Appendix J
Exhaust Emissions Models – Description and Derivation

A. BACKGROUND

The California predictive model (exhaust module), adopted by the Board in May 1995, has been enhanced in two ways. The database has been expanded by about 2000 data points. The new model expands the number of vehicle classes, adding Tech Class 5, vehicles model year (MY) 1994 and later,.

This Appendix describes the procedures used to develop the model for two pollutants, hydrocarbons (THC) and oxides of nitrogen (NOx). The development of the potency-weighted toxics model is the same as described in the ARB report, "Proposed Amendments to the California Phase 2 Reformulated Gasoline Regulations, Included Amendments providing for a Predictive Model."

B. NEW DATA

The available of new tests data have facilitated enhancements to the existing model. Table 1 lists the sources of new data, including number of observations, number of vehicles, and fuel properties tested. There are seven fuel properties of interest: aromatic hydrocarbons (ARO), olefins (OL), oxygen (OX), Reid Vapor Pressure (RVP), sulfur (SU), and 50% as well as 90% distillation temperatures (T50 and T90).

Table 1

Summary of New Data Added

Study	Fuel Properties Tested	Tech Class*	# Vehicles	# Observations
1. AOB17&18	All	3, 4, 5	29	674
2. ARBATLP2	All	3, 4	8	48
3. ARBMSD96	All	3, 4	10	30
4. EPA_ATL2	All	4	40	741
5. EPA_PH3	All	4	19	190
6. CHEVOX99	Oxygen	4	10	32
7. ARBETOH	Oxygen, RVP	4, 5	12	56
8. AAMALOSU	Sulfur	5	21	253
9. CRCLOSUL	Sulfur	5	24	356
10. (Forthcoming)	Sulfur, Oxygen	5	?	?

*See Table 2 for definition

In addition to more than 7,000 data points in the current database, about 2,400 data points have been added. The new data are mostly from vehicles in Tech Class 4 and 5. Unlike the current model, the enhanced model categorized Tech 4 differently. Table 2 presents the vehicle technology group classifications by model year.

Table 2

Vehicle Classification by Model Year

Vehicle Class	Current Model	New Model
Tech 3	MY 1981-1985	MY 1981-1985
Tech 4	MY 1986-1995	MY 1986-1993
Tech 5	N/A	MY 1994 and newer

C. STATISTICAL MODELS:

The main objective of statistical modeling approach here is to find a relationship between emissions (dependent variables) and fuel properties (independent variables) based on vehicle technology groups. A linear model for each pollutant and technology group can be written as follows:

$$\mathbf{y}_{p,t} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + ... + \beta_n\mathbf{x}_n + \boldsymbol{\varepsilon}$$

where

$\mathbf{y}_{p,t}$ (vector) = measured emission for pollutant, p, from vehicles in tech group, t
β_i = parameter to be estimated from the data
$\mathbf{x}_i$ (vector) = fuel property
$\boldsymbol{\varepsilon}$ (vector) = error term

The term 'linear' stems from the fact that the dependent variable is linearly related to fuel properties through β's (model parameters).

1. Mixed Models

Since the emission data were collected from a random sample of the on-road vehicle population tested using gasoline blends that narrowly varied the fuel parameters, and the inference is to be made for a wide array of gasoline blends that meet the California reformulated gasoline standards for the whole vehicle population in the state, the use of mixed models are imminent (1, 2). In this model development procedure the vehicles are considered the random effects and the fuel effects are considered the fixed effects. Having both random effects and fixed effects in the same linear model is referred to as a mixed effects model.

In a classical regression model, only fixed effect is considered. Hence, the inference is restricted to the levels in the experiment. In this report, the SAS system MIXED procedure is used. Version 7 of the SAS Institute's statistical software was used to generate estimates.

2. Forward-Stepwise Regression

The fuel properties as independent variables are not limited to seven linear or first-order terms as indicated earlier. They may include seven squared terms (e.g., OX*OX, T50*T50, etc.), and 21 crossed or interaction terms, where one fuel property is paired with the another (e.g., ARO*OX, RVP*T90, etc.), so a total of 28 possible second-order terms exists. Unlike squared and interaction terms, the linear terms on which the California gasoline regulations are based are always present in the model regardless if they are significant predictors or not, at five percent level; the other terms come in the model only if they are significant predictors.

The model building process is the same as the CaRFG2 Predictive Model and discussed briefly here, will start out with seven linear terms then add each of the remaining 28 terms only one at a time, called forward-stepwise regression. At the end, the most significant term by comparing its t-statistics, at least at 5 percent level, will be added to the seven linear variables that are already in the model. This variable selection is repeated until there is no more significant variable that can be included in the model. However, at any stage when a variable ceases to be significant upon adding another, the variable is removed from the model, and it is potentially reselected at later steps.

3. Random Balance

The working database to build the Predictive Model includes wider range of fuel properties (fuel box) than is allowed by the California reformulated gasoline standards. Thus, the resulted 'raw' models that are developed over the entire range of data may have second-order terms that do not contribute to the predictive power of the model if it was to be constraint the data in a smaller box. For example, a quadratic term could be represented by a straight line over a small range, thus making a simpler model.

In order to simplify the model a technique called the 'random balance' developed by Dr. H. T. Mc Adams of the Advanced Computing Center of Argenta was used by the staff. Then, we estimate the emissions using these randomly generated data that are nearly orthogonal in fuel property terms. The error sum of squares of each term is sorted in descending order, which indicates the relative contribution of the term to the overall explanatory power of the model. Only terms that are included in the 99 percent of cumulative sum of squares will be retained in the final or 'random balance' model. Table 3 shows the Phase 2 reformulated gasoline fuel properties used in the random balance procedure.

D. REGRESSION EQUATIONS BY VEHICLE TECH CLASS

The comparison of regression equations by tech class and pollutant for the current and new models are discussed in the following paragraphs. The emphasis is on the second-order terms since all models contain the seven first-order terms.

Table 3

Fuel Properties Range ('Box')
For Random Balance Procedure

Fuel Property	Unit	Lower Limit	Upper Limit
Aromatic Hydrocarbons	% vol.	10	40
O l e f i n s	% vol.	0	10
O x y g e n	% wt.	0	3.5
Reid Vapor Pressure	psi	6.4	7.5
S u l f u r	ppmw	0	80
50% Distillation Temperature	deg. F	160	225
90% Distillation Temperature	deg. F	260	335

1. Technology Class 3

Since only a small number of observations (55 data points) was added to Tech 3 class database, we do not expect a significant change from the existing models. For the THC model, we have the same number of second-order terms in the raw models as shown in Table 4; however, the random balance is simpler than the existing model, losing one term (T90*ARO). In contrast, the NOx model picked up one new term (RVP*T50) although the raw model started out with two new terms (RVP*T50 and RVP*RVP). Table 5 summarizes the results.

2. Technology Class 4

At the beginning only the first five studies (Table 1) were added to the Tech 4 class database, about 1,300 new observations. The resulting models, THC and NOx, departed significantly from the current models (Tables 6 and 7), not only in the raw models but also in the random balance. THC pollutant shared three common terms in the final model while NOx had only one.

The sole purpose of predictive model is to allow gasoline producers to test their new blends, prior to market, if they would meet the emission standards by comparing the emission change of the blends to a reference fuel. The most common method is to

compute the percent emission change of the proposed gasoline by varying its properties to the allowable range of values.

Figures 1 and 2 are examples of the method. The first figure depicts how the THC emission changes, in percent, when oxygen is varied while all other fuel parameters are kept at certain values. The graph for the new model shows that THC emission increases with oxygen content of gasoline. This result is counterintuitive since engineering tests produce the opposite slope as correctly described in the existing model.

Table 4

Tech Class 3
Hydrocarbons Models
Summary of Model Coefficients in the Regression Equations

Second-Order Term	Current Model		New Model	
	R a w	Random Balance	R a w	Random Balance
Intercept	-0.79246	-0.79455	-0.77651	-0.79147
RVP	0.00450	0.00447	0.00044	0.00047
T50	0.01063	0.01025	0.01112	0.01086
T90	0.01303	0.01786	0.01253	0.00218
ARO	-0.03232	-0.03845	-0.03066	-0.04375
OL	-0.01864	-0.02101	-0.01909	-0.03064
OX	-0.02743	-0.02736	-0.02688	-0.02688
SU	0.00222	0.00193	0.00531	0.00550
T90*ARO	0.01845	0.01823	0.01811	
ARO*SU	-0.04031	-0.04054	-0.04563	-0.04566
RVP*T50	-0.01615	-0.01627	-0.01742	-0.01748
T90*OL	-0.00896		-0.00910	
ARO*OL	0.00982		0.00986	

Table 5

Tech Class 3
Oxides of Nitrogen Models
Summary of Model Coefficients in the Regression Equations

Second-Order Term	Current Model		New Model	
	R a w	Random Balance	R a w	Random Balance
Intercept	-0.15598	-0.15598	-0.13660	-0.07943
RVP	-0.01672	-0.01672	-0.02792	0.01356
T50	-0.01161	-0.01161	-0.01002	-0.00983
T90	0.00342	0.00342	-0.00056	-0.00052
ARO	0.05428	0.05428	0.05314	0.05321
OL	0.02292	0.02292	0.02294	0.02302
OX	0.01440	0.01440	0.01728	0.01724
SU	0.01786	0.01786	0.01601	0.01594
T90*ARO	-0.00978	-0.00978	-0.00808	-0.00968
T50*T90	-0.00858	-0.00858	-0.00971	0.00755
RVP*T50			0.00754	-0.00801
RVP*RVP			-0.00726	

Table 6

Tech Class 4
Hydrocarbons Models
Summary of Model Coefficients in the Regression Equations
(Adding 5 new studies to the existing database)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-1.16114	-1.18304	-1.10829	-1.08720
RVP	0.02484	-0.00850	0.01320	0.01328
T50	0.07649	0.07644	0.05634	0.05637
T90	0.02339	0.03895	0.03167	0.04579
ARO	0.00124	0.00137	-0.00054	-0.00026
OL	-0.00689	-0.00687	-0.00999	-0.00657
OX	-0.01026	-0.01035	-0.01015	-0.01003
SU	0.06909	0.11690	0.06163	0.09290
T50*T50	0.02585	0.02581	0.02074	0.02076
T90*ARO	0.01202	0.01208		
T90*OX	0.01517	0.01511		
T90*T90	0.01819	0.01821	0.01723	0.01724
ARO*ARO	-0.01199	-0.01197	-0.00735	-0.00734
T90*SU	-0.01583		-0.01294	
SU*SU	-0.01700		-0.01105	
RVP*RVP	0.00519			
T50*OX			0.02317	0.02323
RVP*ARO			0.01123	0.01130
T50*ARO			0.01353	0.01354
T90*OL			-0.00440	
RVP*OX			-0.01136	-0.01136

*Adding the first five studies in Table 1 to the existing database.

Table 7

Tech Class 4
Oxides of Nitrogen Models
Summary of Model Coefficients in the Regression Equations
(Adding 5 new studies to the existing database)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-0.62602	-0.58546	-0.59403	-0.62231
RVP	-0.00163	0.03006	0.00583	-0.00373
T50	0.00196	0.00195	0.00159	0.00161
T90	0.00127	-0.00820	0.00560	0.00568
ARO	0.00418	0.00415	0.01340	0.00322
OL	0.02594	0.02595	0.01214	-0.00117
OX	0.01122	0.01132	0.01106	0.02143
SU	0.05912	0.05009	0.05357	0.02610
OX*OX	0.01351	0.01349	0.00940	0.00952
RVP*RVP	-0.00493			
T90*SU	0.00971			
ARO*OX	-0.00577	-0.00579		
RVP*OX	0.00625	0.00628		
SU*OX			-0.01093	
ARO*SU			0.01065	
T50*T50			0.00811	0.00810
T90*ARO			-0.00296	-0.00297
RVP*T90			-0.01030	-0.01027
RVP*OL			0.00554	
ARO*ARO			-0.00608	-0.00607
OL*OL			0.00576	
RVP*SU			0.00776	

*Adding the first five studies in Table 1 to the existing database.

Figure F-1
THC Change Versus Oxygen
(All Other Properties at Flat Limits)

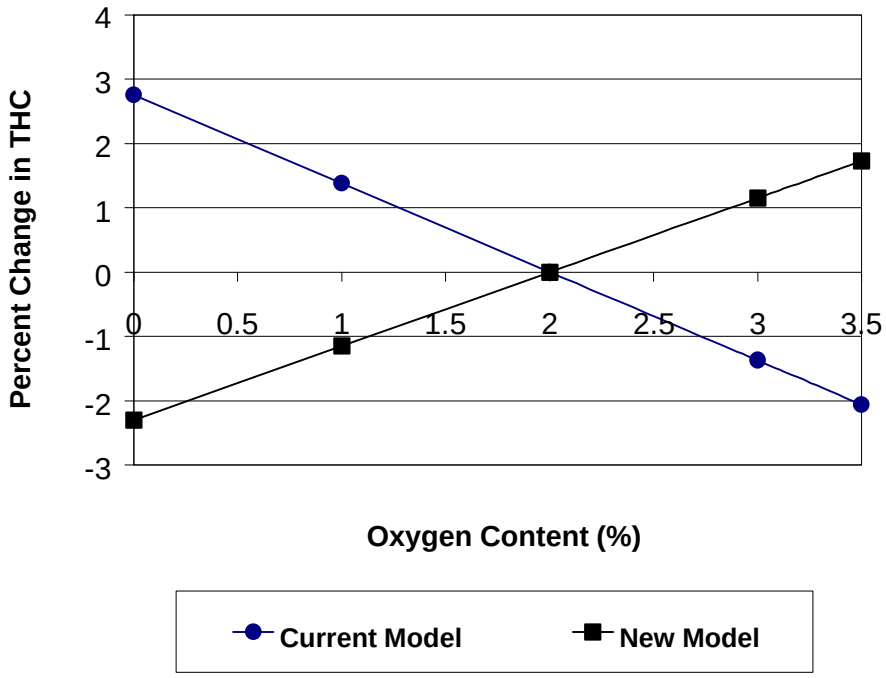
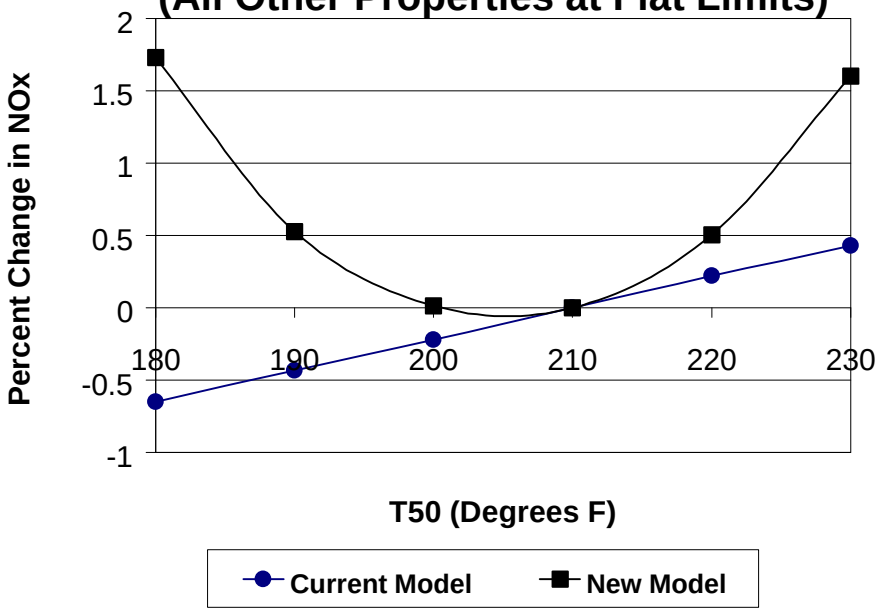


Figure F-2
NOx Change Versus T50
(All Other Properties at Flat Limits)



A similar plot of NOx against T50 (Figure 2) leads to a conclusion that the emission increases invariant to T50 level.

Both results forced us to believe that a few number of vehicles in the new studies had dominated the response. The staff devised a method to indicate the high influence vehicles as follows,

$$PE_i = \frac{\left[PVal_{in} - PVal_{out}\right]^2}{n_i}$$

where

PE_i = mean prediction error of vehicle i

$Pval_{in/out}$ = predicted value for the data of a particular study, to which vehicle i belongs; *in* means the data are included in parameters estimation, and *out* means excluded

n_i = number of observations for vehicle I in the corresponding study

The idea is to identify which vehicles in each new study would significantly change the model coefficients, keeping all terms from the raw model, and hence predicted values when they were included or excluded in model estimations. For instance, those vehicles, from a particular study, which change the prediction error when they are taken out or brought in the working database are potentially high influence. Once the mean prediction error is computed for each study, a histogram is utilized to identify which vehicles have the highest influence.

In THC model, two out of five studies were identified to contain high influence vehicles, five vehicles from ARBMSD96 study and one from EPA_PH3. A sum of 25 observations, ten were from the latter study (Table 8).

Table 8

Tech Class 4
Hydrocarbons Models
High Influence Vehicles

Study	Vehicle	
	ID #	# Observation
ARBMSD96	4	3
	7	3
	8	3
	10	3
	11	3
EPA_PH3	I	10

However, since the ARBMSD90 study has only 21 observations, we decided to remove the study completely from the current tech group.

Refitting the model to the database, without ARBMSD96 study and EPA_PH3 vehicle I, produced a new THC model (Table 9) that is similar to the new model in previous Table 6, except for RVP*OX term. This term is the most dominant in altering the oxygen response in Figure 1, relative to other terms, such as oxygen itself, and T90*OX, or T50*OX. Without RVP*OX term in the model, the hydrocarbons response is in the right direction as the existing model.

A similar approach failed to identify any high influence vehicles in the NOx model. In addition, we computed the mean squared error (MSE) of the models with and without the T50 squared term; however, dropping this term made T90*ARO term insignificant, so both were kept out from the model.

We supposed that both terms contributed a little to the overall fit of the model, especially for the data within the California reformulated gasoline region as displays in Table 3. However, the calculated MSE in Table 10 proves otherwise. The reduced

model increased the prediction error by 6%; hence both terms were reinserted to the model. Table 11 exhibits the results of refitting the model to the database without THC high influence vehicles (Table 8). The results are similar to the new NOx model in Table 7, except for the slight coefficient change.

On stakeholders' suggestion, we later added 70 more observations. Thirty-eight observations were from the ARB ethanol study in 1998 (ARBETOH) while the rest was from Chevron study (CHEVOX99). Unlike the NOx, the addition of new data altered THC model drastically, considering that only 70 observations were annexed. The model brought in 4 new terms, predominantly related to oxygen and olefin RVP*OX, OX*OX, OL*OX, and T50*OL (Table 12). Most importantly, the RVP*OX term was back in the model even after the random balance. As indicated earlier in Figure 1, the term shifted the hydrocarbons response, at odds with engineering expectation. The limited oxygen levels in fuels used to tests the vehicles may have caused the model to behave in such a contradictory fashion. Upon discussion with the stakeholders, we agreed to suspend both studies (ARBETOH and CHEVOX99).

Based on the new NOx model from Table 13 that excluded both ARBETOH and CHEVOX99, its response to olefin is relatively flat compared with the current model (Figure 3); moreover, the response to T90 shows the wrong slope (Figure 4). Unfortunately, these unexpected responses were untraceable since no apparent vehicles could be readily responsible for the culprits. These phenomena coupled with the number of terms that appeared in the equation concerned the staff that the model may overfit the data; it has twice as many terms (10 vs. 5) but only less than 20 percent more observations (about 7,000 vs. 5,700) as the existing model.

Table 9

Tech Class 4
Hydrocarbons Models
Summary of Model Coefficients in the Regression Equations
(Removing High Influence Vehicles)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-1.16114	-1.18304	-1.10950	-1.09097
RVP	0.02484	-0.00850	0.00983	0.00989
T50	0.07649	0.07644	0.05709	0.05712
T90	0.02339	0.03895	0.03144	0.04410
ARO	0.00124	0.00137	-0.00199	-0.00174
OL	-0.00689	-0.00687	-0.00930	-0.00609
OX	-0.01026	-0.01035	-0.01612	-0.01601
SU	0.06909	0.11690	0.05960	0.08675
T50*T50	0.02585	0.02581	0.01937	0.01939
T90*ARO	0.01202	0.01208		
T90*OX	0.01517	0.01511		
T90*T90	0.01819	0.01821	0.01766	0.01767
ARO*ARO	-0.01199	-0.01197	-0.00704	-0.00703
T90*SU	-0.01583		-0.01151	
SU*SU	-0.01700		-0.00948	
RVP*RVP	0.00519			
T50*OX			0.02104	0.02109
RVP*ARO			0.01385	0.01392
T50*ARO			0.01302	0.01303
T90*OL			-0.00414	

*Adding the first five studies in Table 1 to the existing database, except high influence vehicles as identified in Table 8 above.

Table 10

Tech Class 4
Oxides Nitrogen Models
Comparing the Mean Squared Error in California Fuel Box*

Model	DF**	SSE***	MSE****
Full Model (all parameters in Table 7)	411	5.065	0.0123
Reduced Model (w/o T50*T50, T90*ARO)	413	5.387	0.0130

*See Table 3
*DF = degrees of freedom, #observations - #parameters
**SSE = sum of squared error, (observed –predicted value)^2
***MSE = mean squared error, SSE/DF

Table 11

Tech Class 4
Oxides of Nitrogen Models
Summary of Model Coefficients in the Regression Equations
(Removing High Influence Vehicles)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-0.62602	-0.58546	-0.59140	-0.62081
RVP	-0.00163	0.03006	0.00533	-0.00468
T50	0.00196	0.00195	0.00164	0.00166
T90	0.00127	-0.00820	0.00534	0.00542
ARO	0.00418	0.00415	0.01325	0.00266
OL	0.02594	0.02595	0.01196	-0.00155
OX	0.01122	0.01132	0.01144	0.02177
SU	0.05912	0.05009	0.05382	0.02516
OX*OX	0.01351	0.01349	0.00901	0.00913
RVP*RVP	-0.00493			
T90*SU	0.00971			
ARO*OX	-0.00577	-0.00579		
RVP*OX	0.00625	0.00628		
SU*OX			-0.01088	
ARO*SU			0.01108	
T50*T50			0.00803	0.00802
T90*ARO			-0.00288	-0.00289
RVP*T90			-0.00989	-0.00986
RVP*OL			0.00559	
ARO*ARO			-0.00624	-0.00623
OL*OL			0.00591	
RVP*SU			0.00822	

*Adding the first five studies in Table 1 to the existing database, except high influence vehicles as identified in Table 8 above

Table 12

Tech Class 4
Hydrocarbons Models
Summary of Model Coefficients in the Regression Equations
(Adding 2 more new studies)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-1.16114	-1.18304	-1.13210	-1.13122
RVP	0.02484	-0.00850	0.01008	0.00027
T50	0.07649	0.07644	0.05532	0.05531
T90	0.02339	0.03895	0.03103	0.04172
ARO	0.00124	0.00137	-0.00354	-0.01791
OL	-0.00689	-0.00687	-0.00831	-0.00829
OX	-0.01026	-0.01035	-0.01252	-0.01236
SU	0.06909	0.11690	0.06100	0.08796
T50*T50	0.02585	0.02581	0.02068	0.02071
T90*ARO	0.01202	0.01208		
T90*OX	0.01517	0.01511		
T90*T90	0.01819	0.01821	0.01723	0.01719
ARO*ARO	-0.01199	-0.01197	-0.00695	-0.00704
T90*SU	-0.01583		-0.01124	
SU*SU	-0.01700		-0.00960	
RVP*RVP	0.00519			
T50*OX			0.02253	0.02266
RVP*ARO			0.00880	
T50*ARO			0.01368	0.01364
T90*OL			-0.00565	-0.00565
RVP*OX			-0.01080	-0.01079
OX*OX			0.01044	0.01046
OL*OX			0.00662	0.00665
T50*OL			0.00576	0.00576

*Adding the first seven studies in Table 1 to the existing database, except ARBMSD96 study and EPA_PH3 vehicle I.

Table 13

Tech Class 4
Oxides of Nitrogen Models
Summary of Model Coefficients in the Regression Equations
(Adding 2 more new studies)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-0.62602	-0.58546	-0.61130	-0.63955
RVP	-0.00163	0.03006	0.00593	-0.00346
T50	0.00196	0.00195	0.00177	0.00179
T90	0.00127	-0.00820	0.00506	0.00515
ARO	0.00418	0.00415	0.01318	0.00276
OL	0.02594	0.02595	0.01176	-0.00190
OX	0.01122	0.01132	0.01122	0.02173
SU	0.05912	0.05009	0.05309	0.02563
OX*OX	0.01351	0.01349	0.00997	0.01009
RVP*RVP	-0.00493			
T90*SU	0.00971			
ARO*OX	-0.00577	-0.00579		
RVP*OX	0.00625	0.00628		
SU*OX			-0.01107	
ARO*SU			0.01090	
T50*T50			0.00824	0.00823
T90*ARO			-0.00297	-0.00298
RVP*T90			-0.01017	-0.01014
RVP*OL			0.00563	
ARO*ARO			-0.00621	-0.00620
OL*OL			0.00603	
RVP*SU			0.00756	

*Adding the first seven studies in Table 1 to the existing database, except ARBMSD96 study and EPA_PH3 vehicle I.

Figure F-3
NOx Change Versus Oxygen
(All Other Properties at Flat Limits)

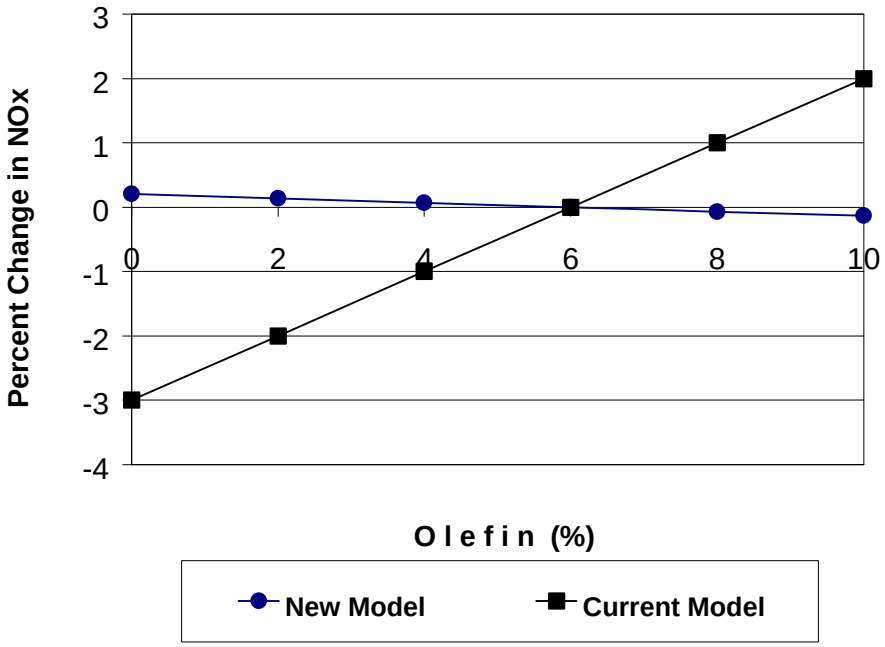
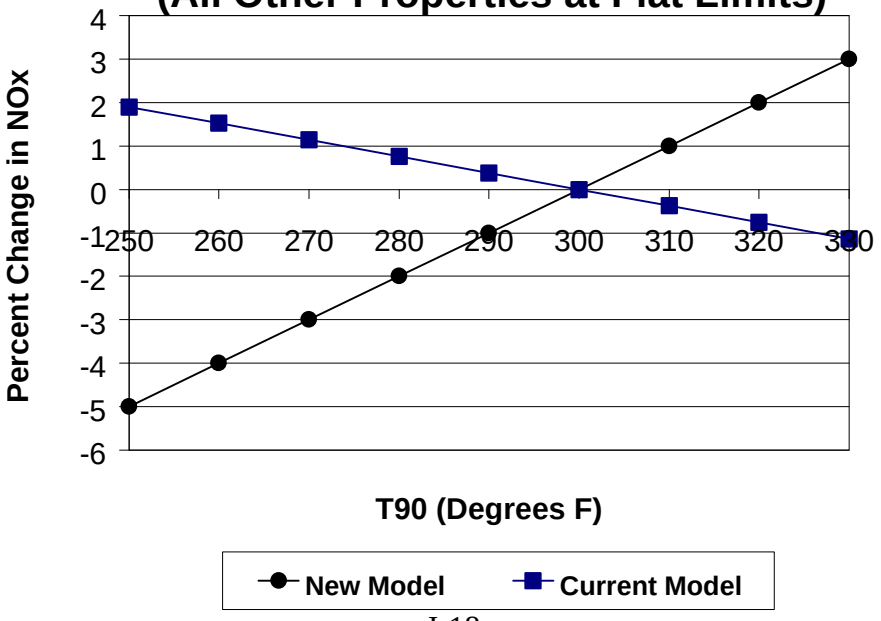


Figure F-4
NOx Change Versus T90
(All Other Properties at Flat Limits)



One of the characteristics of overfitting is when a model contains more terms or variables than necessary. Although the model fits extremely well to the data at hand, it performs poorly for new observations. This may be especially troublesome for a model like California predictive model where its chief application is prediction.

Still, the problem of large number of terms in a model can also be attributed to confounding effect. One variable or term by itself is not significant, but the presence of others can lend support to it to appear significant. For instance, often in our modeling work the last three or so variables entered the model demonstrated the effect.

The excess terms can also be partially attributed to the forward-stepwise regression tendency to artificially draw on relationship between the dependent variable and certain covariates; the procedure assumes the chosen covariates explain the variation in the dependent variable even though it is miniscule. These artifacts are particularly made worse when the procedure is applied to sparse high-dimensional and non-linear data. The case in point is the current working database. The data space was not smoothly populated. In fact, the data points were clustered, creating a lot of voids in between and forcing the model to extrapolate too far.

A solution to this problem is to employ a backward-stepwise regression procedure where the first variable (i.e., second-order term) to enter the model through the forward-stepwise regression is the first taken out, and the significance of the remaining terms are scrutinized. If any of the terms becomes insignificant, it should be dropped. The procedure is repeated for the subsequent term, until the next-to-last term in the model gets its turn. That is when the next-to-last term is removed, only the performance of the last term is evaluated; if significant, it is kept, otherwise jettisoned; and the process stops. At the end, only those terms come through the process stay; consequently, the trimmed down or reduced model is likely more robust. This procedure should be considered for future work.

After further consultation with the stakeholders, the staff refitted the existing models to the new database with the notion that they are more parsimonious, and have been proven to work quite well. Except for two terms (RVP*RVP and RVP*OX) in the NOx model, the rest of the terms was significant as revealed in Tables 14 and 15. However, before the final decision could be made to switch to the existing models structure. The staff was, first, to investigate the extent of overfitting and, second, to compare the predictive power of the corresponding models.

Since the NOx model seemed to overfit the data more so than the THC, the staff computed the information criteria that measure the goodness of fit of the former model as a function of the maximum value of the log likelihood; and number of parameters and data points.

Table 14

Tech Class 4
Hydrocarbons Models
Summary of Model Coefficients in the Regression Equations
(Fitting current model to the new database)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-1.16114	-1.18304	-1.12820	-1.13142
RVP	0.02484	-0.00850	0.01354	-0.01448
T50	0.07649	0.07644	0.06070	0.06068
T90	0.02339	0.03895	0.02745	0.04008
ARO	0.00124	0.00137	0.00011	0.00010
OL	-0.00689	-0.00687	-0.00936	-0.00938
OX	-0.01026	-0.01035	-0.01391	-0.01388
SU	0.06909	0.11690	0.06375	0.09279
T50*T50	0.02585	0.02581	0.02011	0.02010
T90*ARO	0.01202	0.01208	0.00848	0.00847
T90*OX	0.01517	0.01511	0.01046	0.01045
T90*T90	0.01819	0.01821	0.01700	0.01699
ARO*ARO	-0.01199	-0.01197	-0.00861	-0.00860
T90*SU	-0.01583		-0.01324	
SU*SU	-0.01700		-0.01057	
RVP*RVP	0.00519		0.00873	

*Current model fitted to the new database (the first five studies in Table 1, except ARBMSD96 study and EPA_PH3 vehicle I).

Table 15

Tech Class 4
Oxides of Nitrogen Models
Summary of Model Coefficients in the Regression Equations
(Fitting current model to the new database)

Second-Order Term	Current Model		New Model*	
	R a w	Random Balance	R a w	Random Balance
Intercept	-0.62602	-0.58546	-0.59756	-0.60161
RVP	-0.00163	0.03006	0.00640	0.00639
T50	0.00196	0.00195	-0.00020	-0.00020
T90	0.00127	-0.00820	0.00556	-0.00055
ARO	0.00418	0.00415	0.00906	0.00905
OL	0.02594	0.02595	0.01847	0.01847
OX	0.01122	0.01132	0.01379	0.01378
SU	0.05912	0.05009	0.04745	0.04324
OX*OX	0.01351	0.01349	0.01024	0.01024
RVP*RVP	-0.00493			
T90*SU	0.00971		0.00640	
ARO*OX	-0.00577	-0.00579	-0.00587	-0.00587
RVP*OX	0.00625	0.00628		

*Current model fitted to the new database (the first five studies in Table 1, except ARBMSD96 study and EPA_PH3 vehicle I).

Both the AIC and BIC statistics peak almost as the third term (T50*T50) joined the first two second-order terms (SU*OX and ARSU); the AIC is relatively constant afterward, and the SIC that levies heavier penalty for over-parameterized model declines. These suggest that the last several terms could be removed without decreasing the model's predictive power.

Indeed, it is desirable to get a more robust model that will produce smaller error estimate of future observations. To compare the predictive power of two models, the staff used a cross-validation procedure since neither of the information criteria above is appropriate; both criteria are primarily used to evaluate a series of nested model where one model is a subset of the other. The cross-validation procedure can be used to compare any two models as long as the same database is used. The disadvantage of this procedure is that it is numerically intensive and time consuming.

The cross-validation estimate of the prediction error is generated by randomly dividing the database into two: one for estimation (80%) and the other for validation (20%); the first subset is for parameter estimation while the second is for predicting the observed values. The squared difference between the observed and predicted values, called the prediction error, is then calculated. The process is repeated many times to get a better estimate of the mean prediction error. The results of this procedure for the Tech 4 raw models are given in Table 16. As expected, the new NOx model shows three percent less precise than the current model while the new THC model is, the opposite, about one percent more precise. Based on these findings, stakeholders accepted going back to the existing terms for all models, THC and NOx (see Tables 14 and 15).

Table 16

Tech Class 4

Mean Prediction Errors

Model	Existing	New
Hydrocarbons	0.702	0.701
Oxides of Nitrogen	0.780	0.782

3. Technology Group 5

The staff attempted to build stand-alone Tech 5 class models, but the results were contrary to what was expected; the limited data which mainly focused on low sulfur fuel made the task of building full models practically unattainable. The stakeholders concurred that Tech 5 vehicles are similar to Tech 4, so a consensus was reached to nest Tech 5 within Tech 4. Namely, the databases of both technology groups are pooled together and all the terms derived from Tech 4 models (Tables 14 and 15) are retained.

Furthermore, the models are supplanted by sulfur adjustment terms, specific to Tech 5 vehicles, handled by indicator variable.

The following equation describes the structure of the model:

$$\mathbf{y}_p = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + ... + \beta_n\mathbf{x}_n + \beta_{n+1}\mathbf{I} + \beta_{n+2}\mathbf{I}S + \beta_{n+3}\mathbf{I}S^2 + \boldsymbol{\varepsilon}$$

where

$\mathbf{y}_p$ (vector) = measured emission for pollutant, p, from Tech 4 and 5 vehicles
β_i = parameter to be estimated from the pooled data
$\mathbf{x}_i$ (vector) = fuel property with second-order terms as shown in Tables 14 15
I (vector) = indicator variable (zero if tech group is 4, one otherwise)
S and S^2 = sulfur content of the fuel and sulfur squared
$\boldsymbol{\varepsilon}$ (vector) = error term

When the model is refitted to the pooled data, the intercept will be β_0 plus β_{n+1}; similarly, the coefficient for the sulfur term will be combined, as well as sulfur squared if the corresponding term from Tech 4 exist (e.g., Tech 4 THC).

Table 17 shows the results of refitting the above model to the pooled data. Except for RVP*OX in the NOx model, most of the terms from Tech 4 raw models were retained; moreover, they have similar coefficient magnitude and directionally same slope as expected. However, the Tech 4 models will be based on Tables 13 and F14 results.

Table 17

Tech Class 5
Summary of Model Coefficients in the Regression Equations
(Using Tech 4 models and sulfur adjustments)

Second-Order Term	Hydrocarbons		Oxides of Nitrogen	
	R a w	Random Balance	R a w	Random Balance
Intercept	-2.52154	-2.50695	-1.78141	-1.72822
RVP	0.01295	-0.01528	0.00679	0.00680
T50	0.05749	0.05747	-0.00148	-0.00148
T90	0.02796	0.03846	0.00353	-0.00477
ARO	0.00098	0.00098	0.01013	0.01012
OL	-0.00965	-0.00968	0.01883	0.01883
OX	-0.01478	-0.01475	0.01373	0.01371
SU	0.18673	0.18673	0.31524	0.43284
T50*T50	0.01906	0.01905		

T90*ARO	0.00883	0.00882		
T90*OX	0.01015	0.01015		
T90*T90	0.01653	0.01652		
ARO*ARO	-0.00863	-0.00862		
T90*SU	-0.01101		0.00868	
RVP*RVP	-0.03183			
SU*SU	0.00880		-0.06438	
OX*OX			0.01013	0.01013
ARO*OX			-0.00592	-0.00592

The forthcoming data, expected the end of October 1999, will further determine the shape of sulfur curve at the low level, close to zero, and if other fuel property (e.g., oxygen) will need similar adjustments as sulfur.

E. EXHAUST TOXICS MODELING

The basic approach to modeling the exhaust toxics emissions was the same as that used for exhaust THC and NOx emissions. That is, the same basic model form was used. However, a few differences were employed in modeling the toxics and these will be discussed here.

Exhaust toxics models were developed for benzene, 1,3-butadiene, formaldehyde, and acetaldehyde. Two models were developed for each of these four pollutants. The two models developed were a Tech class 3 model and a Tech class 4 and 5 model. In contrast to the models for exhaust THC and NOx, separate models for Tech class 4 and 5 were not developed. The reason for this is that there is very little toxics emissions data for Tech class 5. This lack of data would have resulted in questionable model results for Tech 5 had it been modeled separately. Therefore, the staff decided to model Tech class 4 and 5 together.

In the modeling of exhaust toxics emissions, second order terms (squared terms and interactions) were not included in the models. The reason for this is that there is much less data for toxics emissions than for exhaust THC and NOx. Also, contrary to the approach used for exhaust THC and NOx, not all the first order terms were included in the toxics models. Only first order terms which were significant at the 95 percent significance level (P-value of 0.05 or less) were included in the exhaust toxics emissions models. The same stepwise regression approach used to add second order terms in the exhaust THC and NOx models was used to add first order terms to the exhaust toxics models. The models which resulted from the stepwise procedure appeared, for the most part, to be consistent with engineering expectations. Therefore, these models were retained for purposes of the Phase 3 predictive model and the new data was not fit to models containing the terms contained in the Phase 2 predictive model, as was done with the exhaust THC and NOx models.

A final difference in the modeling approach for exhaust toxics was the inclusion in the models for formaldehyde and acetaldehyde of a term for the amount of oxygen in the gasoline as ethanol, in addition to a term for total oxygen content. This differs from the exhaust THC and NOx models which contain a term for only the total oxygen content. The reason for the oxygen as ethanol term in the aldehyde models is to account for the fact that MTBE and ethanol have different effects on the generation of formaldehyde and acetaldehyde. Specifically, the use of ethanol results in greater acetaldehyde emissions than from the use of MTBE, while MTBE causes greater formaldehyde emissions. The coefficients for the exhaust toxics models are shown below in Table 18.

Table 18
Coefficients for Exhaust Toxics Models

Model Term	Tech Class 3			
	Benzene	**Butadiene**	**Formaldehyde**	**Acetaldehyde**
Intercept	2.9568	0.6717	2.1684	1.1012
RVP				
Sulfur	0.0684			
Aromatic HC	0.1519		-0.0754	-0.0922
Olefins		0.1841		
Oxygen	-0.0330		0.1228	0.00123
Oxygen (as EtOH)			-0.1230	0.5468
T50		0.1139		
T90				
Benzene	0.1203		-0.1423	
	Tech Class 4 and Tech Class 5			
Model Term	**Benzene**	**Butadiene**	**Formaldehyde**	**Acetaldehyde**
Intercept	2.3825	0.4309	1.0589	0.1674
RVP	0.0311			
Sulfur	0.0965		-0.0414	0.0279
Aromatic HC	0.1552	-0.03604	-0.0547	-0.0555
Olefins	-0.0255	0.1035		
Oxygen		-0.0251	0.0637	0.0238
Oxygen (as EtOH)			-0.0982	0.4670
T50	0.0467	0.0371		0.0431
T90		0.0945	0.0604	0.0625
Benzene	0.1169	0.0364		0.0615

Example SAS code:

```
FILENAME IN1 'C:\WIN\PM\MEGA_1.FIN';
OPTIONS LS=80 CLEANUP ;
DATA FINLMEGA;
INFILE IN1           MISSOVER;

INPUT STUDY $ VEHICLE $ FUEL $ MODEL_YR DRYBULB NOX CO THC
      NMHC AR BENZ ETBE ETOH MTBE TAME OL OX RV
      SU T5 T9 EXBENZ EX13BUTD EXFORMAL EXACTALD;

/* FUEL SPECIFICATIONS BASED ON THE ASTM STANDARDS */
IF RV GT 10 OR SU GT 1000 OR OX GT 4.0 OR
   T5 GT 250 OR T9 GT 374 OR DRYBULB LT 68 OR
   DRYBULB GT 95 THEN DELETE;

/* DELETE MISSING VALUES */
IF RV='.' OR AR='.' OR OL='.' OR SU='.' OR
   OX='.' OR T5='.' OR T9='.' THEN DELETE;

/* TECH GROUPS DEFINITION */
IF MODEL_YR LT 1975 THEN TECH =1;
   ELSE IF MODEL_YR LT 1981 THEN TECH = 2;
   ELSE IF MODEL_YR LT 1986 THEN TECH = 3;
   ELSE IF MODEL_YR LT 1994 THEN TECH = 4;
   ELSE TECH = 5;

/* CREATE NEW VARIABLES */
LN_THC = LOG (THC);
LN_NOX = LOG (NOX);
NEW    = STUDY||VEHICLE;

/* TECH GROUPS SELECTION */
IF TECH = 4 OR TECH = 5;
RUN;

PROC STANDARD MEAN=0 STD=1 DATA=FINLMEGA OUT=TEMP000 PRINT;
TITLE1 "FUEL PROPERTY MEANS AND STANDARD DEVIATIONS";
TITLE2 "POOLED DATA, TECH 4 AND 5 COMBINED";
VAR RV T5 T9 AR OL OX SU;
RUN;

DATA TEMP100;
    SET TEMP000;
    /* LIMIT TO TECH 4 ONLY */
        IF TECH=4;

    /* REMOVE TECH 4 HIGH INFLUENCE VEHICLES */
        IF TECH=4 AND STUDY='ARBMSD96' THEN DELETE;
        IF TECH=4 AND STUDY='EPA_PH3' AND VEHICLE='I' THEN DELETE;

    /* INTERACTION TERMS */
        RVRV=RV*RV;
        RVT5=RV*T5;
        RVT9=RV*T9;
        RVAR=RV*AR;
```

```
        RVOL=RV*OL;
        RVSU=RV*SU;
        RVOX=RV*OX;
             T5T5=T5*T5;
             T5T9=T5*T9;
             T5AR=T5*AR;
             T5OL=T5*OL;
             T5SU=T5*SU;
             T5OX=T5*OX;
        T9T9=T9*T9;
        T9AR=T9*AR;
        T9OL=T9*OL;
        T9SU=T9*SU;
        T9OX=T9*OX;
             ARAR=AR*AR;
             AROL=AR*OL;
             ARSU=AR*SU;
             AROX=AR*OX;
        OLOL=OL*OL;
        OLSU=OL*SU;
        OLOX=OL*OX;
             SUSU=SU*SU;
             SUOX=SU*OX;
        OXOX=OX*OX;
RUN;

PROC MIXED DATA=TEMP100 MAXITER=500 CONVH=1E-8 METHOD=REML NOCLPRINT
NOITPRINT;
CLASS NEW;

TITLE "TECH 4 NOX MODEL";

MODEL LN_NOX = RV T5 T9 AR OL OX SU
               SUOX ARSU T5T5 OXOX T9AR RVT9 RVOL ARAR OLOL RVSU
               /S DDFM=RES;

RANDOM         INT RV T5 T9 AR OL OX SU
               SUOX ARSU T5T5 OXOX T9AR RVT9 RVOL ARAR OLOL RVSU
               /SUB=NEW;
RUN;
```

Example SAS output:

Name	Mean	Standard Deviation	N
RV	8.308910	0.846737	7969
T5	207.019049	17.195294	7969
T9	311.785331	21.595186	7969
AR	27.849881	7.004743	7969
OL	6.806801	4.665131	7969
OX	1.355654	1.224639	7969
SU	180.770373	147.006156	7969

The Mixed Procedure

Model Information

Data Set	WORK.TEMP100
Dependent Variable	LN_NOX
Covariance Structure	Variance Components
Subject Effect	NEW
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Dimensions

Covariance Parameters	19
Columns in X	18
Columns in Z Per Subject	18
Subjects	876
Max Obs Per Subject	66
Observations Used	7000
Observations Not Used	0
Total Observations	7000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	NEW	0.4163
RV	NEW	0.000325
T5	NEW	0.000511
T9	NEW	0.001548
AR	NEW	0.001082
OL	NEW	0.000067
OX	NEW	0.000681
SU	NEW	0.001517
SUOX	NEW	0.000337
ARSU	NEW	0.000044
T5T5	NEW	0.000122

OXOX	NEW	0
T9AR	NEW	9.926E-6
RVT9	NEW	0.000396
RVOL	NEW	-55E-22
ARAR	NEW	-238E-24
OLOL	NEW	0
RVSU	NEW	1.48E-20
Residual		0.01248

The Mixed Procedure

Fitting Information

Res Log Likelihood	2666.3
Akaike's Information Criterion	2652.3
Schwarz's Bayesian Criterion	2618.9
-2 Res Log Likelihood	-5332.6

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-0.6113	0.02293	6982	-26.66	<.0001
RV	0.005926	0.003994	6982	1.48	0.1379
T5	0.001766	0.004112	6982	0.43	0.6675
T9	0.005064	0.004002	6982	1.27	0.2058
AR	0.01318	0.004239	6982	3.11	0.0019
OL	0.01176	0.003930	6982	2.99	0.0028
OX	0.01122	0.003307	6982	3.39	0.0007
SU	0.05309	0.004543	6982	11.68	<.0001
SUOX	-0.01107	0.003063	6982	-3.62	0.0003
ARSU	0.01090	0.003129	6982	3.48	0.0005
T5T5	0.008236	0.001930	6982	4.27	<.0001
OXOX	0.009965	0.003674	6982	2.71	0.0067
T9AR	-0.00297	0.001381	6982	-2.15	0.0318
RVT9	-0.01017	0.003131	6982	-3.25	0.0012
RVOL	0.005630	0.001829	6982	3.08	0.0021
ARAR	-0.00621	0.001451	6982	-4.28	<.0001
OLOL	0.006031	0.001857	6982	3.25	0.0012
RVSU	0.007558	0.003112	6982	2.43	0.0152

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
RV	1	6982	2.20	0.1379
T5	1	6982	0.18	0.6675
T9	1	6982	1.60	0.2058
AR	1	6982	9.67	0.0019
OL	1	6982	8.96	0.0028
OX	1	6982	11.50	0.0007
SU	1	6982	136.53	<.0001

SUOX	1	6982	13.07	0.0003
ARSU	1	6982	12.14	0.0005
T5T5	1	6982	18.22	<.0001
OXOX	1	6982	7.36	0.0067
T9AR	1	6982	4.61	0.0318
RVT9	1	6982	10.55	0.0012
RVOL	1	6982	9.48	0.0021
ARAR	1	6982	18.33	<.0001
OLOL	1	6982	10.55	0.0012
RVSU	1	6982	5.90	0.0152