95 percent of both the Observed and Predicted % CE values are between ±50 percent. With restrictions on the Observed % CE, the Observed % CE values for HC range between ±20 percent, and the Predicted % CE values are between ±50 percent.

For the absolute values of the differences between the Observed and Predicted % CE for HC, the frequency distributions are given in Tables 58 and 59. Without restrictions (see Table 58), the difference in the % CE values is less than 20 percent for over 60 percent of the data, and less than 10 percent for approximately 40 percent of the data. With restrictions (see Table 59), the difference in the % CE values is less than 20 percent for over 70 percent of the data, and less than 10 percent for approximately 47 percent of the data.

TABLE 58. FREQUENCY DISTRIBUTION OF THE ABSOLUTE DIFFERENCE IN % CE FOR HC (WITHOUT RESTRICTIONS)

Difference Interval	Frequency	Percent	Cumulative %
< 5%	68	19.9	19.9
5% to 10%	69	20.2	40.2
10% to 15%	51	15.0	55.1
15% to 20%	28	8.2	63.3
20% to 25%	27	7.9	71.3
25% to 30%	37	10.9	82.1
30% to 35%	21	6.2	88.3
35% to 45%	19	5.6	93.8
45% to 55%	11	3.2	97.1
≥55%	10	2.9	100.0

TABLE 59. FREQUENCY DISTRIBUTION OF THE ABSOLUTE DIFFERENCE IN % CE FOR HC (WITH RESTRICTIONS [a])

Difference Interval	Frequency	Percent	Cumulative %
< 5%	49	24.8	24.8
5% to 10%	44	22.2	47.0
10% to 15%	33	16.7	63.6
15% to 20%	20	10.1	73.7
20% to 25%	12	6.1	79.8
25% to 30%	14	7.1	86.9
30% to 35%	11	5.6	92.4
35% to 45%	9	4.6	97.0
≥45%	6	3	100.0

[a] -20% ≤ Observed % CE ≤ +20%

APPENDIX A

DESCRIPTION OF WORK STATEMENT 2-7

STATEMENT OF WORK

WORK ASSIGNMENT 2-7 EPA Contract 68-C-98-169

A. Issuing Office: Environmental Protection Agency
2000 Traverwood Drive
Ann Arbor, Michigan 48105

B. Contractor: Southwest Research Institute
6220 Culebra Road P.O. Box 28510
San Antonio, Texas 78288-0510

C. Statement of Work: Diesel Fuel Impact Model Data Analysis Plan Review

BACKGROUND

Recently, there has been substantial interest on the part of states and others in quantifying the effects of various diesel fuel parameters on emissions to evaluate these as emission reduction strategies. In addition, there have been a number of estimates put forth that in some cases project significant emissions benefits for controlling such parameters as cetane and aromatics levels in diesel fuel. We have concerns with respect to the accuracy, magnitude, and consistency of these projections. Consequently we are preparing to conduct a comprehensive review and analysis of all pertinent, available data. In an attempt to quantify potential reductions in emissions of regulated pollutants from mobile sources that can be associated with diesel fuel parameter control, an effort has been undertaken to study the effects of 2D diesel fuel properties on heavy duty compression-Ignition (CI) engine emissions. EPA's Office of Transportation and Air Quality is undertaking an approach to collect data and provide an assessment of the impact of fuel property controls on emissions using a modeling strategy as the basis for the ultimate assessment.

NATURE OF THE WORK ASSIGNMENT

The purpose of this work assignment is to evaluate the attached proposed data analysis plan for statistical validity /appropriateness with respect to the nature, volume, and method of analysis of the data that will be obtained based on the duty cycle(s) selected which include the highway Federal Test Procedure (FTP), the ISO 8178 CI, and potentially similar duty cycles. The evaluation effort should also aid EPA in determining the need for expansion or consolidation of the data collection based on duty cycle, engine technology or other potential inputs to a fuel impacts model. This work assignment also requires the contractor to assemble a database using data that meets specified criteria. Additionally, the contractor should be prepared to develop a regression model using this database. The purpose of the regression analysis-based model will be to provide as an output the impact on a given regulated emission of changing a single or a set of fuel properties or engine types as independent variables.

WORK PLAN

The contractor shall submit a detailed work plan for EPA approval. The work plan shall include a description of how the Tasks described below are to be satisfied, including an assessment of the sampling plan discussed in Task 1. The work plan shall also include a detailed cost analysis for the effort described here as follows. Additional more detailed workplans will be needed for subsequent tasks as each task is completed and provides the inputs for the subsequent tasks.

Task 1

Objective: The contractor shall assess the scientific and statistical validity and robustness of the proposed sampling strategy / analysis plan as detailed in Attachment 2 of this Work Assignment. This assessment shall include a review of each phase of the analysis plan:

- Data Selection Criteria
- Preparation of Database
- Pre-regression analysis of data
- Regression analysis
- Post regression preparation of model for public consumption

Task Description: The term emissions data refers to Oxides of Nitrogen, Carbon Dioxide, Carbon Monoxide, Particulate Matter by current federal full flow filter methodology. Fuel consumption should also have a bearing on the contractors assessment of the analysis plan. The, contractor shall evaluate the attached proposal on its merit for obtaining a reasonably robust data set for effectively modeling the emissions impacts. We are looking at effects on emissions on conventional heavy-duty diesel engines using conventional 2D diesel fuels as defined in ASTM D975. Transient test data should have been generated based on arid shall conform to 40 CFR 86 and steady state data shall conform to ISO 8178 or alternatively, 40 CFR 89. For the purposes of this analysis we will treat emissions impact data, based on FTP data, as representative of in-use. Other duty cycles may be considered, to the extent the changes in emissions based on fuel quality are statistically similar to the FTP results assuming a 90% confidence. The independent variables of fuel and engine properties arc listed below and the dependent variables as outputs should be NOx, HC, PM, CO, and BSFC.

1) The contractor shall provide an assessment of the adequacy of the attached Diesel Emissions Analysis Plan to facilitate the creation of a model that could effectively predict in-use highway and nonroad engine emissions effects, on a brake specific and / or gram per mile basis, based on the given fuel and engine inputs which could include but not be limited to the following:

   Fuel Properties:

   - T_{90}
   - Total Aromatics
   - Monoaromatics
   - Fuel Sulfur
   - Polyaromatics
   - Density

- API gravity
- Cetane index
- Cetane Improver Type
- Oxygen
- Cetane number

Engine Properties:

- Model Year
- Displacement
- Hot EGR
- DI/IDI
- Rated Speed
- Torque Rise
- Aspiration
- Injection Pressure
- Cooled EGR
- Peak Torque Speed
- Rated Power
- Peak Torque

Test Cycles

- Transient Duty Cycle (assumed FTP)
- Steady State Duty Cycle(s) (assumed ISO 8178-4 C1)
- Modal Steady State Data
- Hot start versus cold start transient data
- Altitude concerns

2) The contractor shall suggest improvements to the attached Diesel Emissions Analysis Plan that would allow for creation of a model that could most effectively predict highway and nonroad engine emissions, on a brake-specific or gram-per-mile basis, based on the given fuel and engine inputs as listed above. Based on contractor expertise, should additional variables be needed as independent model inputs, a list shall be presented to EPA.

3) The contractor shall provide a concise review of the paper: *A Vector Approach to Regression Analysis and Its Implications to Heavy-Duty Diesel Emissions* prepared by H.T. McAdams, November 2000 for Oak Ridge National Laboratory. The contractor shall include in this review an assessment of the appropriateness of the inclusion of the strategy suggested in this paper in the model development effort of subsequent tasks of this Work Assignment.

Task 2

Objective: The contractor shall create a database with data provided by EPA and interested stakeholders. The data that shall be included should contain each of the fields listed in attachment 3 to this Work Assignment. The actual data for input may be found in the papers listed in attachment 4 of this Work Assignment. Additional papers and data sets may be included as identified by the contractor with EPA approval.

Task Description: The contractor shall create a database based on the format and criteria as detailed in attachment 5 of this Work Assignment. All electronic database files shall be saved as .dbf files. The contractor shall include data described by the data analysis plan, with feedback from the EPA Work Assignment Manager. The data from each of the papers

provided or referenced by EPA for this effort should be used as source material for consideration. The specific data sets that are needed from each source will be listed based on priority. Those properties listed as high priority (h) must be included in the data set if available. For those properties listed as low priority (l), they should be included in the data sets however if not *readily* available may be excluded. Data exclusion due to a lack of availability should be approved by the Project Officer (PO), or alternatively, by the Work Assignment Manager (WAM). The contractor shall provide a significant level of post-processing of the data included, to ensure the accuracy of the data and validity of the testing results. EPA will be actively involved and will offer frequent technical guidance in this process as necessary. The data entered into the database shall be included in both the as-received format and converted to standard units for comparison and analysis.

Task 3

Objective: The contractor shall perform a regression analysis or an eigenvector analysis of the available data in the database, with the intent of generating a model appropriate for assessing, on a case by case basis, the impact of a given fuel change on diesel engine emissions.

Task Description: The contractor shall generate a model, based on EPA input, that provides as an output the change in NOx, HC, and PM emissions, based on inputs that include the engine and fuel properties as listed in, but not limited to, Task 1 of this Work Assignment. The model shall utilize either a regression analysis or eigenvector analysis as directed by EPA. The contractor shall include data as described by the data analysis plan and based on feedback from the EPA Work Assignment Manager.

Task 4

Objective: The contractor shall develop a final report that details the contractor's work completed, including any problems encountered, and results from Task 1,2, and 3.

Task Description: The draft final report shall include the following:

1) A detailed description of the contractor's effort undertaken and recommendations for improvement to the diesel emissions analysis plan.

2) A qualitative projection of the resulting model effectiveness with and without the proposed improvements.

3) An assessment of the resulting model effectiveness based on incremental improvements made to the analysis plan based on discussions with the Work Assignment Manager.

A final report shall be submitted to the WAM which addresses and incorporates all EPA comments on the draft final report.

D. Deliverables:

The contractor shall provide at least eight (8) copies of all reports submitted to EPA (excluding weekly reports).

1. Work Plan

The contractor shall submit a detailed work plan to EPA for approval as described above within ten days or receipt of the work assignment.

2. Weekly Reports

The contractor shall provide typed weekly reports that summarize progress to date. (Please see the attachment for required format.)

3. Draft Final Report

The contractor shall provide a draft report summarizing the results of this work assignment. EPA will provide comments on the draft report.

4. Second Draft Report

The contractor shall provide a second draft report which includes modifications approved by EPA based on comments from stakeholders, for review by EPA.

4. Final Report

The contractor shall provide an electronic (MS Word or WordPerfect) and hardcopy final, citable report for use by EPA. The final report shall incorporate comments from EPA regarding the draft final report. The final report shall include a description of and rationale behind the proposed modifications and resulting data analysis plan.

5. Test Data

All data collected while performing the work detailed in this Statement of Work shall become the property of the United States Environmental Protection Agency. Data shall be submitted electronically in ASCII format and in hardcopy format when deemed necessary by the EPA Work Assignment Manager. Additional data formatting requirements shall be continued from previous work, such as Excel data submissions, .dbf formatting for database submissions.

Deliverables shall be submitted as follows:

Deliverable	Proposed Completion Date
Begin Analysis	Upon Delivery
Completion of Analysis	February 2, 2001
Draft Final Report	March 9, 2001
Second Draft Report	April 15, 2001
Final Report	15 days after receiving comments from EPA

E. Task Completion:

The contractor shall prepare a short, typed, weekly status report, due by 1 p.m. each Thursday of the work period, reporting on the progress achieved in the concluded week, technical problems encountered, solutions to those problems (proposed and attempted), and projected activity for the upcoming week. This report shall include an estimate of the percentage of the level of effort expended and a percent of task completed to date. This report shall be submitted to the Project Officer or alternatively to the Work Assignment Manager in the format of the sheet entitled Weekly Report[1]. The contractor representative shall conference call with EPA staff weekly at a time agreed upon by contractor staff and the EPA Project Officer, or alternative the Work Assignment Manager.

In addition, the contractor shall meet with the EPA representative(s) at EPA's facility at least once after the second week, and prior to the submission of the first draft.

F. Work Assignment Manager: Cleophas Jackson (734) 214-4824
(734) 214-4816 FAX
D. Korotney (734) 214-4507

[1] Please see attached

WEEKLY REPORT FORM

TO: Mr. Cleophas Jackson, EPA-AA

FROM: SwRI

SUBJECT: Weekly Progress Report for Work Assignment 2-7
Under EPA Contract 68-C-98-169

PERIOD: From: _____ To: _____

PERCENTAGE OF WORK COMPLETE:

Task or Item	Level of Effort, %	Hours (Total)	Target Completion Date

Problems Encountered:

Projected Activity:

Miscellaneous:

APPENDIX B

DATA ANALYSIS PLAN

DATA ANALYSIS PLAN

A. Data Selection Criteria

1. No. 2 diesel fuel meeting ASTM D-975, made from petroleum
   - No emulsions
   - No pure chemicals
   - Oxygenated fuels will be permitted, including biodiesel up to B20
   - Sulfur doping compounds will be permitted

2. Production-level heavy-duty CI engines
   - Current or expected production
   - Both highway and nonroad engines

3. Engine dynamometer tests
   - Chassis tests will be included in the database, but analyses will determine how and if they should influence the final model

4. Study must have tested at least two different fuels per engine

5. Data is presented in study
   - No reading from graphs
   - We will pursue data from authors as time permits

B. Preparation of Database

1. Select dependent variables for input
   - NO_x, HC, PM, CO, BSFC,
   - Not SO_x, CO_2, or toxics

2. Select independent variables for input

Fuel	Engine	Test
Total aromatics	Model year	Cycle
Polyaromatics	Displacement	Mode/hot/cold
Monoaromatics	Aspiration	Altitude
Sulfur	EGR	
T90	Rated speed	
T95	Rated power	
Density	Injection pressure	
API gravity		
Oxygen		
Cetane index		
Cetane number		
Cetane improver		
Cetane improver type		

3. Enter data
    - Enter data exactly as given in study, without changes, additions, or conversions
    - Multiple entries for averaged results
    - Enter all hot, cold, and/or modal results given in study
    - Separate aromatics entries by test method

4. Adjustments to database
    - Produce composite results from hot and cold if not already available
    - Estimate missing independent variables
        - Fuel properties if only one is missing
        - Engine characteristics if a match with a similar engine can be found
    - Convert API gravity to specific gravity
    - Total aromatics = polyaromatics + monoaromatics
    - Convert aromatics entries to vol% by FIA (ASTM D1319)
        - Will need to develop acceptable correlations between test method results
    - All temps in °F
    - All sulfur in ppmw
    - All emissions in g/bhp-hr

5. Time-drift corrections
    - For studies that included an evaluation of engine drift over time, use a proportional method for adjusting all data so that it better represents reality

C. Pre-regression analysis of data

1. Distribution plots of independent variables
    - Suggestive of valid range of ultimate model
    - Compare to in-use distribution to determine if we have significant data gaps

2. Correlation (fuel property X fuel property) plots of data
    - Highlights collinearities, aids in selection of interactive terms
    - Compare to in-use data to identify data gaps

3. Investigate nonlinearities in fuel property X emission effects
    - Identify terms that should be nonlinear in regression model

4. Investigate differences between tech groups, model year groups, test cycles
    - Informs decisions regarding how to group data for regression analysis
    - Need to know if the slope of the effect is different for different groups
    - ANOVA, analysis of covariance, or discriminate analysis

D. Regression analysis

1. Linear independent variables unless we have a good reason for a nonlinear term

2. Choose to include or exclude specific interactive terms as appropriate

3. Linear (not log) emissions
   - Log(NOx) makes eventual extrapolation complicated

4. Center the independent variables

5. Specify dummy variables for each engine, tech group, and/or study
   - Need to define technology groups according to C.4 results

6. Forward stepwise regression with a p-value cutoff of 0.90

7. Balance over- and under-fitting using, for instance, Mallow's Cp criterion

8. Eliminate terms with significant variance inflation factors

9. Remove outliers (using Rstudent) and influential points (using DFFITS) from the data set

10. Backwards stepwise regression

E. Post-regression preparation of model for public consumption

1. Random balance to eliminate terms that do not contribute significantly to overall explanatory power of the modal

2. Uncenter the independent variables

3. Tech group consolidation

4. Extrapolation

5. Conform model year equations to calendar year estimates

6. Application to nonroad for cases in which nonroad does not exist

APPENDIX C

ASSESSMENT OF THE DATA ANALYSIS PLAN

ASSESSMENT OF THE DATA ANALYSIS PLAN

A. Adequacy of the Data Analysis Plan

Several different aspects of the Data Analysis Plan (given in Appendix B) were examined by SwRI and discussed with EPA. The results of these discussions are given below and the suggested changes are listed in the following subsections.

1. Data Selection Criteria

The data selection criteria appeared to be reasonable. However, SwRI clarified that the data was restricted to what was contained in the papers and reports that EPA provided within the timeframe of the study.

2. Preparation of Database

Initially, five dependent variables (NO_x, HC, PM, CO, and BSFC) were listed for analysis. After discussion, EPA directed that in the interest of time, SwRI should focus on a complete analysis of NOx, PM, and HC. If time permitted, analyses for CO and BSFC were desired, but only after the first three were analyzed.

The Data Analysis Plan listed several engine variables for inclusion in the database. EPA clarified that these variables were added for use in more precisely defining engine technology groups. The Data Analysis Plan also listed several different diesel fuel properties to consider for inclusion in the modeling effort. Since many fuel properties were expected to have some missing values, SwRI did not recommend automatically estimating the missing values. One reason for this argument was because, in the absence of a balanced experimental design, a missing-value estimate could be severely distorted. An alternative suggestion was to investigate the magnitude of missing data for each property and to choose reasonable groups of variables for consideration, namely those with the largest amount of data. Using this approach, it was felt that engine characteristics might be easier to match since similar engines could be expected to have similar properties.

EPA agreed in general that it was not desirable to estimate fuel properties that were not given explicitly in the study. However, EPA noted some limited exceptions including the following:

- If two of the three aromatics variables are given, it is reasonable to estimate the missing aromatics variable. For example, if monoaromatics and total aromatics are given, it is reasonable to estimate polyaromatics.
- If T90 and End Point are given, it may be reasonable to estimate T95.
- Consider using cetane index as a surrogate for cetane number if cetane number is not measured directly and there are no cetane improver additives in the fuel.

SwRI noted that time-drift adjustments were available in several of the engine emissions studies. These ranged from adding a time variable (to represent the engine hours of the test) to the prediction equations, to normalizing the emissions data, to using a proportional method to adjust the emissions data. The last method seemed most reasonable for this project since the final equations needed to be independent of time variables. However, applying this method could be difficult if adequate time data was not available in a particular study paper. Also, there was a question of whether to enter the time-adjusted data in the database or to simply include the original raw data.

After further discussion, EPA decided that a proportional approach should be used for the time adjustments. In this approach, a regression equation of the form

(Predicted) Emissions = b + a × (Time),

where a and b are estimated regression coefficients, is developed based on tests of a single fuel at various times throughout the test program. Using these results, emission measurements for tests on other fuels would be corrected using the following equation:

$$(Corrected)\ Emissions = (Measured)\ Emissions \times \left(\frac{b}{b + a \times (Time)} \right)$$

EPA further suggested that it would be beneficial for those studies having time adjustments to have two sets of emissions data: one adjusted and one unadjusted. The adjusted data could be used in the analysis, and the original data could be stored for later reference. If adequate time information was not available in a study, EPA decided to assume that no time drift was present.

3. Pre-Regression Analysis

SwRI agreed with the steps described in the Data Analysis Plan. Distribution plots of the fuel properties would be helpful in identifying outliers and data gaps, and scatter plots of emissions versus the various fuel properties would help identify any needed variable transformations. In turn, EPA agreed to provide the in-use data for comparisons.

4. Regression Analysis

SwRI reviewed this portion of the plan and suggested a different analysis based on the fitting of a mixed model to the data. These suggested changes are described in the next section.

5. Post-Regression Model Preparation

SwRI agreed with the usefulness of the various steps, but noted that most of them required EPA input and analysis. For example, EPA would need to provide the fuel property limits (i.e., upper and lower bounds) for conducting the random balance analysis. Also, tech group consolidation would depend on EPA decisions. On the issue of extrapolation, SwRI did not favor extrapolation beyond the range of the fuel data. EPA

decided it would handle these aspects of the analysis, as well as the issue of conforming tech groups to calendar-year estimates.

B. Suggested Improvements to the Data Analysis Plan

Several different improvements to the data analysis plan were suggested. These are described below.

1. Changes in Data Entry

In entering the transient test data, EPA requested that only the hot-start or composite data be entered, but not cold-start results. If composite data were available, then it was entered. If only hot-start data were available, then it was entered. If hot-start and cold-start data were available and a composite could be computed, then the computed composite was entered.

If an average emissions result was available, but not the individual test results, the average emissions result was entered "x" times where "x" was the number of repeat tests used to compute the average. If the number of repeats used to obtain an average emissions result was not available, SwRI suggested that a single entry be made. After discussion, EPA recommended that, if there was no way to know how many repeat tests were used to generate the single average value listed in the study, the data should be entered twice. Their rationale was that the single average value is the average of at least two tests, and this is closer to a properly weighted database.

Also, it was suggested that the data initially needed to be converted to the units defined in the EPA database format before it was entered into the database.

2. Changes to Database

Table C-1 lists the database entity definitions as provided by EPA. A complete listing of the variables associated within each entity defined by EPA is provided in Appendix D.

TABLE C-1. DATABASE ENTITY DEFINITIONS

Entity Name	Entity Definition
FBAT_AD	A particular batch of fuel than can be used to power mobile sources during emissions tests. This entity includes fuel properties.
EQUIP_AD	This table represents engine descriptions for the engine tests.
ETEST_AD	Emissions and BSFC results as performed on a particular engine and fuel combination.
EMODE_AD	Steady-state results from a single mode of engine operation.

After reviewing the proposed EPA-defined database entities and translation tables, SwRI recommended and, with EPA approval, implemented the following changes.

a) EQUIP_AD Database Entity

- Added the following categories to the COMPANY translation table:

COMPANY (category)	COMPANY_D (description)
IVECO	IVECO
MAN	Man Euro-II
DEUTZ AG	Deutz engine family 513
HINO	HINO

- Deleted the Test ID Number variable labeled ctr_tst_id
- Added the following variables:

Variable Name	Units	Description
pk_torque	ft-lb	Peak torque
pk_t_speed	rpm	Peak torque speed
cyl_valves		No. valves per cylinder
stroke	in	Piston stroke
bore	in	Diameter of cylinder bore
inj_ctrl		Injection Control Type: ESSCE=Electric SS cruise enabled ESSCD=Electric SS cruise disabled(default) M=mechanical
inj_pcat		Injection equipment/pressure category: R=Rotary P=Pumpline nozzle U=Unit C=Common rail

b) FBAT_AD Database Entity

- Added the following categories to the OXY_TYPE translation table:

OXY_TYPE (category)	OXY_TYPE_N (number)	OXY_TYPE_D (description)
BIODIESEL	8	BIODIESEL
ISOBUTANOL	9	ISOBUTANOL
C11	10	C11
MGLYME	11	MONOGLYME
DGLYME	12	DIGLYME
AROALCOH	13	AROMATIC ALCOHOL
ALIALCOH	14	ALIPHATIC ALCOHOL
POLYETHER	15	POLYETHER POLYOL
GLYETHA	16	GLYCOL ETHER A
GLYETHB	17	GLYCOL ETHER B
GLYETHC	18	GLYCOL ETHER C

- Added the following category to the CETANE_T translation table:

CETANE_TYPE (category)	CETANE_T_N (number)	CETANE_T_D (description)
U	4	UNKNOWN

- Deleted the Work Assignment variable labeled wa_id
- Added the following variable:

Variable Name	Units	Description
cetane_diff		Difference in cetane number between the test fuel with the cetane improver additive and the base fuel without the additive

c) ETEST_AD Database Entity

- Added the following categories to the TEST_PRO translation table:

TEST_PROC (category)	TEST_PRO_D (description)
9MODE	Steady state 9-mode test
ISOD2	IDO "D2" cycle for generator sets with intermittant load
EPA13	Steady state EPA 13-mode test

- Changed the Work Assignment variable labeled wa_id to ENG_MS_ID (engine serial number). In cases where the serial number was unknown, an engine ID was assigned which included the number of the SAE paper or report.

3. Changes to Data Analysis Plan

The Work Statement indicated that a regression analysis or an eigenvector analysis (see Appendix E for details on this methodology) should be performed on the available data. SwRI recommended an alternate approach based on the use of mixed models. This approach assumes the engines in a tech group are random effects and representative of a larger population of engines, while the fuel properties are fixed effects. After discussion, EPA agreed that a mixed model would be appropriate, and should be applied to all tech groups. More details on mixed models are contained in Appendix F.

SwRI suggested standardizing the fuel properties (i.e., subtracting the mean and dividing by the standard deviation) prior to modeling. This removes the scale effects, and it still allows for unstandardization at the end of the analysis phase. EPA accepted this change. SwRI also suggested performing an eigenanalysis of the fuel correlation matrix to determine if any collinearities were present. EPA also accepted this approach.

In the data analysis plan, linear emissions variables were recommended. SwRI suggested the data needed to be checked to see if a log transformation might be useful. After discussion, EPA recommended that SwRI investigate the need for a log or any other type of transformation of the emissions variables.

At this stage, SwRI also suggested fitting an initial model and performing model diagnostics to determine if there were any normality assumption violations, or any outliers present in the data. This approach would include constructing residual plots to aid in making these decisions. In addition, SwRI suggested that severe collinearities be identified prior to the model fits by examining the eigenvalues and eigenvectors of the fuel property correlation matrix. Given this approach, SwRI recommended using only a forward stepwise procedure, rather than a backwards stepwise procedure. It was suggested that the backwards approach be used only if time permitted. EPA concurred with this plan.

In the absence of non-road data, one suggestion was to try to map highway-engine technologies onto non-road engines so that the obtained emissions equations could be used to predict emissions for the non-road fleet. One method for doing this was to use the engine information in the database to obtain typical characteristics of each technology group, and then to determine if any of the non-road engines approximately match these characteristics. EPA decided that its staff would follow-up on this approach since they had more knowledge of the non-road engine technologies.

APPENDIX D

ENTITY NAME/TABLE NAME DEFINITIONS

TABLE D-1. ENTITY NAME/TABLE NAME

Entity Name	Entity Attribute Name	Entity Attribute Definition
EQUIP_AD	eng_ms_id	Mobile source identifier. For engines, their serial number, probably in conjunction with their manufacturer code.
	study_id	Identification number assigned to the analysis/paper/report of interest.
	veh_ms_id	Mobile source identifier. For equipment this would be the serial number which best identifies the equipment as a whole.
	vehclass	Vehicle class. Will have a translation table. Values defined by translation table for this field.
	vehcompany	Vehicle manufacturer. Is designed to align with the MFR_ fields in CFEIS. Has extended translation table in which COMPANY_N will contain the same numeric code as CFEIS for this manufacturer. Values defined by Company translation table for this field.
	engcompany	Engine manufacturer. Is designed to align with the MFR_ fields in CFEIS. Has extended translation table in which COMPANY_N will contain the same numeric code as CFEIS for this manufacturer. Values defined by Company translation table for this field.
	highway	Yes if mobile source is intended for highway use. No for non-road mobile sources
	model_name	model name
	model_yr	If a prototype, enter representative model year.
	make	Vehicle make e.g. Buick, as distinct from vehicle manufacturer, GM. Legal values defined by MAKE translation table. Values defined by translation table for this field.
	disp_liter	Nominal engine displacement expressed in liters.
	fi_type	Type of fuel injection PFI (port fuel injection) TBI (throttle body injection) INDIR (Indirect injection) DIRECT (direct fuel injection e.g. as for most diesel engines.) Values defined by translation table for this field.
	aspirated	Indicates how engine is aspirated. CHARGED if turbocharged or supercharged. NATURAL if not. Values defined by translation table for this field.
	cylinder	Number of cylinders or rotors.
	cat_type	What type catalyst, if any, is present on the mobile source. Values are: 3WAY Three-way catalyst OX3W Oxidation plus three-way catalyst OXID Oxidation Catalyst NONE No catalyst NULL Unknown Values defined by translation table for this field.
	egr_type	Type of exhaust gas recirculation (EGR). Values defined by translation table. Values defined by translation table for this field.
	engseries	Engine series or product line name.
	cooling	Type of after_cooling. (Legal values defined by translation table.) Values defined by translation table for this field.
	fi_meth	Method of fuel injection. (Legal values defined by translation table.)
	fi_press	Fuel injection pressure, expressed in kPa.
	parttrap	Is particulate trap used? "YES", "NO", or "NUL". Values defined by translation table for this field.

TABLE D-1 (CONT'D). ENTITY NAME/TABLE NAME

Entity Name	Entity Attribute Name	Entity Attribute Definition
	eng_cycle	Engine cycle, 2 = 2-stroke, 4 = 4-stroke, 0 = Unknown. Values defined by translation table for this field.
	ratedpower	Rated horsepower of engine.
	ratedspeed	Rated rpm of engine
	idle_rpm	Idle rpm as declared by the oem.
	proc_odom	Approximate odometer reading in miles at time of vehicle recruitment.
	hour_meter	Hours of operation (usually available only for off-road mobile sources). Null value is 0.
	gvwr	Gross vehicle weight rating in pounds. The value specified by the manufacturer as the loaded weight of a single vehicle.
	pk_torque	Peak torque of the engine expressed in ft-lb.
	pk_t_speed	Peak torque speed expressed in rpm.
	cyl_valves	The number of valves per cylinder.
	stroke	Piston stroke expressed in inches. (not ready to be stored in msod database at this time)
	bore	The diameter of the cylinder expressed in inches.
	inj_ctrl	Code of the Injection control type. Values defined by translation table for this field.
	inj_pcat	Code of the injection equipment/pressure category. Values defined by translation table for this field.
ETEST_AD	p_co	CO emissions. Expressed in grams per bhp-hr.
	p_thc	Total HC emissions. Expressed in grams per bhp-hr.
	p_ch4	Methane emissions. Expressed in grams per bhp-hr.
	total_work	Total work performed in test. Expressed in bhp-hrs.
	p_nox	NO_x emissions. Expressed in grams per bhp-hr.
	p_pm	Total particulate emissions. Expressed in grams per bhp-hr.
	bsfc_meas	Measured brake-specific fuel consumption. Expressed in pounds per bhp-hr.
	study_id	Identification number assigned to the analysis/paper/report of interest.
	fbatch_id	Fuel batch identification.
	test_id	Identification number assigned to the engine test.
	No_modes	Number of test modes involved in this result. Data for individual chassis test modes is stored in the DYNOMODE table; data for individual engine dynomometer test modes is stored in the EMODE table.
	ms_type	General kind of mobile source: 1 = Vehicle 2 = Engine.
	eng_ms_id	Mobile source identifier. For engines, their serial number, probably in conjunction with their manufacturer code.
	test_proc	Identifies the specific test procedure used. Values defined by translation table for this field.

TABLE D-1 (CONT'D). ENTITY NAME/TABLE NAME

Entity Name	Entity Attribute Name	Entity Attribute Definition
FBAT_AD	fbatch_id	Fuel batch identification.
	study_id	Identification number assigned to the analysis/paper/report of interest.
	cetane_num	Cetane number of complete fuel.
	cetane_idx	Cetane index of complete fuel.
	cetane_imp	Amount of cetane improver added, expressed as percentage by volume
	cetane_typ	Type of cetane improver used, e.g. "N" for nitrate type or "P" for peroxide type. Exact set of legal values defined and described by translation table for this field.
	sulfur	Sulfur content, expressed in parts per million.
	nitrogen	Nitrogen content, expressed in parts per million.
	tarom	Total aromatics content of fuel, expressed as a percentage by volume. This is a measured value, as opposed as being calculated as the sum of the monoaromatics and polyaromatics fields.
	marom	Monoaromatics content of fuel, expressed as a percentage by weight. This is a measured value, as opposed as being calculated as the difference of the total aromatics and polyaromatics fields.
	parom	Polyaromatics content of fuel, expressed as a percentage by weight. This is a measured value, as opposed as being calculated as the difference of the total aromatics and monoaromatics fields.
	IBP	Initial boiling point expressed in degrees F.
	T10	10% distillation boiling point, expressed in degrees Fahrenheit.
	T50	50% distillation boiling point, expressed in degrees Fahrenheit.
	T90	90% distillation boiling point, expressed in degrees Fahrenheit.
	T95	95% distillation boiling point, expressed in degrees Fahrenheit.
	EP	End point of distillation curve, expressed in degrees Fahrenheit.
	spec_grav	Specific gravity.
	viscosity	Viscosity, expressed in centistokes @40 degrees C.
	hcratio	Molecular ratio of hydrogen to carbon.
	oxygen	Amount of oxygen in the fuel, expressed as a percentage by weight.
	oxy_type	Type of oxygenate. "NONE" if no oxygenate was added to the base fuel. Values defined by translation table for this field.
	heat	Net heating value of the fuel, expressed in btu/pound.
	ash	Ash content of fuel, expressed as a percentage.
	cetane_dif	This is the difference in cetane number between the described fuel (with additive) and a baseline fuel without additive.
STUDY_AD	study_id	Identification number assigned to the analysis/paper/report of interest.

APPENDIX E

REVIEW OF VECTOR APPROACH TO REGRESSION ANALYSIS

REVIEW OF VECTOR APPROACH TO REGRESSION ANALYSIS

A. Summary of Report

The following comments are based on a review of the paper entitled "A Vector Approach to Regression Analysis and Its Implications to Heavy-Duty Diesel Emissions" which was prepared by H.T. McAdams, R.W. Crawford, and G.R. Hadder, in November 2000 for Oak Ridge National Laboratory. Their approach is based on Principal Component Regression (PCR) analysis (e.g., for more details on PCR see Jackson, J.E. (1991) *A User's Guide to Principal Components*, John Wiley & Sons: New York; or Gunst, R.F. and Mason, R.L. (1980) *Regression Analysis and Its Application: A Data-Oriented Approach*, Marcel Dekker: New York). The PCR procedure has been in existence for over 40 years, and has been thoroughly explored by many different researchers. Generally, its usage is advocated in situations where there are severe collinearities (i.e., linear dependencies) among the predictor variables in a regression analysis.

For simplicity, we will use the symbol PCR to designate a principal component regression. In the PCR procedure a response variable (typically an emissions variable in this application) is regressed against a set of eigenvectors (typically labeled as eigenfuels in the McAdams report). The eigenfuels are obtained from decomposing a correlation matrix, which is constructed using the pairwise correlations existing among the chosen fuel properties. In essence, the eigenfuels represent a transformation of the individual fuel properties to terms consisting of linear combinations containing all the fuel properties. Thus, the fuel-property data space is rotated so that its axes point in the direction of the eigenfuels (i.e., in the direction of the linear combinations of fuel properties) rather than in the direction of the individual fuel properties. Since data will be more dense in some directions than others, it is possible to gain efficiencies by deleting eigenfuels that provide little information about the relationship between the emissions variable and the fuel properties.

The McAdams approach, though lacking a precise protocol in the report, includes the following basic steps.

- Collect a set of emissions data as a function of engine and fuel data
- Assure that the assumptions of a correct model and a normal distribution are valid
- If any assumptions are invalid, consider appropriate data transformations or additional terms in the model (such as nonlinear or interactive terms) and apply as necessary
- Compute the correlation matrix of the fuel properties
- Determine the eigenvectors (i.e., eigenfuels) of the fuel properties
- Regress the emissions variable on the eigenfuels, and, if appropriate, include engine variables in the model
- Delete "inappropriate" eigenfuels from the analysis
- Regress the emissions variable on the remaining eigenfuels
- Delete non-contributing fuel properties
- Compute the correlation matrix of the remaining fuel properties

- Determine the eigenfuels of the remaining fuel properties
- Regress the emissions variable on these eigenfuels
- Repeat the process if there remain any inappropriate eigenfuels or non-contributing fuel properties; otherwise stop.

There are many different issues to consider in completing the above tasks. These include defining criteria for deleting "inappropriate" eigenfuels and "non-contributing" fuel properties; addressing how to incorporate engine variables and non-linear fuel properties in the model; and determining appropriate transformations for the emissions variable. Even after all the steps are completed, McAdams indicates that "further study and exercising of the models are necessary" (i.e., see p.102). These additional steps may entail such procedures as cross-validating the model. All of the steps are discussed in the McAdams report, and various solutions are presented. The report is very thorough and contains a useful data example, which is used to illustrate the various techniques. This presentation helps in the understanding of the application of the methodology.

B. Advantages and Disadvantages of the Methodology

The PCR procedure has many desirable properties. Some of these are listed below.

- If none of the eigenfuels are deleted, a PCR will produce exactly the same regression coefficients for the fuel properties as would a standard regression.
- If severe collinearities exist among the fuel properties, a reduced set of eigenfuels may produce a better solution than standard regression. This occurs because the standard errors of the estimated coefficients, associated with the fuel properties involved in the collinearities, may be substantially reduced.
- The principal component transformation yields uncorrelated and orthogonal eigenfuels, and the sum of squares explained by each is additive. These results facilitate the data analysis. For example, there are no collinearities among the eigenfuels, and one can easily determine the important coefficients.
- If meaningful interpretations can be assigned to the eigenfuels (e.g., if the eigenfuels can be associated with particular diesel blend stocks and/or refinery streams), the resultant regression equation may be easier to understand and utilize.

In contrast to the above advantages, PCR has several drawbacks. Some of these are listed below.

- Deleting eigenfuels can reduce some of the variances of the regression coefficient estimators, but the tradeoff is that bias is introduced to the resulting fuel-property coefficients. The bias increases as more eigenfuels are deleted. This procedure is termed a biased-regression approach in the statistics literature as it is a dimension-reduction technique.

- With increased bias, but reduced variance, one must make tradeoffs to determine the choice of the number of eigenfuels to retain. However, this can be difficult to assess.
- The t-statistics associated with the regression coefficients, which remain after deletion of eigenfuels, are approximate, and thus it is not possible to state significance levels for the results. This may be the reason why the McAdams report stresses use of "non-contributing" fuel properties rather than "non-significant" fuel properties.
- Meaningful interpretation of eigenfuels is not always possible. Thus, it can be difficult to interpret the structure of the regression relationship.

There are many issues associated with conducting a PCR, and not all of them are addressed in the McAdams report. For example, the report does not discuss the issue of outliers and data anomalies. However, in many PCR references, suggestions are made to plot the eigenfuels against one another in scatter plots in order to isolate extreme observations. One also could use residual analysis with the final model similar to what is done in a standard regression analysis.

The McAdams report discusses the complexities associated with transforming the data and adding nonlinear terms to the model. The suggested procedures for solving these problems are brief and not as well developed as the rest of the report. In these situations, the report indicates (on p.164) that it “could well be that the most valuable contribution PCR can make is as a variable screening procedure in a manner not unlike existing ‘screening designs’, in which the primary purpose of the model is to identify predictor variables but not necessarily how they interact in their effects on the response variable”.

Engine effects are briefly addressed in the McAdams report but only for the situation where the engines represent fixed effects and are represented by categorical variables. More complicated situations can be encountered in practice where engines are often represented as random effects, and where actual engine properties are included in the model. The procedure possibly could be extended to these situations, but the steps to follow are not developed.

The McAdams report includes printed MatLab software code to run this procedure in its most basic form. One alternative is to use SAS with its PROC PRINCOMP and PROC REG procedures. However, additional SAS code would be required in this latter approach in order to obtain several of the components needed in the final model.

C. Appropriateness of Inclusion of This Approach

SwRI was tasked with assessing the appropriateness of the inclusion of the McAdams PCR strategy in its model development effort. It would seem that the usefulness of the procedure hinges on the degree of collinearity that is present among the fuel properties. The more severe the collinearities, the more there is to be gained by using PCR. The procedure might also serve best by using it as a screening tool for the fuel properties to be included in the standard regression model. Thus, there may be situations where the method is appropriate as well as situations where it would not be appropriate.

For example, if engine effects are to be included in the model and treated as random components, it might not be possible to use this procedure due to the complexities of the model. However, if engines are treated as fixed effects, then it might be possible to use this approach as a screening method to help determine the useful fuel properties. It also might be of interest to separately run this procedure on the fuel terms to determine which ones are important. In either of the latter two cases, we could include the chosen fuel properties in a standard regression analysis.

A suggested strategy would be to use standard regression procedures to model emissions, but to include a PCR analysis to help screen for the fuel effects. Since stepwise regression often produces many viable options as models, the results from the PCR analysis may help select the final model and lend substance to its selection. More applications of this procedure are needed to verify the usefulness of this suggestion.

APPENDIX F

MIXED MODEL METHODOLOGY

MIXED MODEL METHODOLOGY

The mixed-effects model approach chosen for this study is based on fitting a model having both fixed effects and random effects. Fuel properties have been designated as fixed effects because, in the source documents from which they were determined for the database, the individual fuel values were usually controlled, to some degree, by the experimenter. Thus, these properties represented the whole population of possible values of interest. In contrast, the engines within a tech group have been defined as random effects as their values represent a sample of engines from a large population of possible engines. Although the engines were not necessarily selected in a probabilistic manner, as would be required to have a random sample, the group of engines used in this project have been treated as a random sample since they are representative of the larger group of engines from which they were sampled.

A mixed model is given as

$$Y = X\beta + Zu + e$$

where Y represents the response variable of interest, Xβ represents the fixed effects, Zu represents the random effects, and e represents the error terms. The X represents the known design matrix for the fixed effects (i.e., the fuel terms), and β represents the corresponding unknown coefficients that are to be estimated. The Z represents the known design matrix for the random effects (i.e., the engine terms and the engine-by-fuel interaction terms), and u represents the unknown random-effects coefficients. The u values are assumed to have a normal distribution with a mean of zero and an unknown covariance structure. The e represents the unknown random error terms and their values are assumed to have a normal distribution with a zero mean and an unknown covariance structure. Also, the error terms are assumed to be uncorrelated with the random effects. With these assumptions, Y is assumed to have a normal distribution with a mean of Xβ and a covariance that is a function of the covariances of u and e.

In a mixed model, there is interest in estimating the unknown coefficients, but there also is interest in estimating the unknown variance components associated with u and e. Due to these variance components, a least-squares procedure is not best for use in estimation. An alternative is to use maximum-likelihood-based methods. Maximum likelihood estimates of the unknown coefficients are obtained by maximizing the likelihood function over the parameter space. This is based on an iterative process and therefore the estimation procedure must be repeated several times before it converges to a solution. The likelihood-based solutions for the coefficients, β and u, in this project are based on estimates of the covariance matrices for u and e. Because of this, the estimate of β is labeled the empirical best linear unbiased estimate, and the predictor of u is labeled the empirical best linear unbiased predictor.

Due to the complexity of the analysis for a mixed model, the computations require a computer program. The one used in this project was the PROC MIXED procedure available in the SAS computing software (version 8.01 on a PC platform). Several options are available in PROC MIXED and those implemented in this analysis are listed below.

- The maximum number of iterations for convergence was set to 500.
- The residual (restricted) maximum likelihood was used to estimate the covariance parameters.
- The unique engine-study combinations were defined as the class variable.
- The residual degrees of freedom were used for the tests of the fixed effects.
- The solution for the fixed-effects parameters was used to obtain the coefficient estimates.
- The random engine-by-fuel terms were nested within the engine-study variable.

Several statistical measures for model comparisons are provided in the SAS program. Those used in this project include Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). The AIC measure estimates the discrepancy between the distribution that generated the data and the model that approximates it. It is considered a criterion for the best model fit, and is used when the objective is to find the best approximating model. The AIC can be used to compare models with the same fixed effects, but different variance structures. Models with the smallest AIC values are preferred. In contrast, the BIC measure provides a consistent estimator of the true order of the model at the expense of assuming that a true model exists and is low-dimensional. Models with smaller BIC values are preferred, but BIC penalizes models with increased numbers of covariance parameters more than does the AIC, and the two may not agree as to which covariance model is best.

More details on mixed models can be found in the following reference:

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). SAS System for Linear Models, SAS Institute Inc: Cary, North Carolina.

APPENDIX G

TIME-DRIFT CORRECTION EQUATIONS

TIME-DRIFT CORRECTION EQUATIONS

Four studies used in the database provided time-adjusted emissions data or the equations to convert to time-adjusted data. This appendix summarizes the equations used in the studies.

A. SAE2000-01-2890

This paper did not provide the correction equations for emissions, but did provide the time-adjusted values for NO_x and PM on the 95CAT 3404E engine and time-adjusted values for NO_x on the 96 Series 50 engine.

B. CAPE32-80, VE-1 (Phase I)

A regression equation of the form:

$$\text{Emissions}_{\text{Predicted}}\ (\text{g/bhp-hr}) = a \times (\text{Time}) + b$$

where a and b are estimated regression coefficients was developed based on tests of a single fuel at various times throughout the test program. Time is measured in engine hours. Using these results, emissions measurements on other fuels were corrected for time-drift corrections using the following equation:

$$(\text{Emissions})_{\text{corrected}} = (\text{Emissions})_{\text{measured}} \times [b/(a \times (\text{Time}) + b)]$$

The following table lists the coefficients for the transient composite test for the three engines.

	NOx		PM		HC		CO	
Engine	a	b	a	b	a	b	a	b
NTCC 400	9.721E-4	4.5698	5.175E-4	0.49781	-4.386E-4	0.60770	6.0697E-4	2.3140
DDC 60	-1.818E-3	4.9213	5.024E-5	0.29499	1.990E-4	0.14058	-1.963E-4	2.3376
NIC 7.3	-1.821E-3	4.4717	-2.922E-4	0.30601	-6.038E-6	0.37103	-8.444E-4	1.4024

C. CRCVE-1 (Phase II)

A regression equation of the form:

$$\text{Emissions}_{\text{Predicted}}\ (\text{g/bhp-hr}) = a \times (\text{Time}) + b$$

where a and b are estimated regression coefficients was developed based on tests of a single fuel at various times throughout the test program. Time is measured in engine hours. Using these results, emissions measurements on other fuels were corrected for time-drift corrections using the following equation:

$$(\text{Emissions})_{\text{corrected}} = (\text{Emissions})_{\text{measured}} \times [b/(a \times (\text{Time}) + b)]$$

The following table lists the coefficients for the transient composite FTP test for the single engine.

Engine	NOx		PM		HC		CO	
	a	b	a	b	a	b	a	b
DDC 60	0.00749	4.5812	-0.0002416	0.1871	-0.002748	0.4808	-0.003322	2.0369

D. CRCVE-10

A regression equation of the form:

$$\text{Emissions}_{\text{Predicted}}\ (\text{g/bhp-hr}) = a \times (\text{Time}) + b$$

where a and b are estimated regression coefficients was developed based on tests of a single fuel at various times throughout the test program. Time is measured in engine hours. Using these results, emissions measurements on other fuels were corrected for time-drift corrections using the following equation:

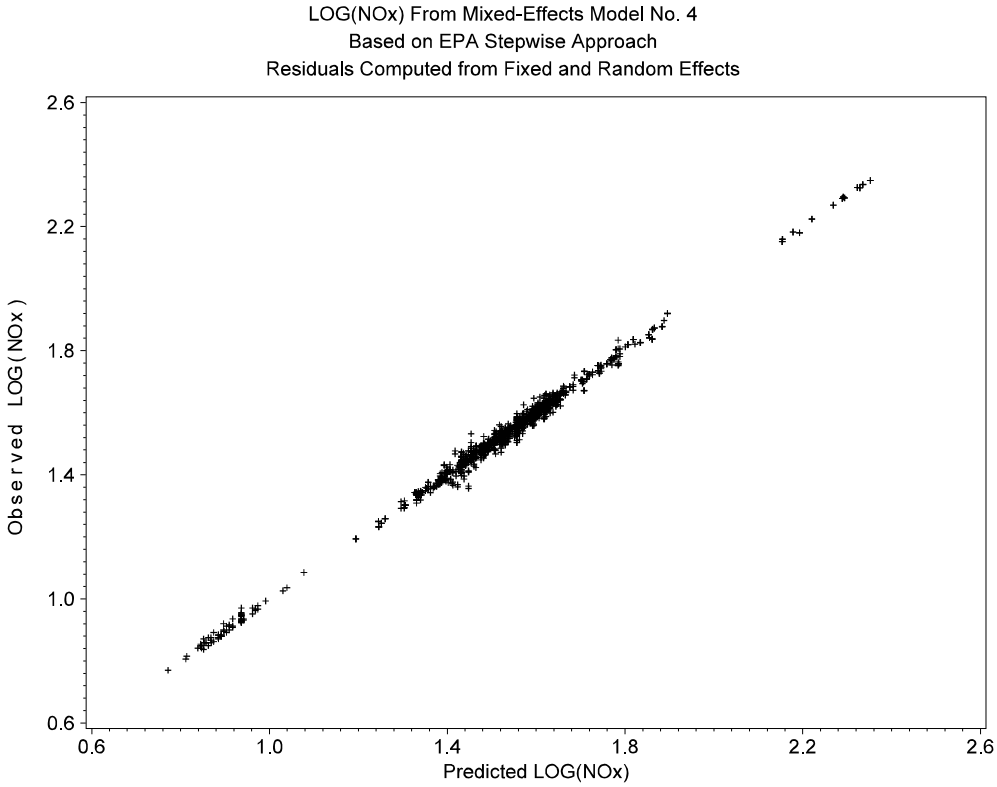
$$(\text{Emissions})_{\text{corrected}} = (\text{Emissions})_{\text{measured}} \times [b/(a \times (\text{Time}) + b)]$$

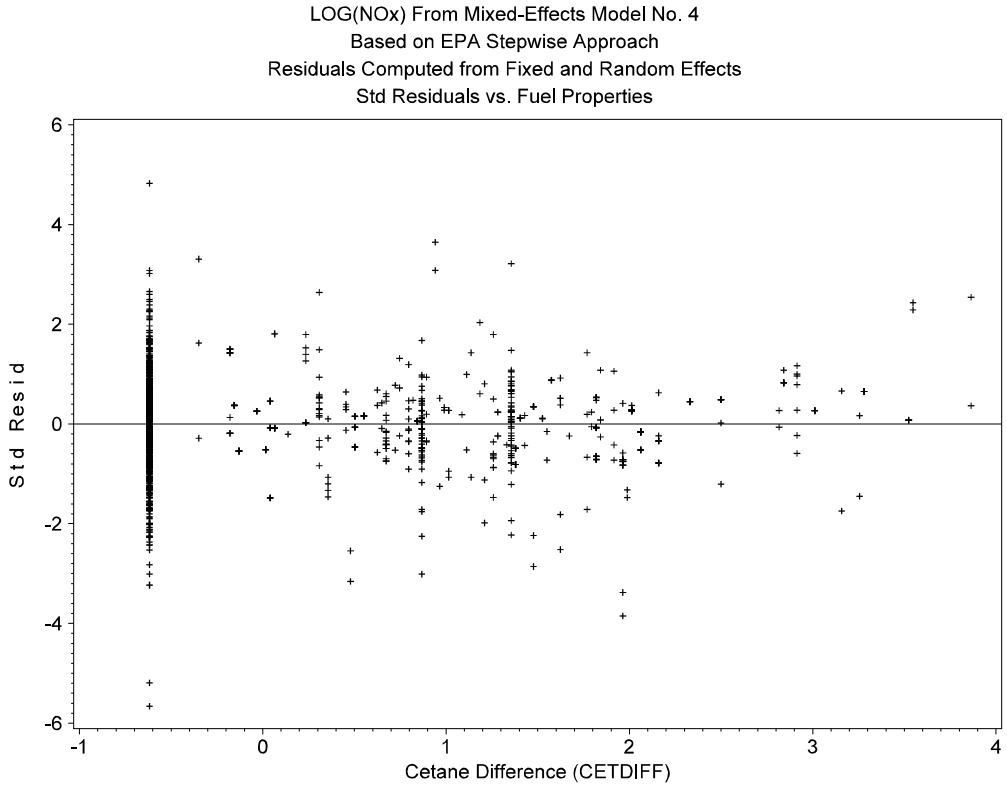
The following table lists the coefficients for the FTP composite test for the five engines.

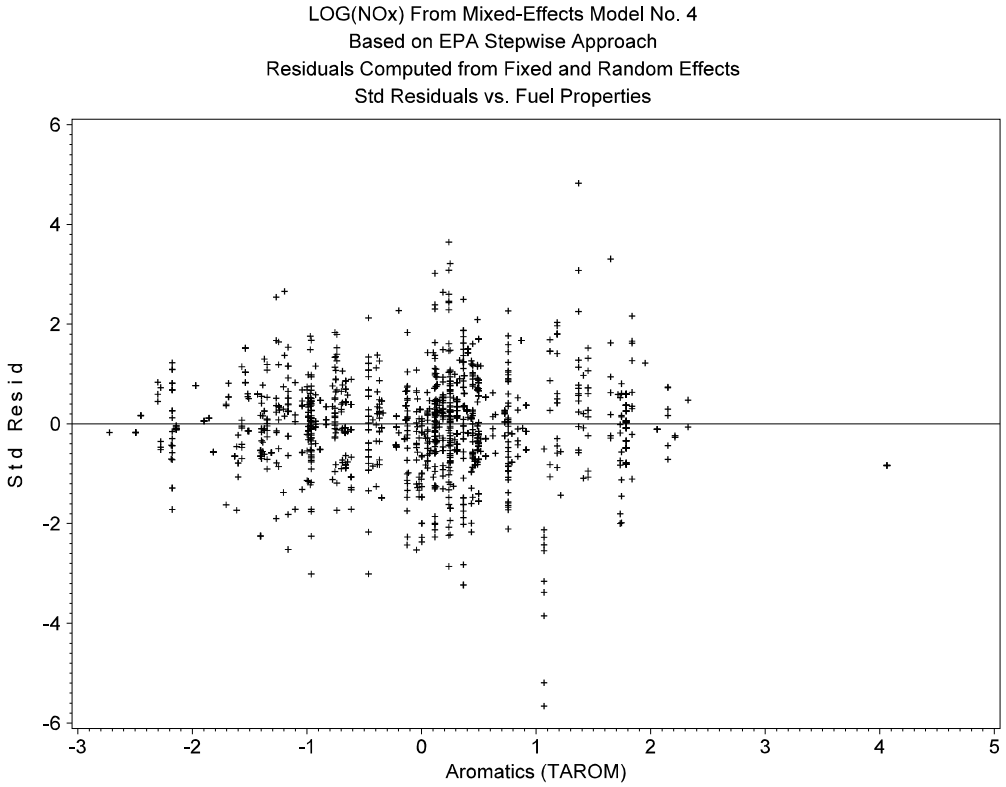
Engine	NOx		PM		HC		CO	
	a	b	a	b	a	b	a	b
VE_10_1	5.836E-4	4.813	-9.413E-5	0.1166	1.773E-4	0.0627	-5.090E-4	1.3817
VE_10_2	n/a	n/a	n/a	n/a	4.439E-4	0.0488	-7.673E-4	1.3875
VE_10_3	1.931E-3	4.528	-1.240E-4	0.1009	n/a	n/a	1.861E-3	0.5623
VE_10_4	2.155E-3	3.818	n/a	n/a	n/a	n/a	n/a	n/a
VE_10_5	n/a	n/a	-9.874E-5	0.0889	-5.761E-4	0.1420	n/a	n/a

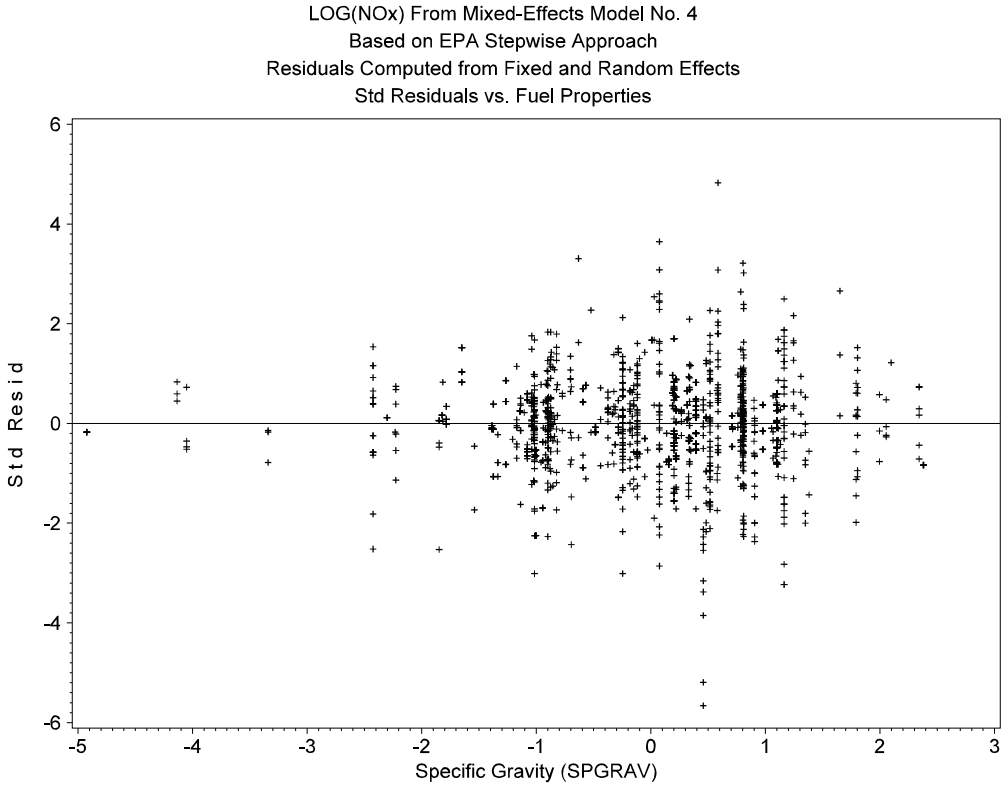
APPENDIX H

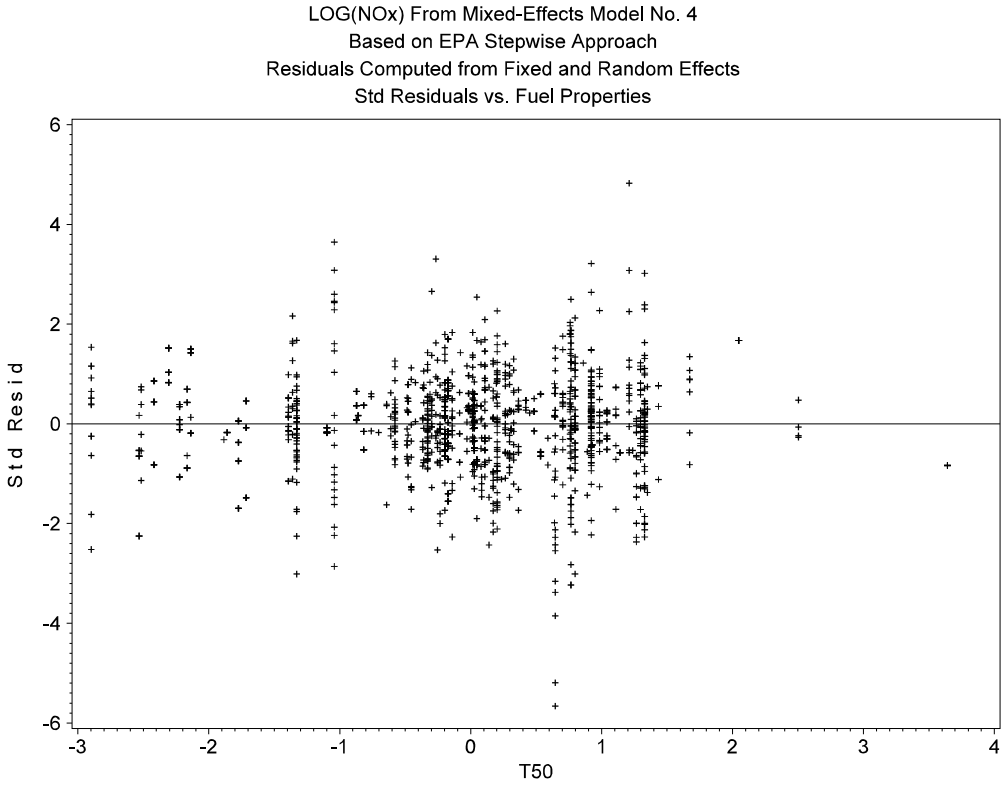
RESIDUAL PLOTS

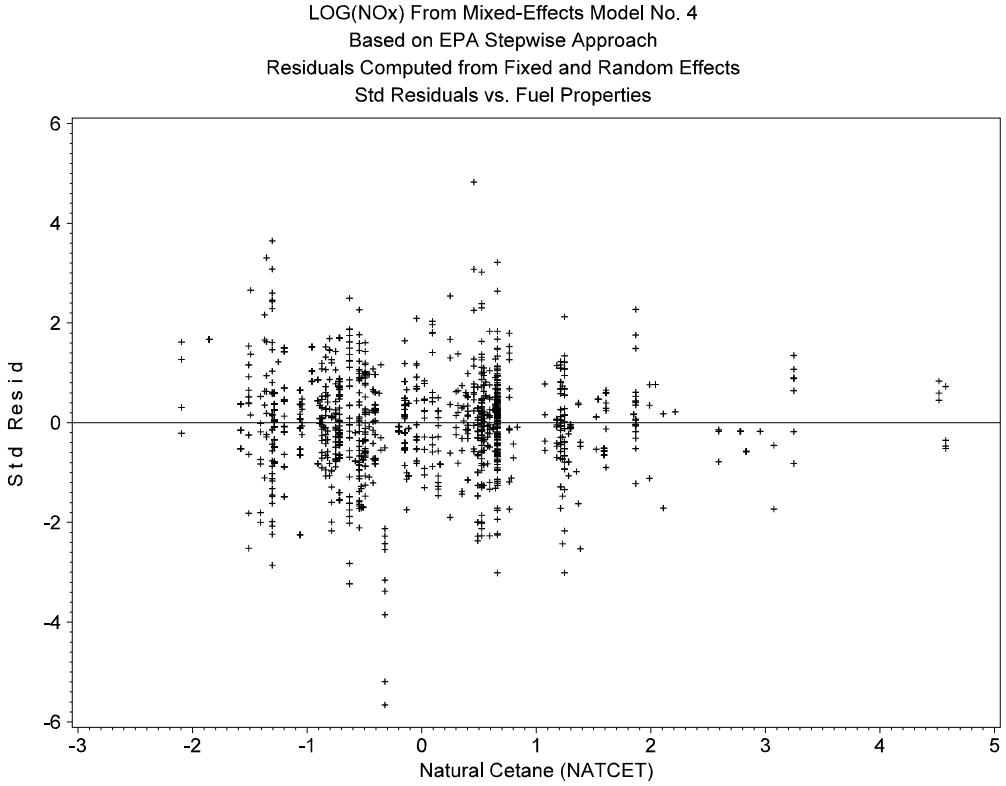

LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Observed LOG(NOx)
Predicted LOG(NOx)
2.6
2.2
1.8
1.4
1.0
0.6
0.6
1.0
1.4
1.8
2.2
2.6


LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
Cetane Difference (CETDIFF)
6
4
2
0
-2
-4
-6
-1
0
1
2
3
4

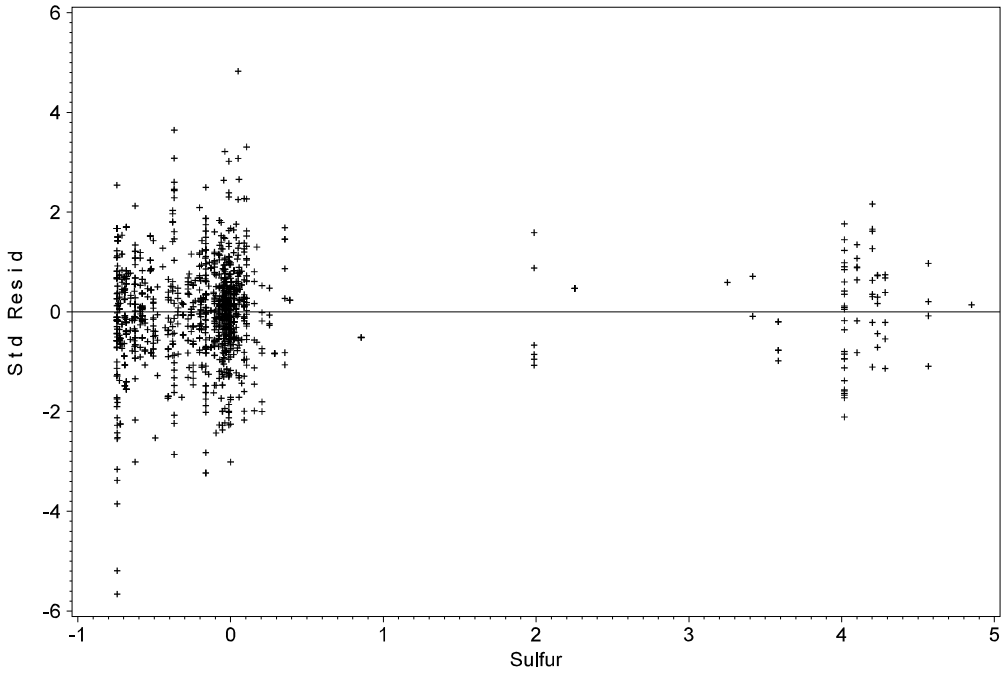

LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
4
5
Aromatics (TAROM)


LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
6
4
2
0
-2
-4
-6
-5
-4
-3
-2
-1
0
1
2
3
Specific Gravity (SPGRAV)

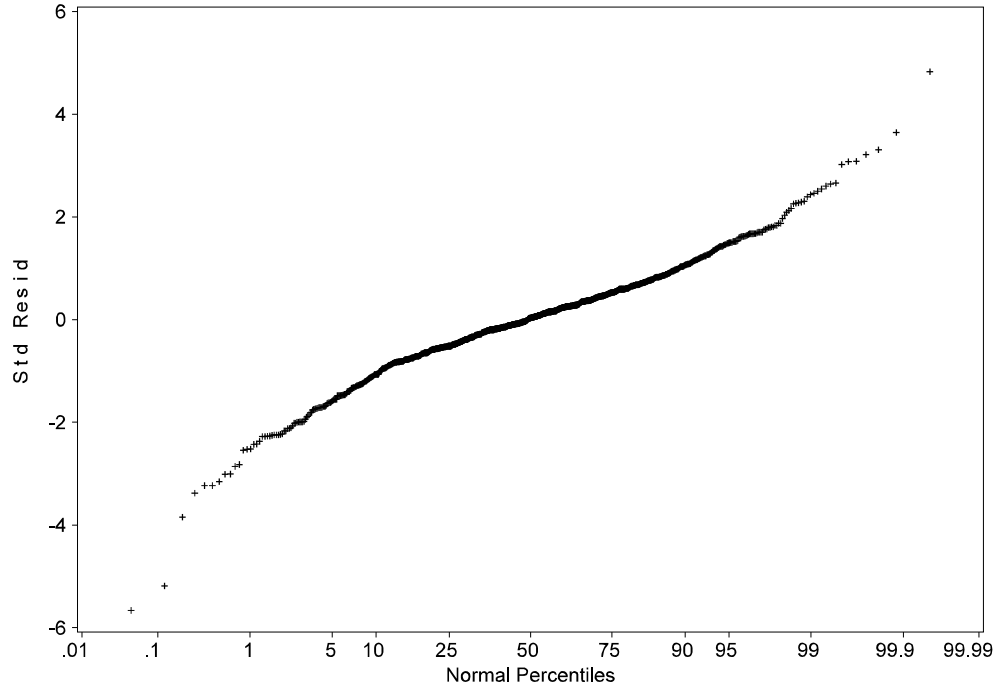

LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
4
T50

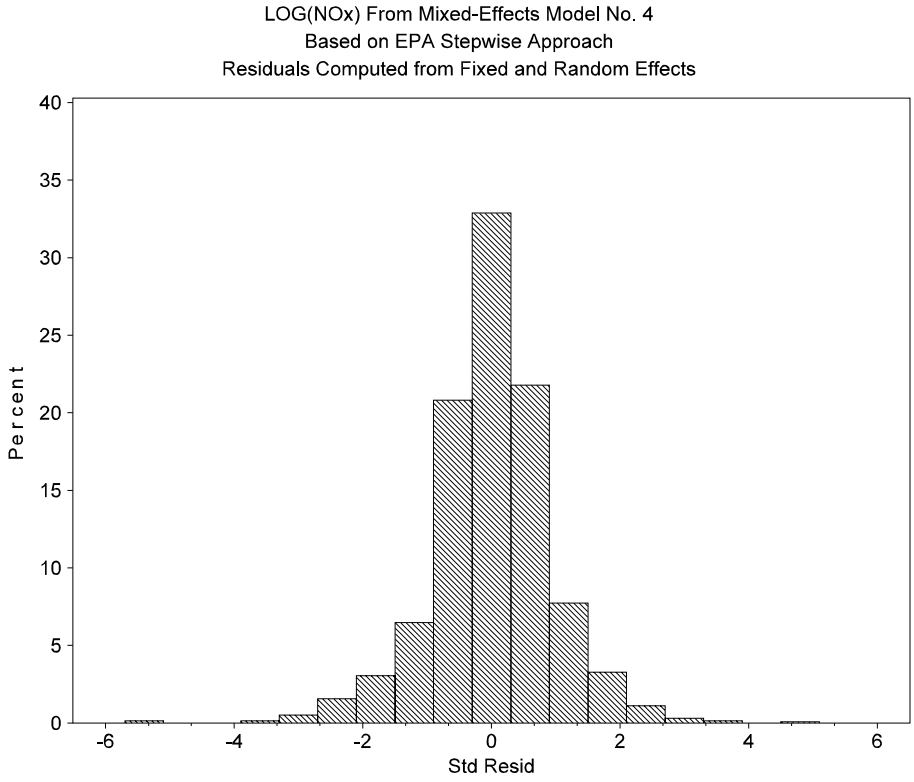

LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
4
5
Natural Cetane (NATCET)

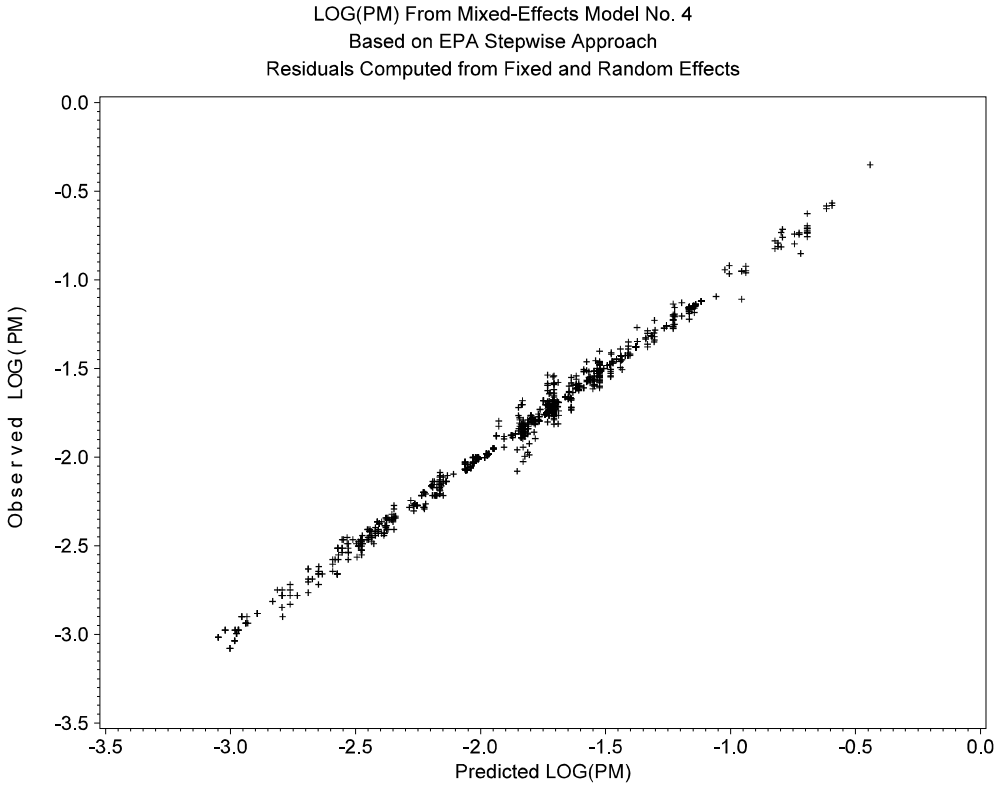
LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties

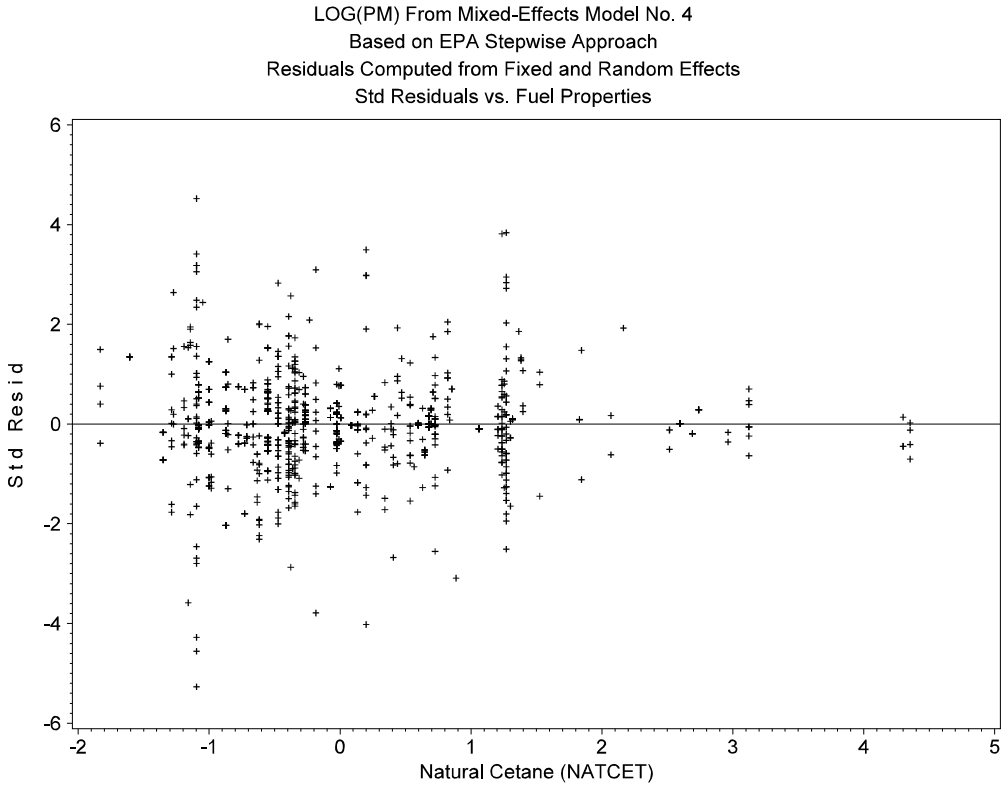


LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects

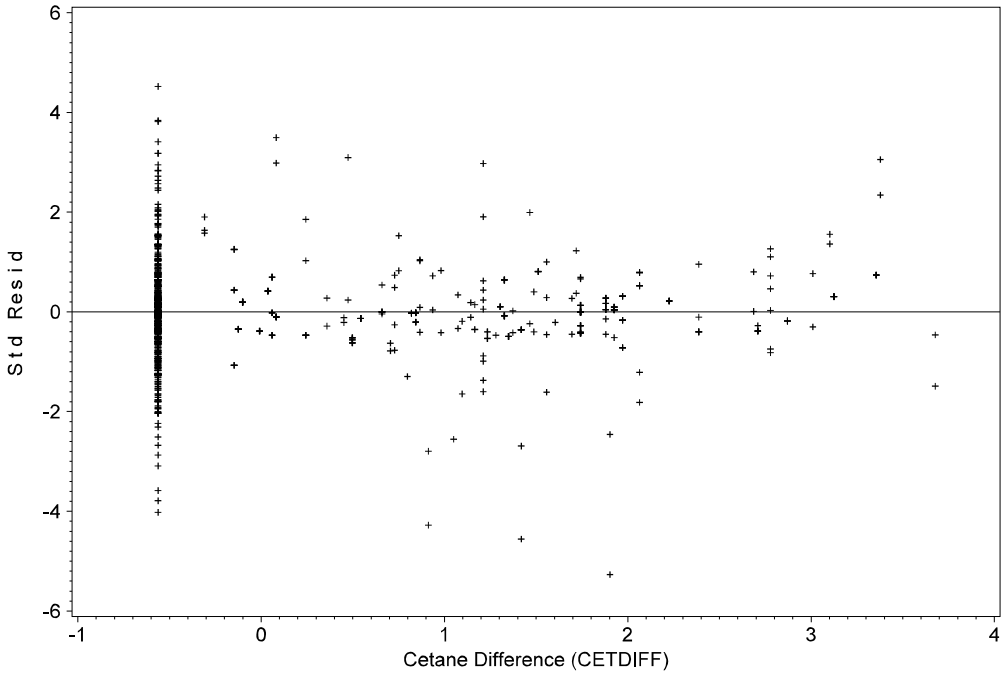



LOG(NOx) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Percent
40
35
30
25
20
15
10
5
0
-6
-4
-2
0
2
4
6
Std Resid

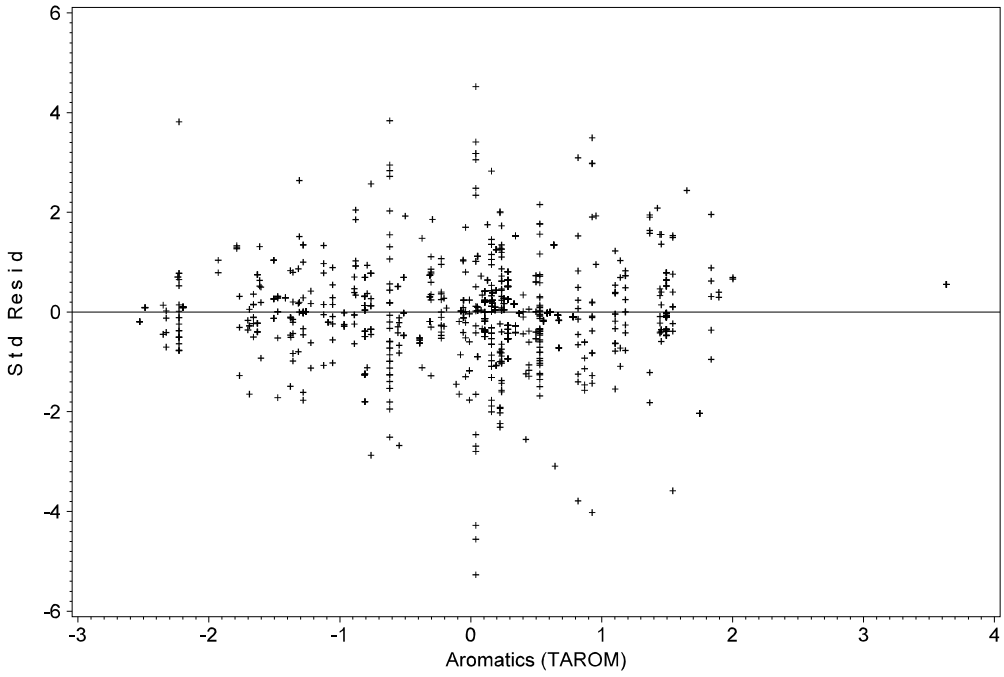

LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Observed LOG(PM)
Predicted LOG(PM)
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5


LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
Natural Cetane (NATCET)
6
4
2
0
-2
-4
-6
-2
-1
0
1
2
3
4
5

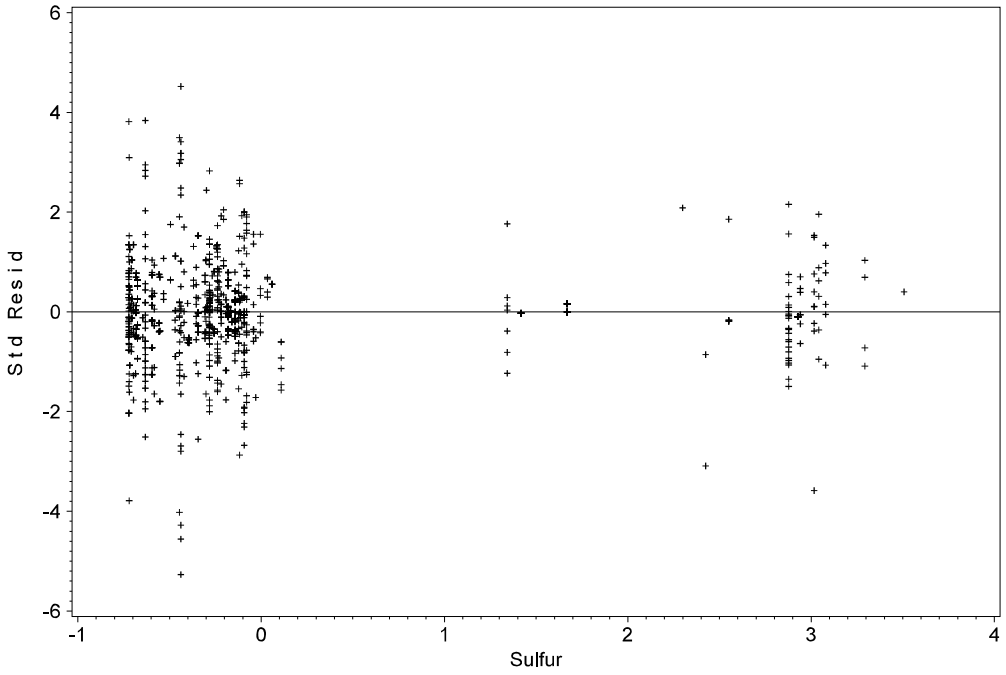
LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties



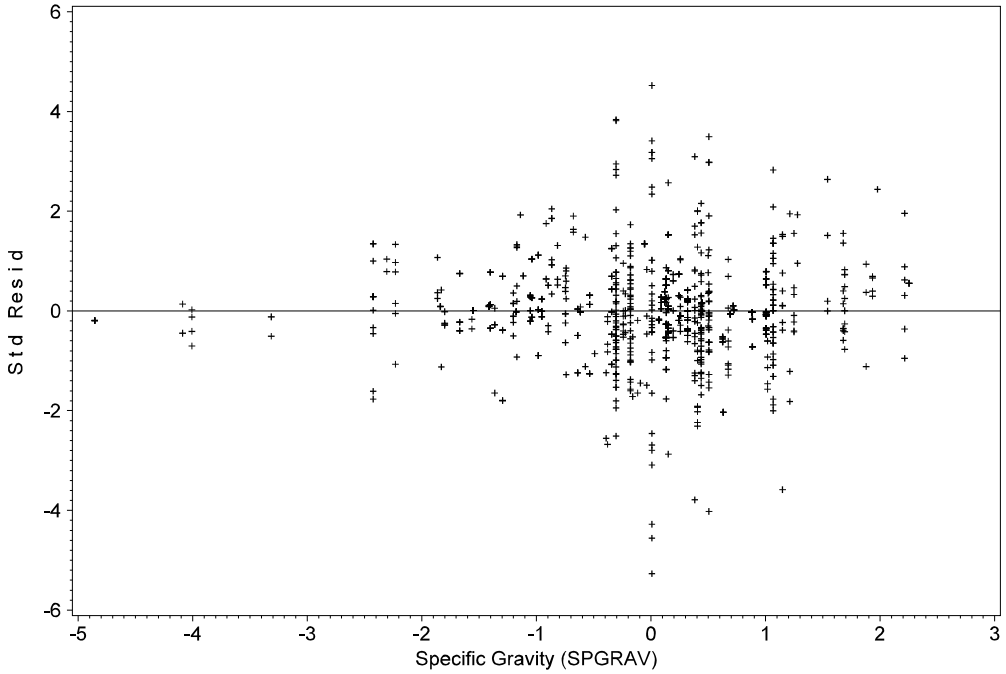
LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties



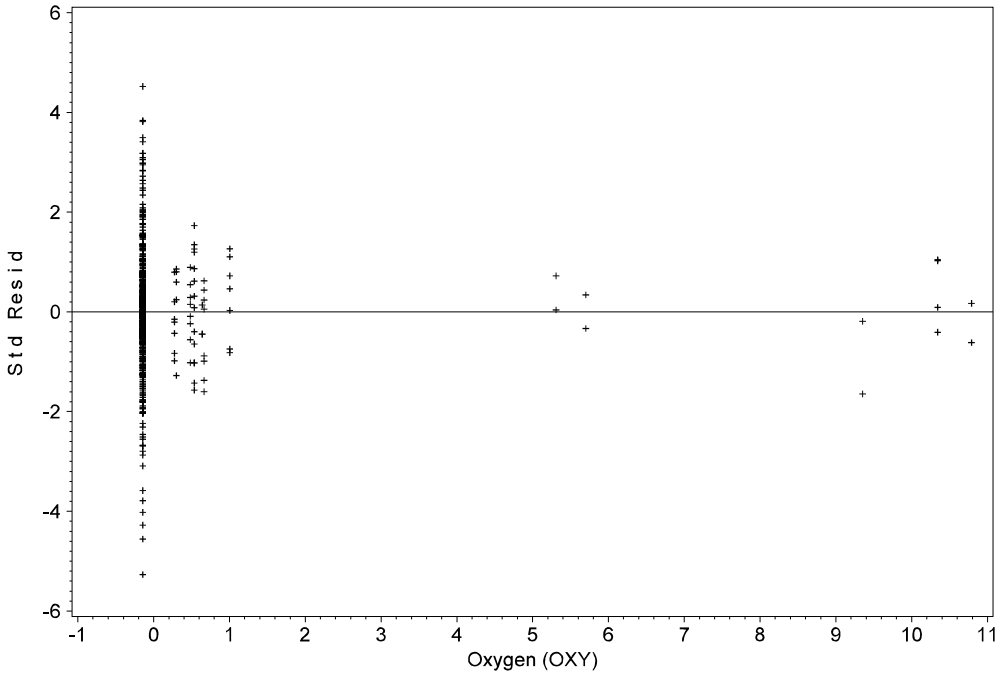
LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties



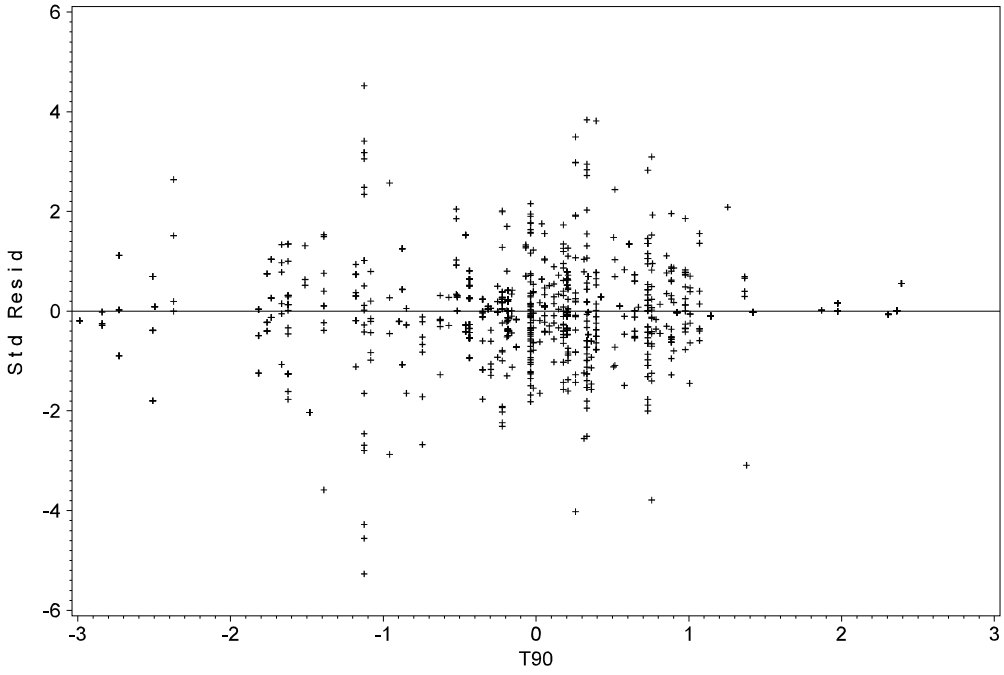
LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties

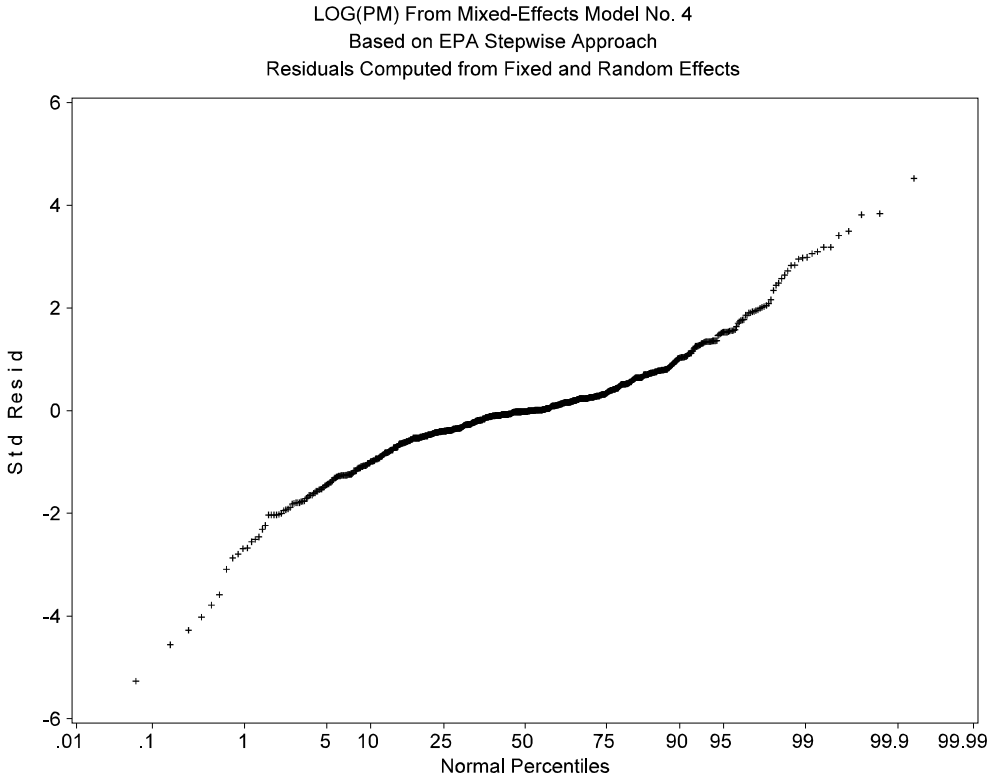


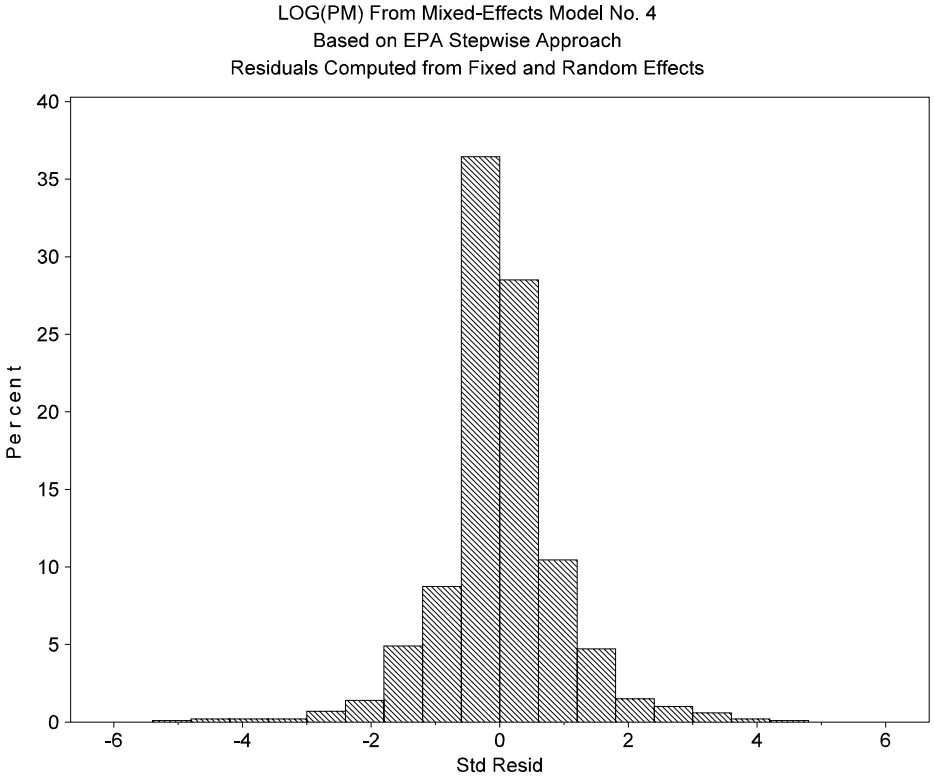
LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties

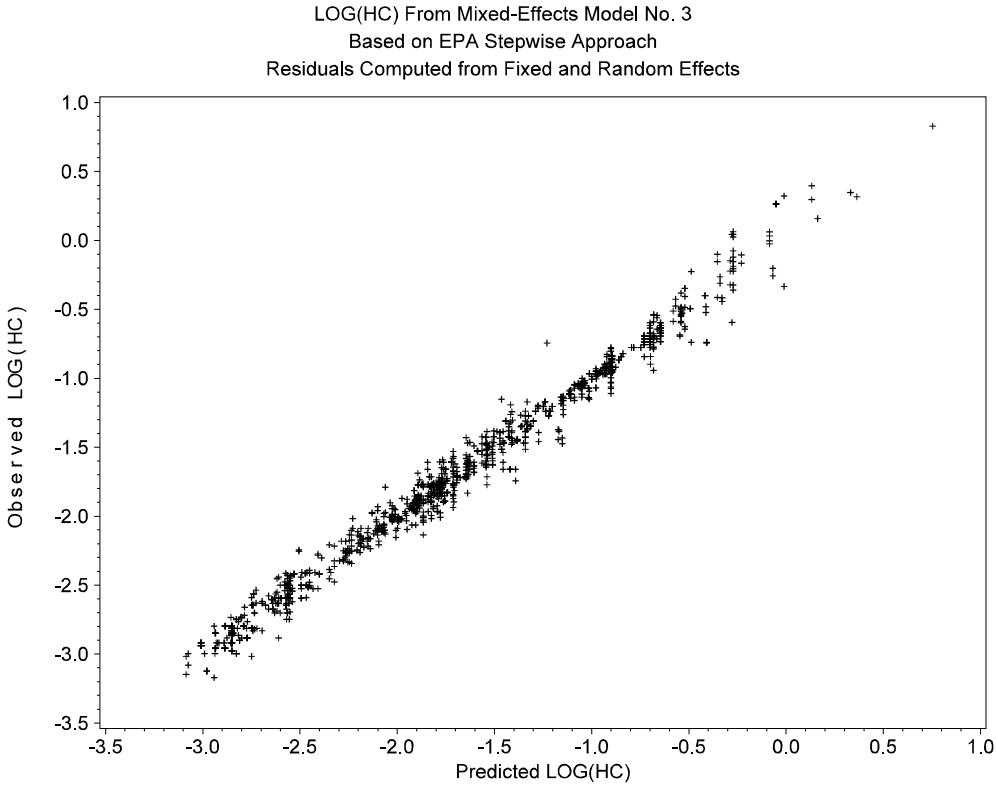


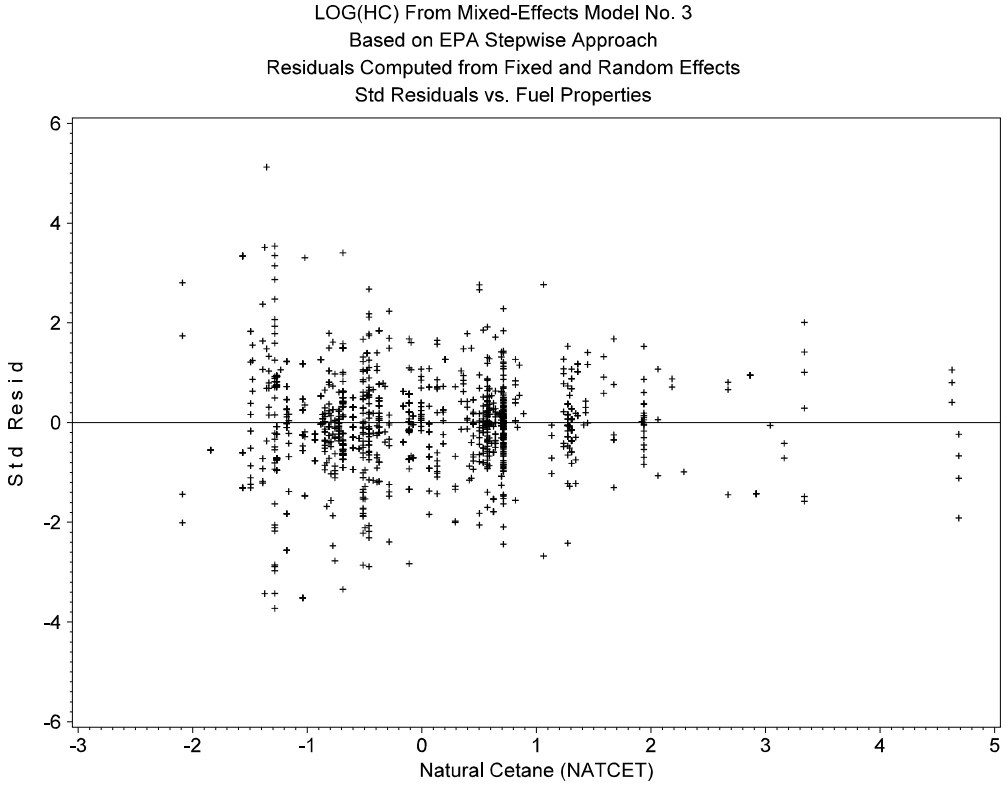
LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties



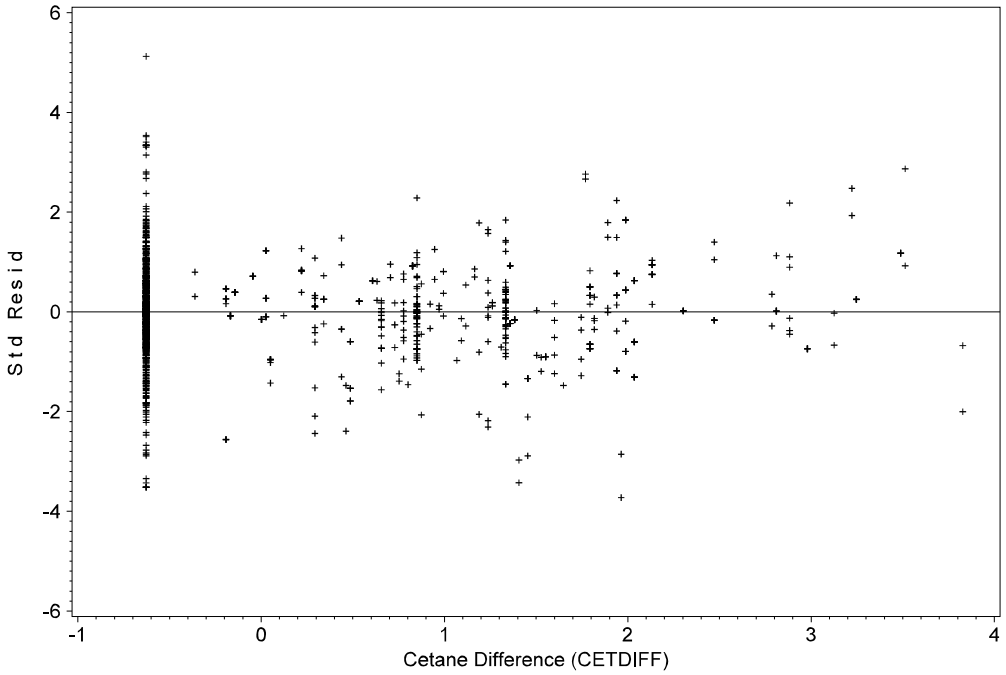

LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Resid
6
4
2
0
-2
-4
-6
.01
.1
1
5
10
25
50
75
90
95
99
99.9
99.99
Normal Percentiles


LOG(PM) From Mixed-Effects Model No. 4
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Percent
40
35
30
25
20
15
10
5
0
-6
-4
-2
0
2
4
6
Std Resid

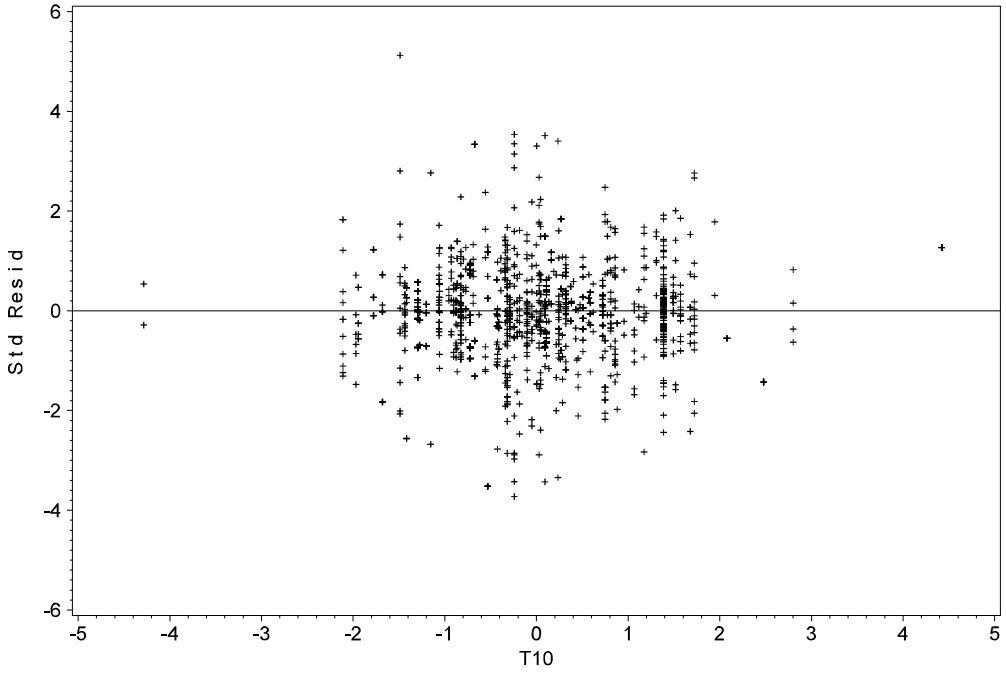

LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Observed LOG(HC)
Predicted LOG(HC)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

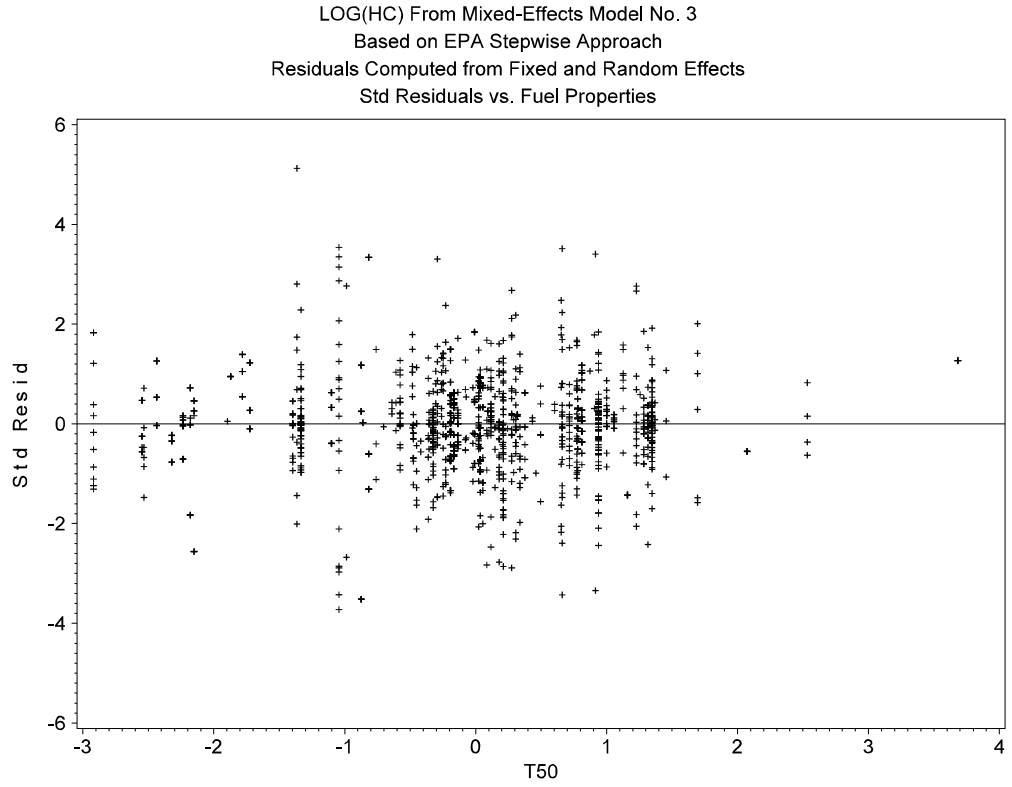

LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
Natural Cetane (NATCET)
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
4
5

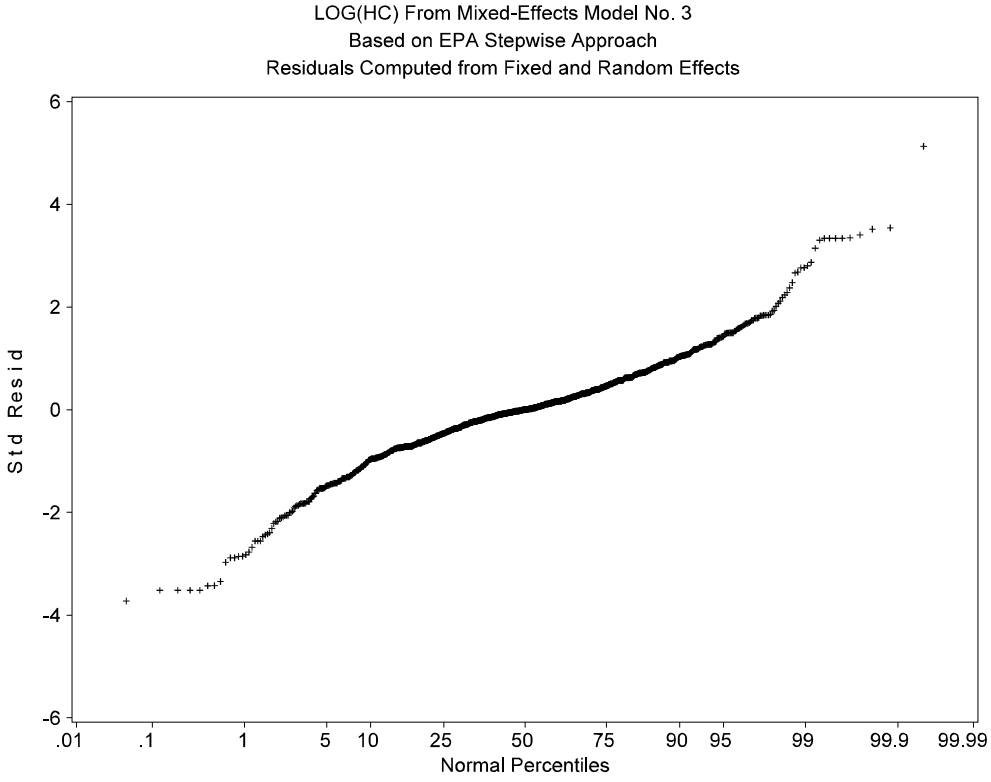
LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties

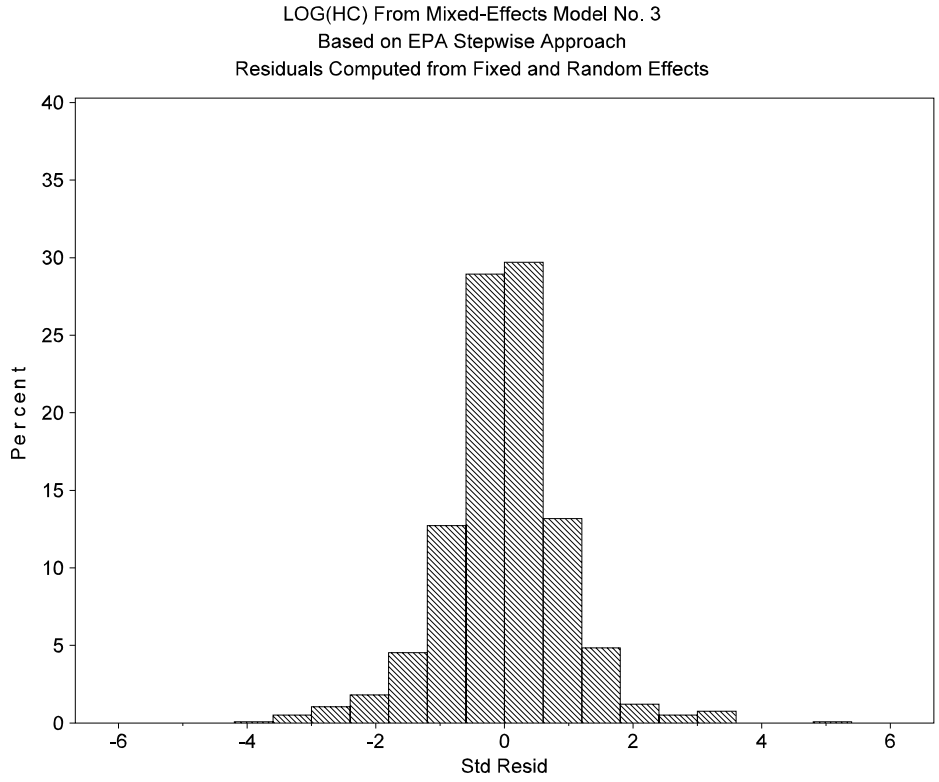


LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties



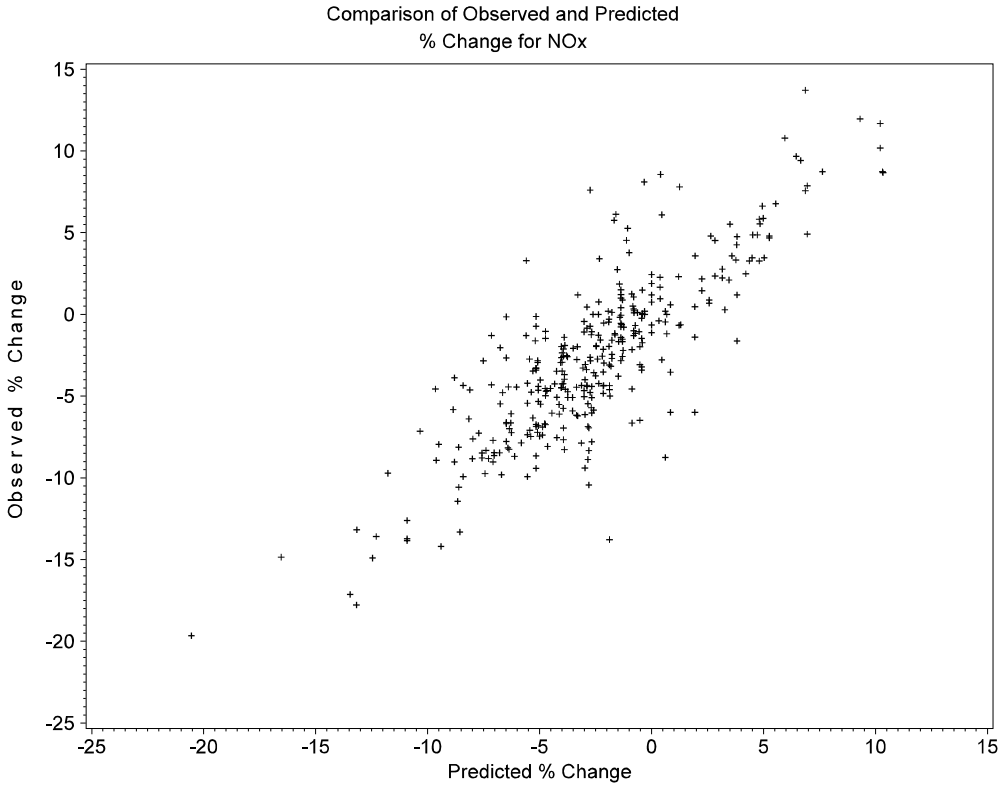

LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Residuals vs. Fuel Properties
Std Resid
6
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3
4
T50

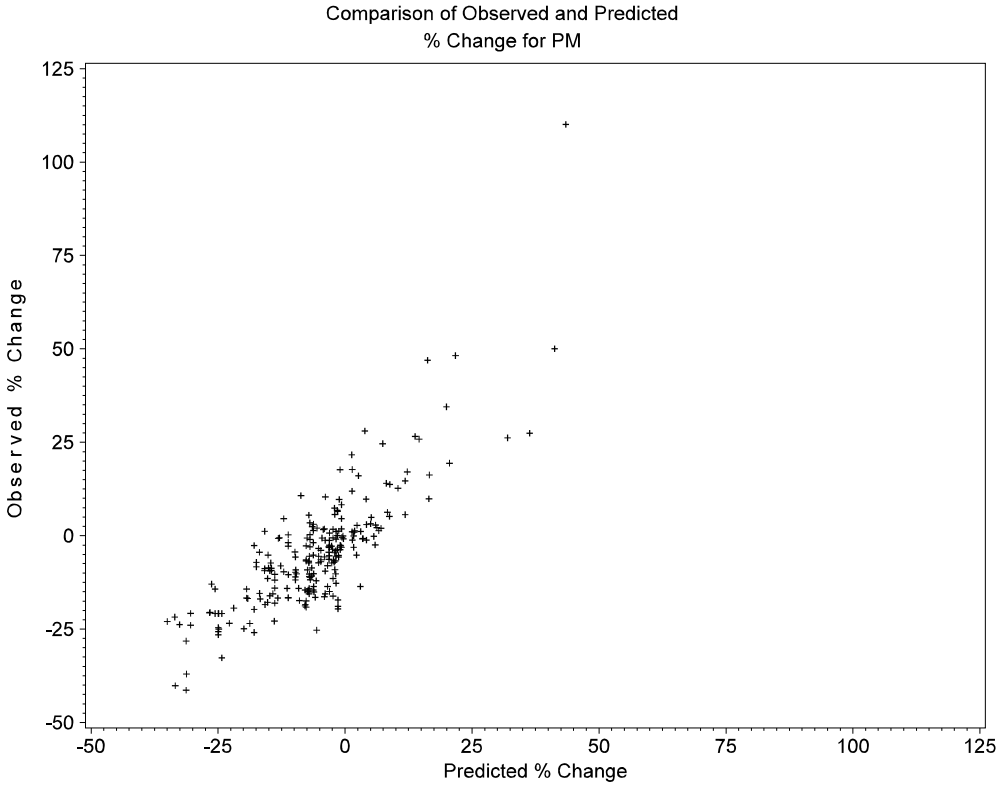

LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Std Resid
6
4
2
0
-2
-4
-6
.01
.1
1
5
10
25
50
75
90
95
99
99.9
99.99
Normal Percentiles

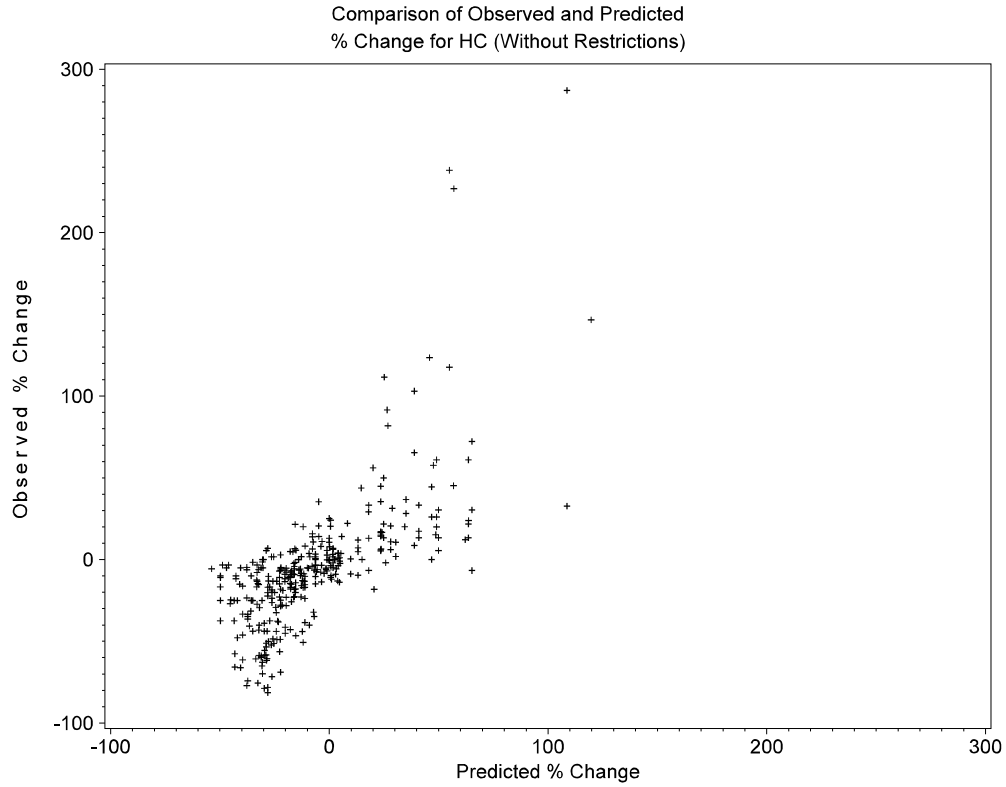

LOG(HC) From Mixed-Effects Model No. 3
Based on EPA Stepwise Approach
Residuals Computed from Fixed and Random Effects
Percent
40
35
30
25
20
15
10
5
0
-6
-4
-2
0
2
4
6
Std Resid

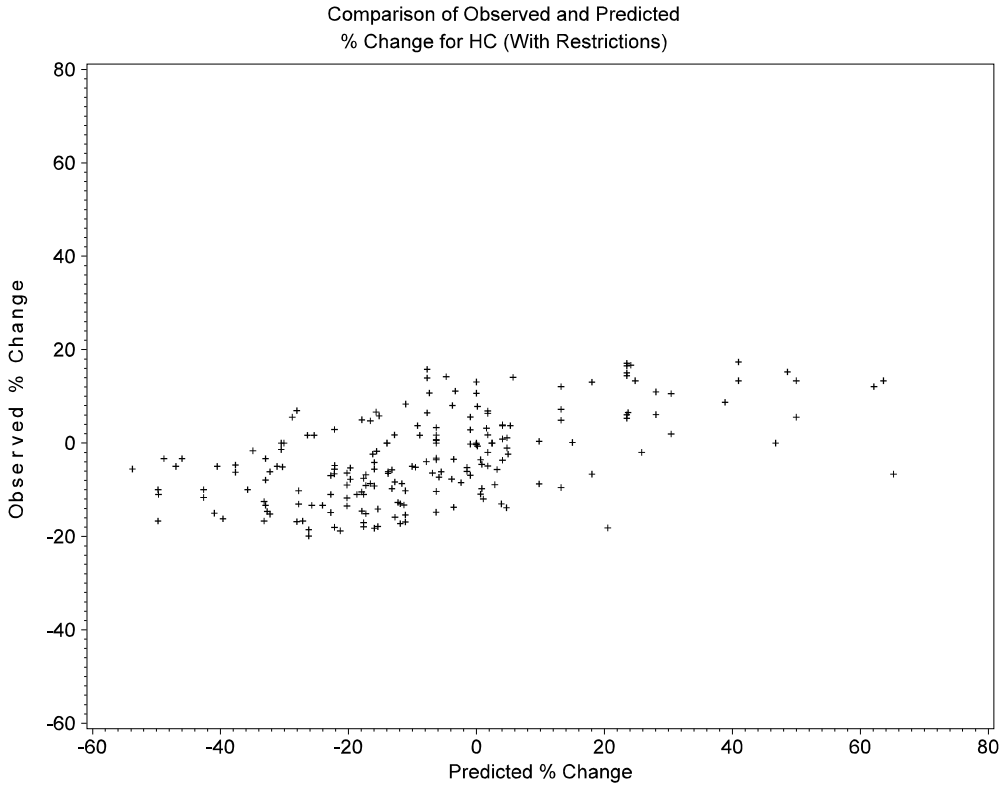
APPENDIX I

SCATTER PLOTS OF PERCENT CHANGE


Comparison of Observed and Predicted
% Change for NOx
Observed % Change
Predicted % Change
15
10
5
0
-5
-10
-15
-20
-25
-25
-20
-15
-10
-5
0
5
10
15


Comparison of Observed and Predicted
% Change for PM
Observed % Change
Predicted % Change
125
100
75
50
25
0
-25
-50
-50
-25
0
25
50
75
100
125


Comparison of Observed and Predicted
% Change for HC (Without Restrictions)
Observed % Change
Predicted % Change
300
200
100
0
-100
-100
0
100
200
300


Comparison of Observed and Predicted
% Change for HC (With Restrictions)
Observed % Change
Predicted % Change
80
60
40
20
0
-20
-40
-60
-60
-40
-20
0
20
40
60
80