Multiple Regression Analysis of Hydrocarbon Degradation in a Vapor-Phase Biofilm Reactor

by

Darcie H. Dunlap

Statistical Writing Project
Montana State University
April 21, 1995

1. INTRODUCTION

The treatment of contaminated groundwater, a major undertaking in the U.S. and abroad, is accomplished with a wide variety of technologies. One of the most common and inexpensive techniques uses a vapor extraction system to remove air from subsurfaces laden with hydrocarbons. While effective and relatively efficient, this technique produces a contaminated air stream which requires treatment before the air can be released to the atmosphere.

In a collaborative study, researchers at Montana State University and the Orange County Water District in California investigated characteristics of biological soil vapor treatment in a biofilm reactor. The soil vapor can contain many different hydrocarbons. The investigators were particularly interested in the effects of various hydrocarbon combinations on the biotransformation of those hydrocarbons. An extensive data set was collected and used to develop an empirical model of the system. The model can aid in the design and scale-up of future systems.

2. DATA

The data were based on samples removed from a vapor-phase biofilm reactor located in Orange County, California. The purpose of the experiment was to obtain information about the degradation rates for six hydrocarbons: hexane, benzene, trimethylpentane 2-2-4, toluene, octane, and p-xylene. The six hydrocarbons could simultaneously flow into the reactor; the loading rates were controlled by the operator. After setting the loading rates of the hydrocarbons, the operator would wait until the system reached steady state, then measure (i) the associated degradation rate for each hydrocarbon and (ii) the CO_2 production. The operator also recorded six relevant physical characteristics of the reactor: influent humidity, pH balance, differential pressure, redox potential, conductivity, and reactor temperature. Each setting of hydrocarbon

loadings and associated measurements constitutes a "case." The data set for the total experiment comprises a large number of cases, each case consisting of the *response variables* (degradation rates and CO_2 production) and *explanatory variables* (the six hydrocarbon loading rates and six physical measurements).

Don Phipps of the Orange County Water District collected the data. He sent the data to the Center for Biofilm Engineering on a computer diskette in the form of Symphony™ spreadsheet files. His goal was to reach conclusions about the effects of the various hydrocarbon loading rates and physical characteristics on hydrocarbon degradation rates and CO_2 production.

3. BACKGROUND

Don Phipps conducted initial statistical analyses in Orange County. He sent the results of his analyses; viz., multiple linear regression models (see Figure 1 for a hexane model example) and partial correlation plots (see Figure 2 for a hexane model example) for each of the hydrocarbon degradation rates. He had chosen to eliminate the intercept term since hydrocarbon degradation rate should be zero if that same hydrocarbon's loading rate is zero. He wanted to quantitatively interpret the estimated coefficients for the degradation models by observing the size and sign of the coefficients. The interpretations seemed complicated so he decided to seek the advice of a statistician, specifically, Dr. Marty Hamilton. After reviewing the data, Dr. Hamilton decided that more extensive statistical analyses were required. At that time, I was hired to do those analyses.

4. INITIAL ANALYSES

4.1 Translating the Data

After translating the data files into a form acceptable to the statistical software, I repeated the multiple

regression analyses to see if I could reproduce Don's results. This effort provided confidence that the data matrix had been translated correctly and had the right labels attached to the data columns. During the course of these analyses, I discovered that Don's models were not based on all the data. It turned out that his computer program unintentionally deleted the first case during each analysis.

4.2 Choosing the Relevant Subset of the Data

I broke the data into three time period groups as suggested in the material Don included with the data. Appendices A, B, and C correspond to the three time periods: **period A** which represents operation using only indigenous nitrogen, **period B** representing the addition of 1.8mM of NH4Cl to the reactor, and **period C** which contained continually declining nitrogen.

The plots for period A indicated large correlations between the loading variables (Figure 3). Dr. Hamilton and I judged that it would be impossible to separate the effects of the different loading variables. The plots for period B showed that each loading variable varied over a narrow range (note the axes of Figure 4). We judged that over such narrow ranges, the response variables did not change enough to provide modelling information. For these reasons, we chose to do the analyses only on period C data. In time period C, different loads were turned on and off at different times, resulting in less collinearity than for period A (see Figure 5a). **The statistical analyses in this report pertain only to the 291 cases of period C.**

4.3 Important Characteristics of the Data

After plotting the data and attempting to model the degradation rates with linear explanatory variables, I discovered some important characteristics of the data for period C. First, the scatterplots for two explanatory variables at a time indicated that several of the variables were highly correlated (see Figure 5a).

With multicollinearity present, it is nearly impossible to quantitatively interpret coefficients of models. A correlation matrix (see Table 1 below) also shows the strong amount of correlation present in the explanatory variables, particularly between benzene, toluene, trimethylpentane, and octane loads. Secondly, scatterplots of the response variables by the explanatory variables (see Figures 5b and 5c) presented some curvature, which indicated that a quadratic model was probably more appropriate than a model in which the explanatory variables are entered in linearly. Finally, an intercept is possible because the clients are interested in modelling points away from the origin, and they are not interested in the points at or near the origin.

Table 1 Correlation Coefficients for the Hydrocarbon Loading Rates

LOADS	Hexane	Benzene	Toluene	TMP	Octane	P-xylene
Hexane	1.000	-.046	-.029	-.022	-.019	-.408
Benzene		1.000	.796	.796	.797	.360
Toluene			1.000	.796	.800	.355
TMP				1.000	.801	.254
Octane					1.000	.358
P-xylene						1.000

5. MULTIPLE REGRESSION MODELS: METHODOLOGY

My objective was to construct multiple regression models for each of the six hydrocarbon degradation rates. The models provided a means for determining which hydrocarbon loadings affected each degradation rate.

Dr. Hamilton and I conducted the statistical analysis on this project in consultation with Dr. Warren Jones. The statistical software packages S-plus and SAS were used to complete the analyses.

5.1 Initial Discoveries

After studying the plots and initial analyses, four decisions guided the subsequent modelling effort.

1. If a hydrocarbon loading rate is zero, or nothing is going into the system, then none of that hydrocarbon should be degrading and the corresponding degradation rate should be zero. A model for hydrocarbon degradation is helpful only to the extent that it describes degradation when there is a positive loading of that hydrocarbon. Therefore, I did not use data points where the hydrocarbon loading rate was zero when modelling that hydrocarbon's degradation rate. For example, in modelling hexane degradation (see Table 2 on page 12), all cases where hexane loading was zero were removed from the data set and the remaining 247 cases were used in the modelling process. The same steps were used for the other five hydrocarbons. Thus, having an intercept term in the models was possible.

2. Dr. Jones explained that the six hydrocarbons could be split into two classes: the aromatic class consisting of benzene, toluene, and p-xylene, and the aliphatic class containing hexane, trimethylpentane 2-2-4, and octane. In interpreting the models, I gave special attention to the effects within and between these two classes. For comparison purposes, I also decided to build additional models that used total aromatic degradation and total aliphatic degradation as response variables.

3. When modelling the degradation of a single hydrocarbon, I had included loading interaction terms and squared terms, which meant there was a total of 33 potential explanatory variables. Ultimately, we decided that it was not necessary to include in the model all possible pair-wise interactions among the six hydrocarbon loadings. Although base models for total hydrocarbon degradation and CO_2 production included 33 explanatory variables, the base models for each hydrocarbon degradation contained only the five interactions involving the hydrocarbon being modelled. For example, in the

model for hexane degradation, I only considered the five pair-wise interactions of hexane loading with each of the other hydrocarbon loading rates.

4. We discussed the potential problem of a time lag in the degradation rates and CO_2 production. By the nature of the experiment and the time allotted between measurements, we decided that there should be no time effect in the degradation rates, but there was a possibility of a time effect in CO_2 production. To test this possibility, I rearranged the data so that CO_2 production would be lagged behind the other variables by *n* time measurements. Then I graphed the correlation values for *n* (values 1 to 5) and CO_2 production (see Figure 6). The graphs demonstrated no obvious lag in the CO_2 production responses. Therefore, our models do not include any time lagged variables.

5.2 Models

The initial models contained the same twelve linear explanatory variables that Don Phipps included in his regression analysis. Those twelve variables were the following:

Loads	Physical Traits
Hexane	Influent Humidity
Benzene	Differential Pressure
Trimethylpentane 2-2-4	pH
Toluene	Redox Potential
Octane	Reactor Temperature
P-xylene	Conductivity

Curvature in the degradation versus loading variable plots (see Figure 5b) indicated that a quadratic model should be considered in modelling the degradation rates. This led us from the original linear model with twelve explanatory variables to 33 potential explanatory variables. The 33 explanatory variables contained an additional six variables that were the squares of the hydrocarbon loadings and fifteen additional interaction variables. The interaction variables were developed by multiplying each hydrocarbon loading by

the other five hydrocarbon loadings. By including these extra variables, we created parabolic models that corresponded to the curvature indicated in the scatterplots. The 33 variables were as follows:

Loads	Loads Squared	Interactions	Physical Traits
Hexane	Hexane Squared	P-xylene by Hexane	Influent Humidity
Benzene	Benzene Squared	P-xylene by Benzene	Differential Pressure
TMP 224	TMP Squared	P-xylene by TMP	pH
Toluene	Toluene Squared	P-xylene by Toluene	Reactor Temperature
Octane	Octane Squared	P-xylene by Octane	Redox Potential
P-xylene	P-xylene Squared	Octane by Toluene	Conductivity
		Octane by TMP	
		Octane by Benzene	
		Octane by Hexane	
		Toluene by TMP	
		Toluene by Benzene	
		Toluene by Hexane	
		TMP by Benzene	
		TMP by Hexane	
		Benzene by Hexane	

The 33 variables were used in models for total hydrocarbon degradation and for CO_2 production. But when modelling the degradation of a single hydrocarbon, we reduced the 33 potential explanatory variables to 23 variables: the six hydrocarbon loads and their six squared terms, the six physical characteristics, and the five interactions of the loading for the hydrocarbon being modelled with the other five loading variables. Hence, the 23 variable model became our base model for single hydrocarbon degradation.

For each of the single hydrocarbon degradation rates, we used backward, stepwise, forward, and R-square selection procedures in the SAS statistical package (see Littell, Freund, and Spector, 1991) to reduce the 23 variable base model to a smaller model, designated as the "final model." The backward selection procedure begins with the 23 variable model and removes variables, one at a time, that have the least amount of influence on the model for the response variable. The forward selection procedure begins with no variables and then adds variables, one at a time, that have the most influence on the model. The

stepwise selection procedure is a combination of backward and forward selection in that at each step it will either add a new variable with the next largest amount of influence on the model, or eliminate a variable previously entered into the model because it is no longer needed in the model. The R-square procedure provides a list of the best models for different numbers of explanatory variables. "Best" was based on higher adjusted-R^2 values and appropriate Mallow's Cp values. In each of the procedures, p-values for F tests and increases in the R^2 values (or decreases in the residual sum of squares) were used to decide how much influence each variable had on the model. Mallow's Cp statistic was used to approximate the number of variables appropriate for each model (see Myers, 1990). Mallow's Cp statistic should be close in value to the number of explanatory variables in the model. See Figures 7a and 7b for graphs of the various Cp values and adjusted-R^2 values that indicate how these two statistics were used to decide on the appropriate-sized model for benzene degradation. In this example, it appears that better models for benzene degradation contain about nine terms. Using the results of these automated selection procedures and considering the associated R^2 values, adjusted-R^2 values, Mallow's Cp statistics, and plots, I selected the model that I believed best fit the data. I also used individual variable investigation of t-test results to aid in my selection process. If there existed two or more models that met the above criteria, I chose the model with the fewest terms according to a hierarchical structure where linear terms were favored over interaction and squared terms. Thus, the model chosen was parsimonious and easier to interpret.

Although it makes intuitive sense that a zero loading rate for a hydrocarbon should result in a zero degradation rate, we chose not to force the intercept term to be zero because wee are not interested in modelling points near the origin. Rather, we are interested in modelling a cluster of points situated away from the origin, and it is reasonable for the chosen model to contain an intercept term other than zero. Nevertheless, if the p-value for an intercept term was not significant (about $p \geq .15$), I forced the model to

exclude an intercept term by using SAS commands such as NOINT and RESTRICT INTERCEPT=0. The latter term gives R^2 values that are comparable to R^2 values of models containing an intercept. In the NOINT cases, the R^2 value is equivalent to calculating the squared correlation between responses and predictions (see Freund and Littell, 1991).

6. MULTIPLE REGRESSION MODELS: RESULTS

6.1 Description of the model tables

There are six tables in the "Results" section, one for each hydrocarbon degradation. Each table contains adjusted-R^2 and intercept information for five different models for single hydrocarbon degradation. Descriptions of the five types of models are as follows:

- The first is the model containing twelve linear terms and all 291 cases. This is essentially the analysis that Don Phipps had done.

- The second model shows how the adjusted-R^2 changed when the cases where loading rate equalled zero were removed. For example, in Table 2 the adjusted-R^2 for the first model containing 291 cases is .3155. When the number of cases was reduced to 247, the adjusted-R^2 decreased to 0.1650. In essence, the inclusion of 44 cases with zero loading doubles the adjusted-R^2. However, these 44 cases are irrelevant to our modelling goal.

- Next is a model reduced from the 12 linear variable model to fewer terms. This model for each of the six degradations shows an increased adjusted-R^2, indicating the benefit of removing extraneous variables from the model. In the hexane degradation model (see Table 2), the adjusted-R^2 increases

from 0.1605 to 0.1805.

- The fourth set of values demonstrate how the adjusted-R^2 values increased when I changed to 23 variables, indicating that the additional variables contained useful information about the degradation rates. In Table 2, notice how the adjusted-R^2 for the hexane degradation model changes from 0.1805 to 0.4117.

- Extraneous variables were removed from the 23 variable base model using previously described methods. The goal here was to arrive at a smaller model (the final model) that still fit the data nearly as well as the 23 variable model. For the hexane degradation model, the final model contained 16 explanatory variables and the adjusted-R^2 was 0.4167.

In the tables, N refers to the number of cases used in the model. There are a total of 291 cases, but in modelling each hydrocarbon degradation, the cases where the loading rate for that particular hydrocarbon was zero were removed, so N is less than 291.

In the intercept column in the tables, a Y or N was recorded, indicating whether or not an intercept term was included in the model. If a Y, for Yes, was recorded, a p-value associated with the intercept is also included.

The most valuable part of the tables to the researchers at Orange County is the "Effect" section, because it includes a qualitative analysis of the effects of our explanatory variables on the responses. **The "Effects" section of the table indicates whether or not each hydrocarbon load and physical trait included in the final model effects the degradation rate being modelled.** Reduced model F-tests were

used to test how important certain variable groups were to the model. A p-value less than 0.01 indicates that a variable, or related group of variables, has a statistically significant effect on the degradation rate. The adjusted-R^2 values in the column provided refer to the new model after all explanatory variables containing that hydrocarbon load or physical trait have been removed. This adjusted-R^2 can be compared to the adjusted-R^2 of the final model to see how much that hydrocarbon load or physical trait effects the final model.

6.2 Model Results for the Six Single Hydrocarbon Degradations

Table 2

Response: **HEXANE DEGRADATION**

Model Description	Adj R^2	Intercept
12 linear variables, N=291	0.3155	Y .3397
12 linear variables, N=247	0.1650	No
7 linear variables, N=247	0.1805	No
23 variables, N=247	0.4117	No

The Final Model:

16 variables, N=247	**0.4167**	**No**
Predictors: Hexane Load	**P-xylene Load Squared**	
Benzene Load	**Benzene by Hexane**	
Trimethylpentane Load	**Toluene by Hexane**	
P-xylene Load	**Octane by Hexane**	
Benzene Load Squared	**P-xylene by Hexane**	
Trimethylpentane Load Squared	**Influent Humidity**	
Toluene Load Squared	**Reactor Temperature**	
Octane Load Squared	**Conductivity**	

Effects of:	Variables removed	New Adj R^2	P-value
Hexane	Hexane Load, Benzene by Hexane Toluene by Hexane, Octane by Hexane P-xylene by Hexane	0.1352	0.0000
Benzene	Benzene Load, Benzene by Hexane Benzene Load Squared	0.2195	0.0000
TMP 224	Trimethylpentane Load Trimethylpentane Load Squared	0.3842	0.0007
Toluene	Toluene Load Squared Toluene by Hexane	0.3926	0.0050
Octane	Octane Load Squared, Octane by Hexane	0.3810	0.0004
P-xylene	P-xylene Load, P-xylene by Hexane P-xylene Load Squared	0.3754	0.0003
Influent Humidity		0.4025	0.0105
Reactor Temperature		0.3906	0.0009
Conductivity		0.3784	0.0001

Summary: Table 2 shows that hexane load, benzene load, trimethylpentane load, octane load, p-xylene load, reactor temperature, and conductivity all have a strong statistically significant effect on hexane degradation. Hexane load has the most effect on hexane degradation, but this is probably because hexane load is included in five variables. Toluene load also influences hexane degradation, but not as significantly. Influent Humidity does not have a significant effect at the 0.01 level.

Table 3

Response: **BENZENE DEGRADATION**

Model Description	Adj R^2	Intercept
12 linear variables, N=291	0.9098	Y .2383
12 linear variables, N=232	0.7462	No
8 linear variables, N=232	0.7506	Y .0001
23 variables, N=232	0.8267	No
The Final Model:		
9 variables, N=232	**0.8250**	**No**

Predictors: Benzene Load
Toluene Load
P-xylene Load
Benzene Load Squared
P-xylene Load Squared
Toluene by Benzene
P-xylene by Benzene
Reactor Temperature
Redox

Effects of:	Variables removed	New Adj R^2	P-value
Benzene	Benzene Load, P-xylene by Benzene Benzene Load Squared Toluene by Benzene	0.7638	0.0000
Toluene	Toluene Load, Toluene by Benzene	0.8143	0.0005
P-xylene	P-xylene Load, P-xylene by Benzene P-xylene Load Squared	0.7458	0.0000
Reactor Temperature		0.8116	0.0000
Redox Potential		0.7868	0.0000

Summary: Table 3 indicates that benzene load, toluene load, p-xylene load, reactor temperature, and redox potential all have a strong, statistically significant effect on benzene degradation. The aliphatic hydrocarbon loads have very little influence on the rate of benzene degradation. Also, it is interesting to note that removing all four variables containing benzene load only slightly decreased the R^2 value.

Table 4

Response: **TRIMETHYLPENTANE 2,2,4 DEGRADATION**

Model Description	Adj R^2	Intercept
12 linear variables, N=291	0.2695	Y .3524
12 linear variables, N=232	0.3196	Y .0772
7 linear variables, N=232	0.3233	Y .0001
23 variables, N=232	0.5545	Y .0008

The Final Model:

7 variables, N=232	**0.5597**	**Y .0001**

Predictors: Trimethylpentane Load
Toluene Load
Octane Load
Toluene Load Squared
Octane Load Squared
Reactor Temperature
Difference in Pressure

Effects of:	Variables removed	New Adj R^2	P-value
Toluene	Toluene Load, Toluene Load Squared	0.4484	0.0000
Octane	Octane Load, Octane Load Squared	0.5291	0.0002
Trimethylpentane Load		0.0251	0.0000
Reactor Temperature		0.4922	0.0000
Differential Pressure		0.5039	0.0000

Summary: According to Table 4, trimethylpentane load, toluene load, octane load, reactor temperature, and differential pressure all have strong, statistically significant effects on trimethylpentane degradation. Notice that trimethylpentane load is the most important explanatory variable in the model (see the adjusted-R^2 values).

Table 5

Response: **TOLUENE DEGRADATION**

Model Description	Adj R^2	Intercept
12 linear variables, N=291	0.9560	Y .0100
12 linear variables, N=232	0.9109	Y .0384
6 linear variables, N=232	0.9133	Y .0001
23 variables, N=232	0.9352	Y .0483

The Final Model:

7 variables, N=232	**0.9228**	**No**

Predictors: Toluene Load
P-xylene Load
Toluene Load Squared
P-xylene Load Squared
Differential Pressure
Redox Potential
Conductivity

Effects of:	Variables removed	New Adj R^2	P-value
Toluene	Toluene Load, Toluene Load Squared	0.8886	0.0000
P-xylene	P-xylene Load, P-xylene Load Squared	0.9167	0.0001
Differential Pressure		0.9169	0.0001
Redox Potential		0.9175	0.0001
Conductivity		0.8931	0.0000

Summary: Table 5 shows that toluene load, p-xylene load, differential pressure, redox potential, and conductivity all have strong, statistically significant effects on toluene degradation. The aliphatic hydrocarbon loads do not appear in this model. As with the benzene model, removing both toluene load terms does not drastically change the R^2 value.

Table 6

Response: **OCTANE DEGRADATION**

Model Description	Adj R^2	Intercept
12 linear variables, N=291	0.3085	No
12 linear variables, N=232	0.1429	No
5 linear variables, N=232	0.1568	Y .000
23 variables, N=232	0.2570	No
The Final Model:		
11 variables, N=232	**0.2559**	**No**

Predictors: Octane Load
P-xylene Load
Hexane Load Squared
Toluene Load Squared
Octane Load Squared
P-xylene Load Squared
Octane by Hexane
Octane by Toluene
Influent Humidity
Reactor Temperature
Redox

Effects of:	Variables removed	New Adj R^2	P-value
Hexane	Hexane Load Squared Octane by Hexane	0.2136	0.0008
Toluene	Toluene Load Squared Octane by Toluene	0.1392	0.0000
Octane	Octane Load, Octane Load Squared Octane by Hexane, Octane by Toluene	0.0003	0.0000
P-xylene	P-xylene Load, P-xylene Load Squared	0.2133	0.0008
Influent Humidity		0.1752	0.0000
Reactor Temperature		0.2035	0.0001
Redox Potential		0.1811	0.0000

Summary: The information in Table 6 implies that hexane load, toluene load, octane load, p-xylene load, influent humidity, reactor temperature, and redox potential have strong, statistically significant effects on octane degradation. Octane load was the most important predictor in the model, followed by toluene load (see adjusted-R^2 values).

Table 7

Response: **P-XYLENE DEGRADATION**

Model Description	Adj R^2	Intercept
12 linear variables, N=291	0.8235	No
12 linear variables, N=250	0.6392	No
7 linear variables, N=250	0.6393	Y .0433
23 variables, N=250	0.6890	No

The Final Model:

13 variables, N=250	**0.6920**	**No**
Predictors: Hexane Load	**P-xylene by Hexane**	
Toluene Load	**P-xylene by Trimethylpentane**	
Hexane Load Squared	**P-xylene by Toluene**	
Trimethylpentane Load Squared	**Influent Humidity**	
Toluene Load Squared	**Reactor Temperature**	
Octane Load Squared	**Redox**	
P-xylene Load Squared		

Effects of:	Variables removed	New Adj R^2	P-value
Hexane	Hexane Load, Hexane by P-xylene Hexane Load Squared	0.6313	0.0000
TMP 224	Trimethylpentane Load Squared P-xylene by Trimethylpentane	0.6561	0.0000
Toluene	Toluene Load, P-xylene by Toluene Toluene Load Squared	0.6365	0.0000
P-xylene	P-xylene Load Squared P-xylene by Hexane, P-xylene by Toluene P-xylene by Trimethylpentane	0.6327	0.0000
Octane Load Squared		0.6816	0.0030
Influent Humidity		0.6864	0.0213
Reactor Temperature		0.5906	0.0000
Redox Potential		0.6684	0.0000

Summary: In Table 7, hexane load, trimethylpentane load, toluene load, p-xylene load, octane load, reactor temperature, and redox potential have strong, statistically significant effects on p-xylene degradation. Benzene load's effect is not significant at the 0.01 level. P-xylene load is no more important to the prediction of p-xylene degradation than hexane load or toluene load.

6.3 Diagnostic plots

Figures 8a and 8b contain five types of diagnostic plots for each of the six hydrocarbon degradation models: observed degradation rates versus predicted degradation rates; predicted values versus residuals; and each of the residuals, observed values and predicted values versus observation number.

The observed-by-predicted plots show how well a model predicts the actual values. The ideal plot would show points along the line of equality (this line is included in these plots), indicating a correlation near one. A high correlation indicates that a model fits the data fairly well. In some of the plots, a pattern of points falling above the line for the smaller values and below the line for the larger values indicates that our model over-predicts the smaller degradation rates and under-predicts the larger prediction rates. This pattern is particularly obvious for the aliphatic hydrocarbon degradations. The aromatic degradation models tend to have a stronger linear relationship between the predicted and observed degradation values than the aliphatic models.

The residual-by-predicted value and residual-by-observation number plots should appear fairly random. Some of these plots tend to show slight patterns that may indicate that homogeneity of variance assumptions necessary to construct linear models may be violated. However, there are no strong, obvious patterns evident. Also, I chose not to delete any points as outliers since I had little opportunity to consult with the client on the consequences of deleting particular points. There did not seem to be any unusually large outliers to be really concerned about.

The observed-by-observation number plots and the predicted-by-observation number plots can be compared to see how closely the predicted values follow the observed values. Also, some of these plots contain curvature that resulted from turning on and off the various loads. The aromatic hydrocarbon plots, in particular, display this pattern as well as a consistent decreasing trend in the degradation rates over time. The observed and predicted values over time for the aliphatic degradation plots appear more scattered and random than the aromatic degradation plots, but the they also have a smaller scale along the vertical axis.

6.4 Total Hydrocarbon Degradation Models

A new response variable, "total hydrocarbon degradation," was formed by adding the degradation rates of all six hydrocarbons. To model total hydrocarbon degradation rate, two models were used: one reduced from 33 initial variables, and one reduced from 11 grouped variables. By grouping the hydrocarbons into their two separate classes, total aliphatic and aromatic degradation rates were modeled as well. The 11 grouped variables were as follows:

The six original physical traits plus,
Aliphatic Load (Hexane+TMP+Octane)
Aromatic Load (Benzene+Toluene+P-xylene)
Aliphatic Load Squared
Aromatic Load Squared
Aliphatic by Aromatic Interaction

Below are the tables for models of total hydrocarbon (Table 8), aromatic (Table 9) and aliphatic (Table 10) degradation rates.

Table 8

Response: **TOTAL HYDROCARBON DEGRADATION**

Model Description	Adj R^2	Intercept
12 variables (of 33), N=291	**0.9862**	**Y .0001**

Predictors:
- Hexane Load
- Trimethylpentane Load
- Toluene Load
- P-xylene Load
- Toluene Load Squared
- Octane Load Squared
- P-xylene Load Squared
- Toluene by Benzene
- Influent Humidity
- Reactor Temperature
- Redox Potential
- Conductivity

8 grouped variables (of 11), N=291	**0.9826**	**Y .0001**

Predictors:
- Aliphatic Load
- Aromatic Load
- Aliphatic by Aromatic
- Influent Humidity
- Reactor Temperature
- Differential Pressure
- Redox Potential
- Conductivity

Table 9

Response: **TOTAL AROMATIC DEGRADATION**

Model Description	Adj R^2	Intercept
14 variables (of 33), N=262	**0.9516**	**Y .0002**
Predictors: Hexane Load		
Benzene Load		
Trimethylpentane Load		
Octane Load		
P-xylene Load		
Hexane Load Squared		
Trimethylpentane Load Squared		
Toluene Load Squared		
Octane Load Squared		
Toluene by Benzene		
Toluene by Trimethylpentane		
Differential Pressure		
Redox Potential		
Conductivity		
6 grouped variables (of 11), N=262	**0.9350**	**Y .0001**
Predictors: Aromatic Load		
Aliphatic by Aromatic		
Influent Humidity		
Reactor Temperature		
Redox Potential		
Conductivity		

Table 10

Response: **ALIPHATIC DEGRADATION**

Model Description	Adj R^2	Intercept
12 variables (of 33), N=280	**0.4472**	**Y .0026**
Predictors: Hexane Load		
Benzene Load		
Trimethylpentane Load		
Octane Load		
P-xylene Load		
Hexane Load Squared		
Trimethylpentane Load Squared		
Toluene Load Squared		
P-xylene Load Squared		
Toluene by Trimethylpentane		
Octane by Benzene		
P-xylene by Hexane		
3 grouped variables (of 11), N=280	**0.1760**	**Y .0024**
Predictors: Aliphatic Load		
Redox Potential		
Conductivity		

6.5 CO_2 Production Models

When the hydrocarbons are being degraded by the biofilm, CO_2 is produced. The amount of CO_2 in the reactor was measured throughout the experiment. The CO_2 production rate was modeled in the same manner as total hydrocarbon degradation, first reducing from a 33 variable model and then an 11 grouped variable model.

Table 11

Response: **CO_2 PRODUCTION**

Model Description	Adj R^2	Intercept
13 variables (of 33), N=291	**0.7385**	**Y .0194**
Predictors: Benzene Load		
Trimethylpentane Load		
Toluene Load		
Benzene Load Squared		
Toluene Load Squared		
Octane Load Squared		
P-xylene Load Squared		
Toluene by Benzene		
Toluene by Trimethylpentane		
Reactor Temperature		
Differential Pressure		
pH		
Redox Potential		
6 grouped variables (of 11), N=291	**0.6887**	**No**
Predictors: Aromatic Load		
Aliphatic Load		
Influent Humidity		
Reactor Temperature		
Differential Pressure		
Redox Potential		

No patterns in the chosen explanatory variables are evident, other than there are more aromatic hydrocarbons included than aliphatic hydrocarbons.

7. CONCLUSION

7.1 Client Conclusions

A regression modelling study was conducted to determine the effects of the six hydrocarbon loading rates on the degradation rates of those same hydrocarbons. The data were taken from a vapor-phase biofilm reactor. Plots of the hydrocarbon degradation rates by the loading rates indicated that some explanatory terms needed to added to the model quadratically, rather than linearly. The addition of quadratic and interaction terms improved the fit of the models tremendously, compared to the original linear models. The analysis used only the most informative subset of the data, period C.

Unremovable multicollinearity was present in the data. A high degree of multicollinearity does not allow reliable interpretation of the models phenomenologically, such as estimation of partial correlation and multiple regression coefficients. Nevertheless, we could arrive at qualitative conclusions by examining the effect of a hydrocarbon loading rate on the ability to predict a degradation rate. The regression models can be used to predict degradation rates for various loading patterns within the range of loadings used in the experiment. Adding other explanatory variables to the final model does not improve the model; that is, adding other explanatory variables increases negligibly the adjusted-R^2. We cannot assume that the explanatory variables not in the final model do not effect the degradation response. We can only assume, that GIVEN the variables already in the final model, the other variables not in the model do not improve our prediction capabilities.

The aromatic hydrocarbons (benzene, toluene, and p-xylene) were easier to model than the aliphatics (hexane, trimethylpentane, and octane). Benzene and toluene degradation rates were easiest to model, and their models yielded the highest adjusted-R^2 values (0.83 and 0.92 respectively). The largest adjusted-R^2 for an aliphatic hydrocarbon degradation model was 0.56.

There are several results on how different loadings effected individual hydrocarbon degradations.

- In general, the individual aliphatic hydrocarbon loading rates had little influence on an aromatic hydrocarbon degradation rate. Aliphatics do not appear at all in the models for benzene and toluene.
- The intercept term was zero in five of the six individual hydrocarbon models (the exception being trimethylpentane).
- Toluene loading had a significant effect on the degradation rate of each hydrocarbon.
- Hexane load, trimethylpentane load, and toluene load each effect p-xylene degradation as much as does p-xylene load.
- Only nine of the total fifteen interaction terms appeared in the six hydrocarbon degradation models, and specific interaction terms appeared in no more than two models.

PH balance was the only measured physical characteristic that did not appear in any of the models. There was no pattern as to which models each of the other five physical characteristics appeared in. Reactor temperature appeared in five of the models (excluding the toluene model), while redox potential was present in four of the models.

For the total hydrocarbon degradation model, no zero load cases were deleted (N=291) since there was always at least one hydrocarbon being entered into the system. For the aliphatic or aromatic degradation models, cases were excluded if the total aliphatic or total aromatic loading rate was zero. This action reduced the data used in these models to 280 and 262 cases, respectively. When the degradation rates and loading rates were separated into two groups, aromatics and aliphatics, the aromatic degradation model did not depend on the aliphatic load rates. When the two groups were combined to form total hydrocarbon degradation, the associated model was very similar to the aromatic model. More aromatics were degraded in the system than aliphatics. Of the total hydrocarbon degraded, 85% to 99% is aromatic degradation. A

logical explanation for this difference in degradation rates stems from the fact that the reactor system contains water and the aromatics are water soluble while the aliphatics are not.

The adjusted-R^2 values are large for the total hydrocarbon and the aromatic hydrocarbon degradation models compared to the aliphatic hydrocarbons. For the aromatic and total hydrocarbon degradation models, the adjusted-R^2 is greater than 0.9, indicating that most of the variation in the degradation rates is explained by the models. The models using the grouped variables have comparable adjusted-R^2 values, so it is just as appropriate to use the grouped variable models when predicting degradation rates for aromatics or total hydrocarbons. The grouped models will be easier to use since the equations are simpler and the explanatory variables are easy to understand.

For the aliphatic degradation models, the adjusted-R^2s are much lower, particularly for the grouped model (adjusted-R^2=0.18). Again, this is due to a small range of degradation rates for these compounds because they are not water soluble. The grouped data model for aliphatic degradation is not appropriate for prediction purposes.

The CO_2 production model contained thirteen variables, six of which are benzene or toluene load terms, and had an adjusted-R^2 of 0.74. I did not research CO_2 production any further other than developing a model because I did not know how to interrupt the results. The interpretation of the model results were left entirely to the client.

7.2 Statistical Conclusions

The first step in any data analysis project should involve plotting the data. By plotting the data, I realized that quadratic explanatory variables needed to be entered into the models. I was also able to see the problems with multicollinearity in the explanatory variables. Because of the multicollinearity, a

quantitative analysis, such as partial correlations, regression coefficients, etc., cannot be accomplished. The following example illustrates this problem.

Table 12

Two Models for Toluene Degradation:

Model 1: 10 variables and an intercept term Adjusted-R^2=.9106

Model 2: 12 variables and an intercept term Adjusted-R^2=.9172

Predictor:	Coefficient in Model 1:	Coefficient in Model 2:
Intercept Term	2.3824	2.0337
Hexane Load	NA	-0.2348
Toluene by Hexane	NA	0.0391
Benzene Load	-2.0549	-1.9298
Toluene Load	-1.0259	-0.9785
Octane Load	-0.6915	-0.6808
P-xylene Load	**0.0263**	**-0.1764**
Toluene by Benzene	0.9260	0.8754
Toluene by TMP	**0.0153**	**-0.0023**
Octane by Toluene	0.3046	0.2997
P-xylene by Toluene	**-0.0340**	**0.0481**
Redox Potential	0.0026	0.0013
Conductivity	0.0015	0.0022

Table 12 shows two different models for toluene degradation with nearly the same adjusted-R^2 values. Model 2 uses the same explanatory variables as Model 1 and includes hexane load and the interaction, toluene by hexane. The highlighted lines indicate variables for which the regression coefficient changes sign when the two hexane variables were added. This illustrates that, when there is multicollinearity, the slightest change in a model can reverse the signs and drastically change the values of some regression coefficients. The partial correlation coefficient is similarly sensitive. This example shows why we cannot quantitatively analyze the coefficients of the explanatory variables, and hence we cannot evaluate partial correlation coefficients. However, the final model can be used for prediction purposes (as long as one does not extrapolate away from the loadings and physical characteristics of this experiment). Also, a qualitative assessment is possible: if there is a qualitative drop in prediction accuracy, as measured by adjusted-R^2,

when an explanatory variable is removed from the final model, then that variable is important.

Zero loading rates for each degradation rate were removed from the data to assure us that the models were applicable to relevant positive loadings. Throwing out the zero-loading cases reduced the adjusted-R^2 values by as much as two tenths in some cases. This was due to the fact that the cases at the point zero load, zero degradation were highly influential to the regression. When modeling a cluster of points, it is difficult to fit a line through the cluster, but if several cases occur at one point outside of the cluster (specifically, the zero loading and zero degradation point), the regression line will go through this point. It is somewhat surprising, therefore, that the adjusted-R^2 for trimethylpentane degradation increased when we removed the cases where the trimethylpentane load equalled zero.

The data contain negative degradation rates, which we were having trouble interpreting. Since degradation rates are a measure of the difference between the amount of a hydrocarbon going into the system and the amount coming out, negative rates appear to result from transient peaks of hydrocarbon leaving the system. Further, negative rates are attributable to inherent variability in the methods of chemical analysis. For example, if hexane load were terminated, and significant quantities of hexane remained in the reactor, whether in gas phase, dissolved in the liquid, or absorbed by the reactor contents, then measurable hexane might still be present in the reactor effluent for some time. Hence, a negative degradation rate (influent minus effluent rates) would result. It is inconceivable that hexane would actually be produced inside the reactor. Some data entries, particularly for toluene degradation, appear to be impossible as degradation rates exceed loading rates. Finally, comparison of the data columns labelled "PPMV_C" and "umoles-C/M2*DAY" point to the use of a conversion factor which was not constant throughout the experiment.

In order to develop regression models in the first place, two main assumptions need to be upheld. The first is homogeneity of variance in the errors of each model. To evaluate this assumption, residuals can be plotted by time and by predicted values. If patterns appear in the plots then the homogeneity assumption

is not appropriate, and it would be best to look at a possible transformation, or maybe generalized linear models. To review these plots, see Figures 7a and 7b. The second assumption is that the errors are normally distributed. To evaluate this assumption, a normal probability plot of the residuals is effective. If the residuals appear fairly linear on the plot, then the normality assumption is acceptable. See Figure 9 for the normal probability plots of the single hydrocarbon degradation models. These plots appear slightly heavy in the tails which may indicate that the normality assumption is not upheld. Dr. Hamilton and I recognized this problem, but due to time constraints, we chose not to pursue methods other than ordinary least squares procedures.

7.3 Extensions

The main extension would be to rerun the experiment without the multicollinearity in the explanatory variables. Experiments such as this can be designed to prevent multicollinearity among explanatory variables. One class of designs is called the Box-Behnken design (see Box and Draper, 1987). If one were planning a new experiment with six hydrocarbons, the standard Box-Behnken design would require only 54 runs, or cases. Figure 10 shows the structure of this design.

A principal component analysis of the variables would be appropriate and it would eliminate the multicollinearity problems. However, the resulting models would be difficult to analyze in terms of the hydrocarbon loading variables so the Box-Behnken design is a solution to the multicollinearity problems that is more beneficial to the client.

Transformations, generalized linear models, or robust estimation procedures may help eliminate problems with the original model assumptions. Nonadditive models or data smoothing would be other options to use in working with this data.

Because of the range of the loading rates (between zero and five units), the linear and quadratic terms for each loading are collinear. A reanalysis using centered and scaled predictors would eliminate within-loading collinearity by creating linear and quadratic predictors that have small (near zero) correlation.

I cannot guarantee that a reanalysis would result in significantly improved models, but it may be worth future considerations.

My advisor, Dr. John Borkowski, mentioned that using CuSum charts on the residuals would be another way to test the residuals for independence over time and homogeneity assumptions. Given more time, this would probably be my next step since these graphs can be very informative of the overall process.

Figure 1: Linear Model For Hexane Degradation

Model fitting results for: B:GCSUMC.HEXDEG

Independent variable	coefficient	std. error	t-value	sig.level
B:GCSUMC.HEXLOAD	0.130607	0.01838	7.1061	0.0000
B:GCSUMC.BNZLOAD	0.016044	0.017583	0.9125	0.3623
B:GCSUMC.TMPLOAD	0.016314	0.019216	0.8490	0.3966
B:GCSUMC.TOLLOAD	-0.005143	0.013375	-0.3845	0.7009
B:GCSUMC.OCTLOAD	-0.009644	0.013711	-0.7034	0.4824
B:GCSUMC.PXLLOAD	0.003231	0.010886	0.2968	0.7669
B:GCSUMC.INFHUM	0.003699	0.001439	2.5713	0.0107
B:GCSUMC.RTEMP	-0.000412	0.005889	-0.0700	0.9442
B:GCSUMC.DIFFPRESS	-0.020244	0.047913	-0.4225	0.6730
B:GCSUMC.pH	0.021146	0.028374	0.7453	0.4567
B:GCSUMC.REDOX	0.00051	0.000252	2.0239	0.0439
B:GCSUMC.CONDUCT	-0.000384	0.000164	-2.3387	0.0201

R-SQ. (ADJ.) = 0.6865 SE= 0.102313 MAE= 0.072058 DurbWat= 1.453
Previously: 0.8653 0.075859 0.041274 2.810
290 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

Figure 2: Partial Correlations for Hexane Degradation

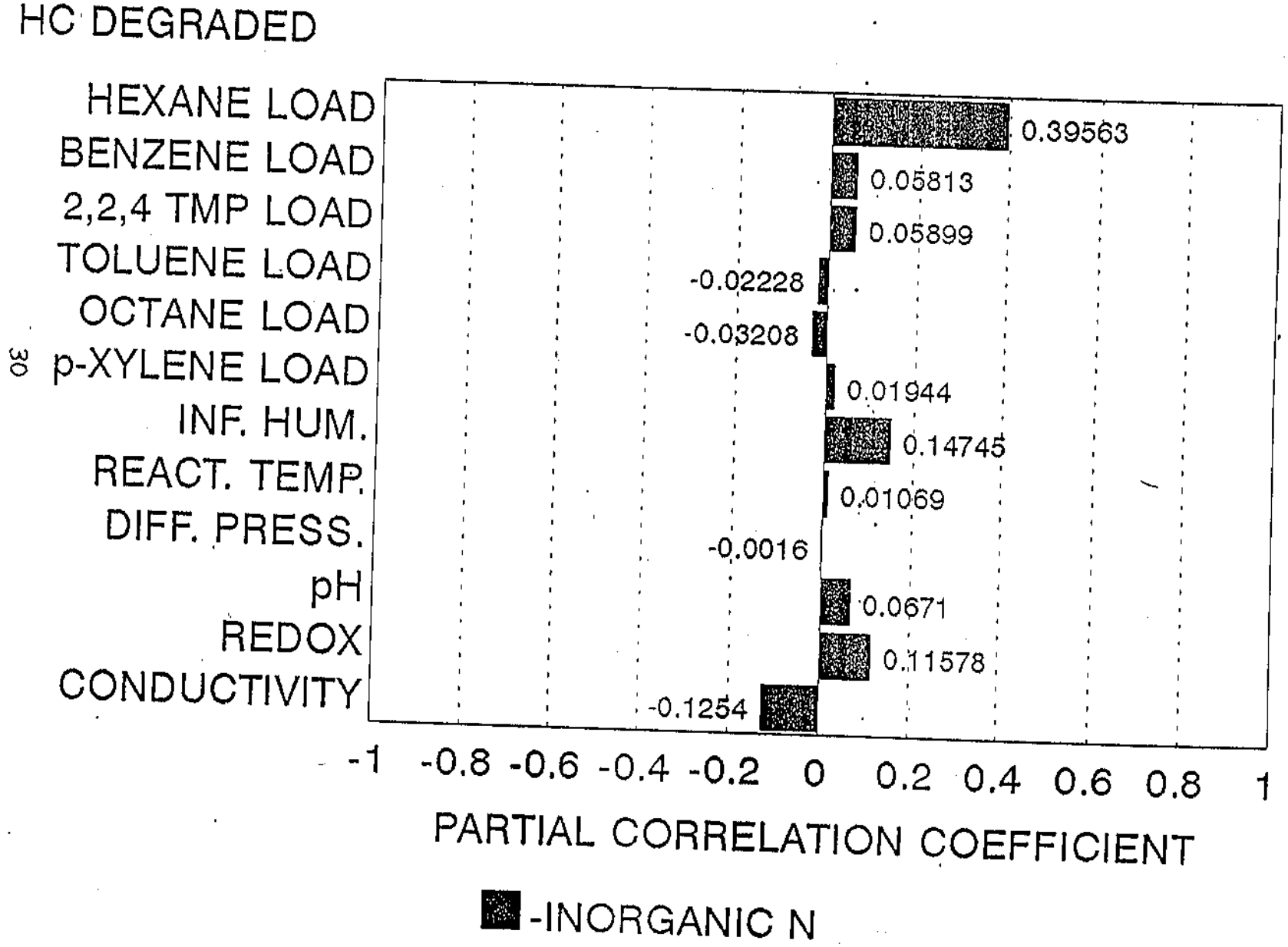


Figure 3: Scatterplots of Loads for Period A

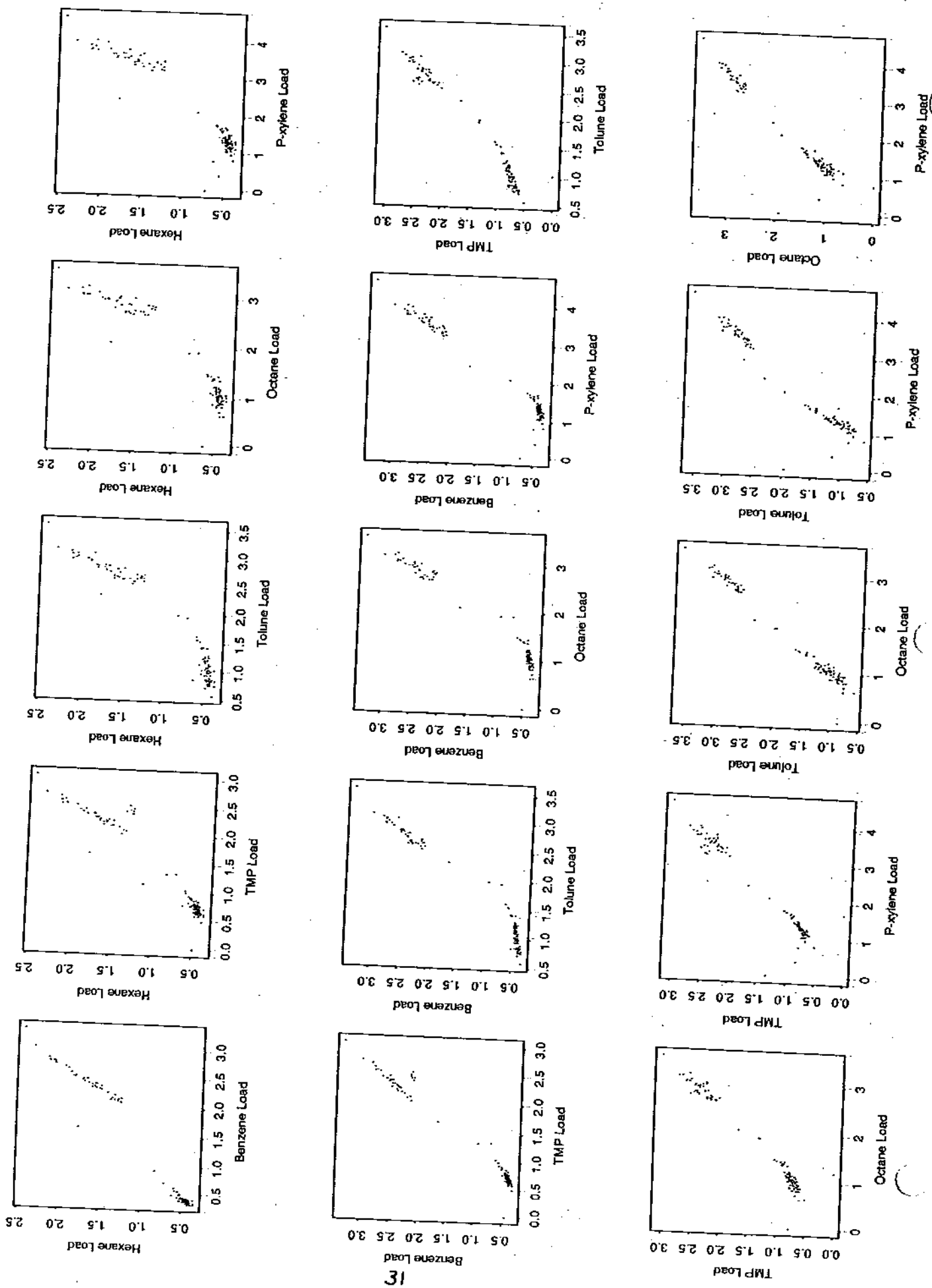


Figure 4: Scatterplots of Loads for Period B

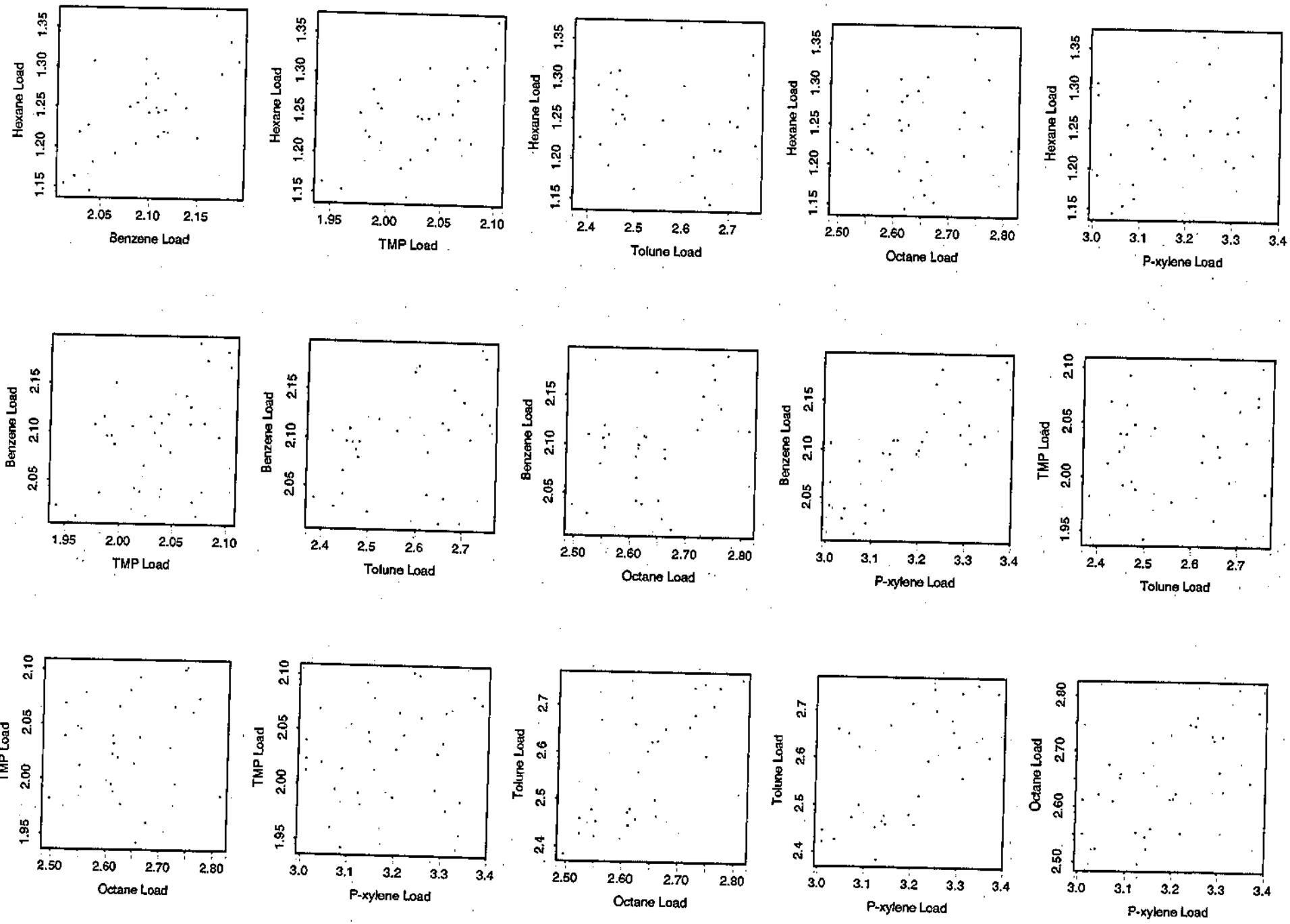


Figure 5a: Scatterplots of Loads for Period C

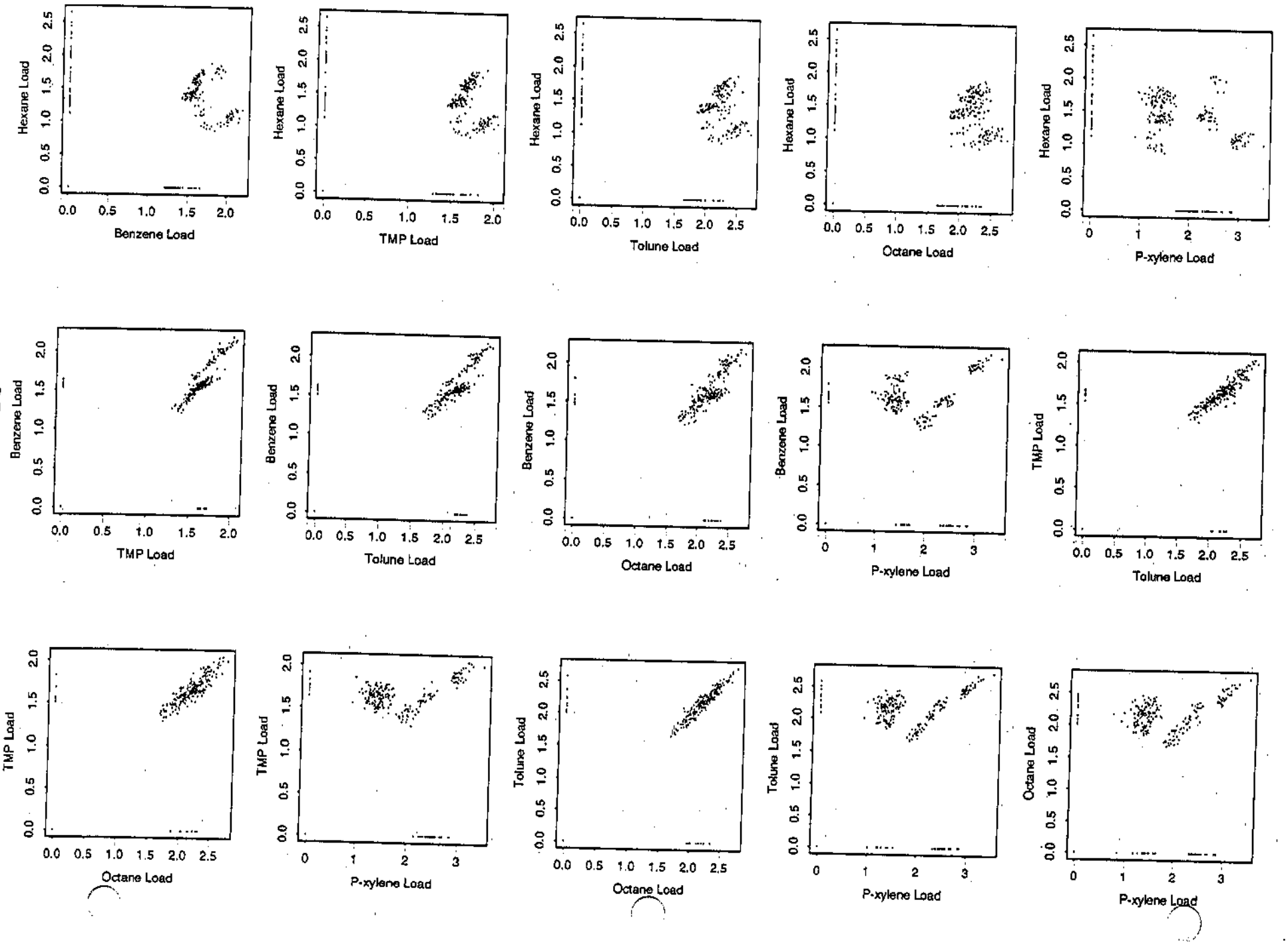


Figure 5b: Scatterplots of Degradations by Loads for Period C

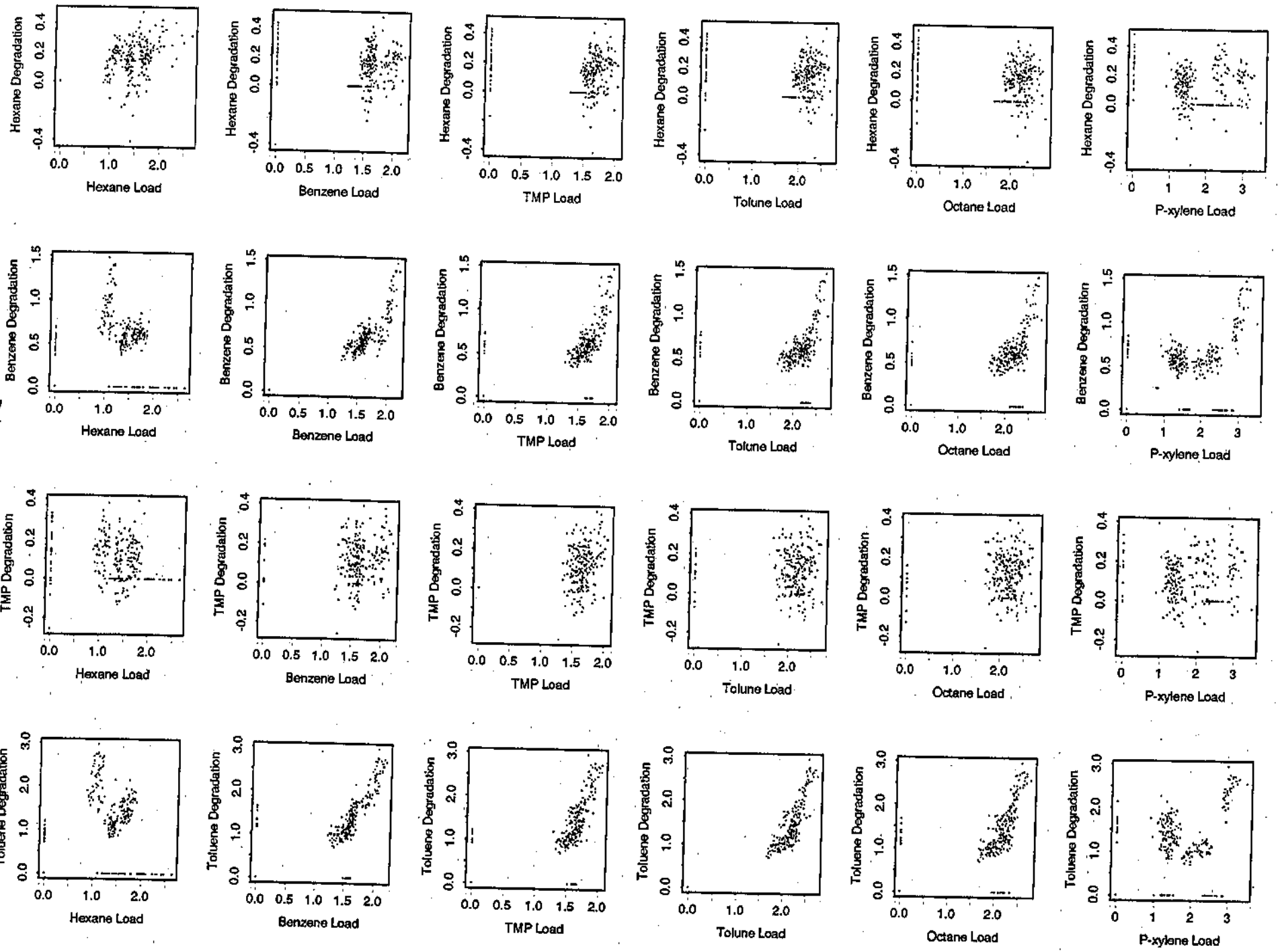


Figure 5c: Scatterplots of Degradations by Loads for Period C

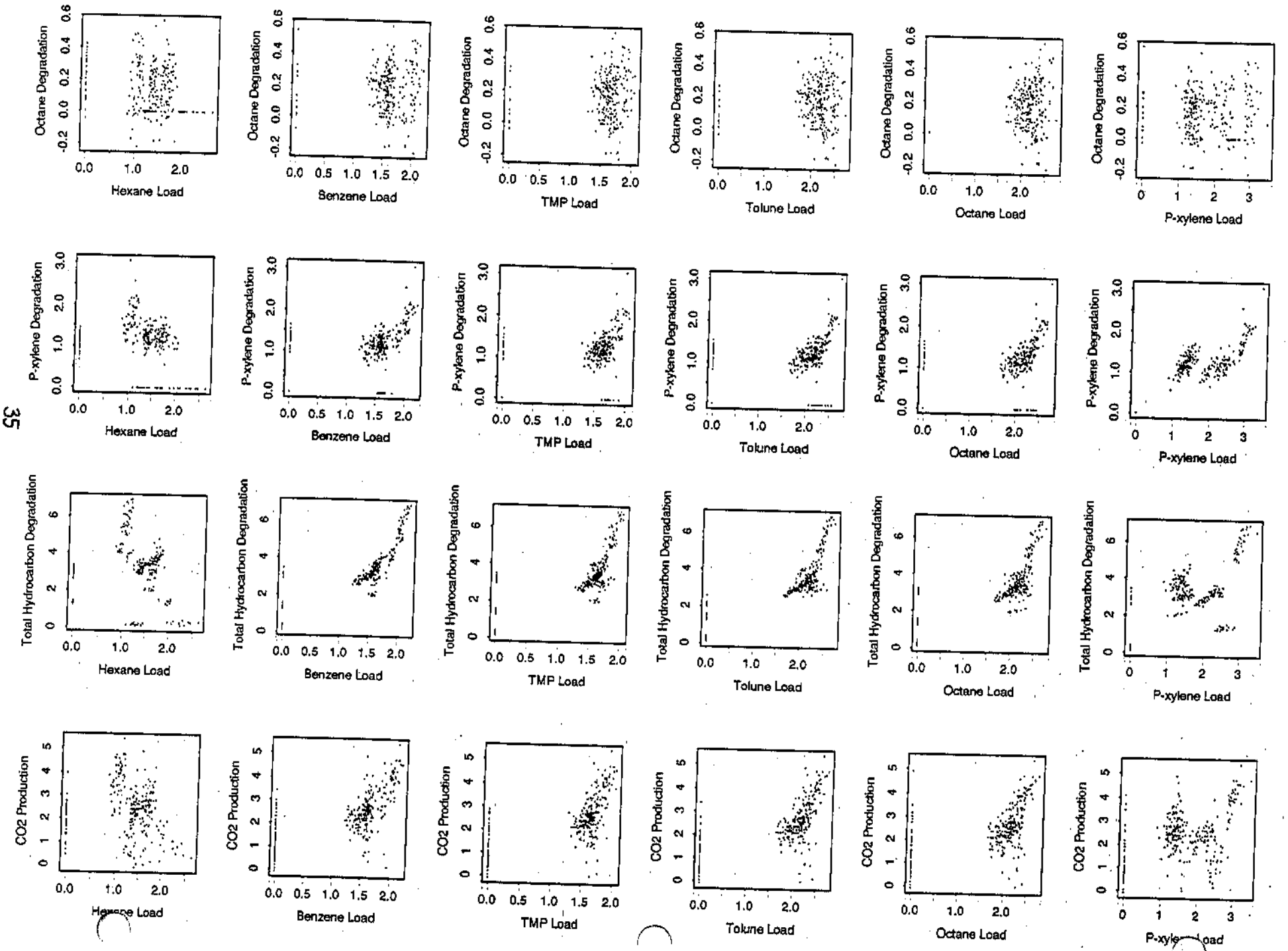
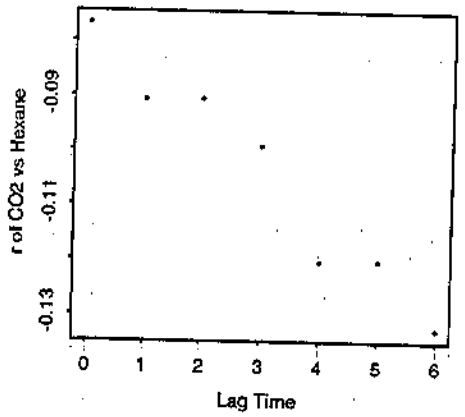
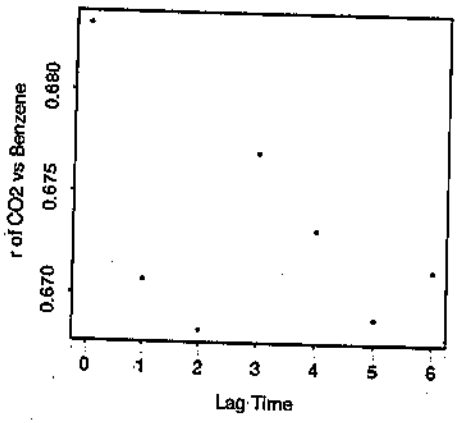
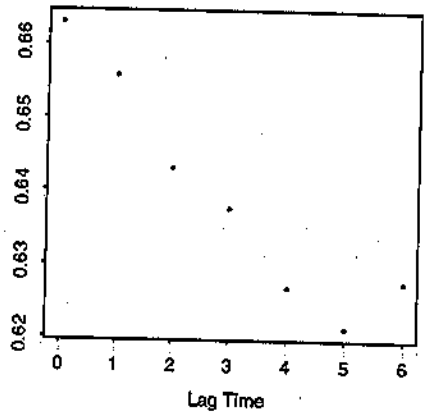
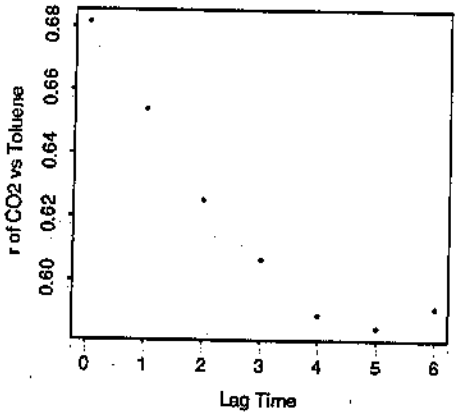


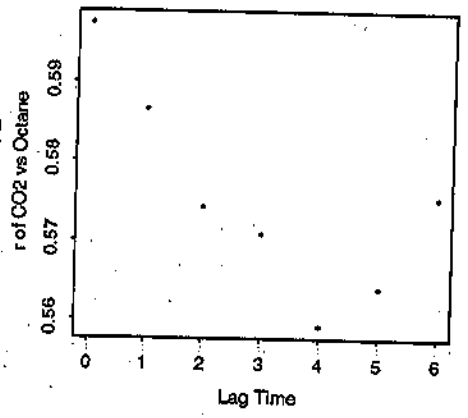
Figure 6: Correlations of CO2 and Predictors by Lagged Time

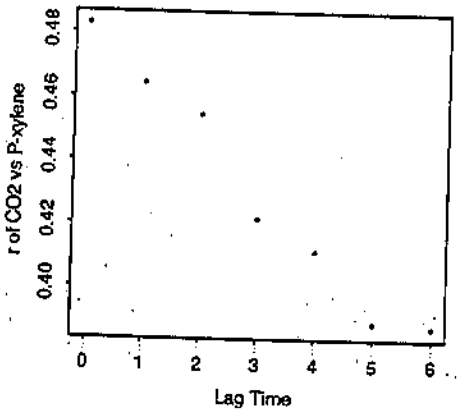


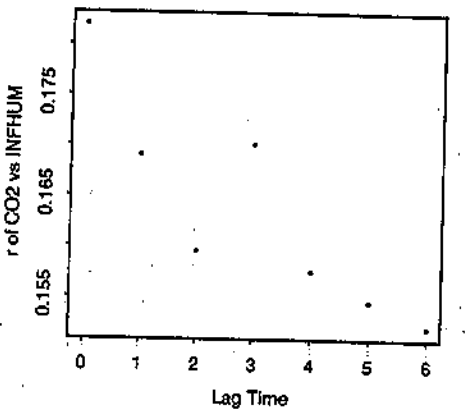


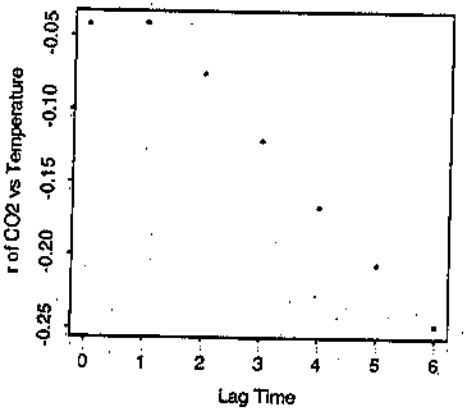


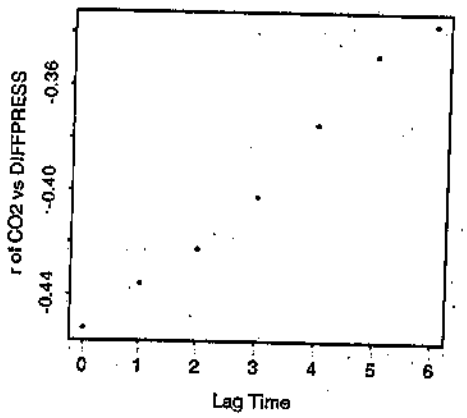


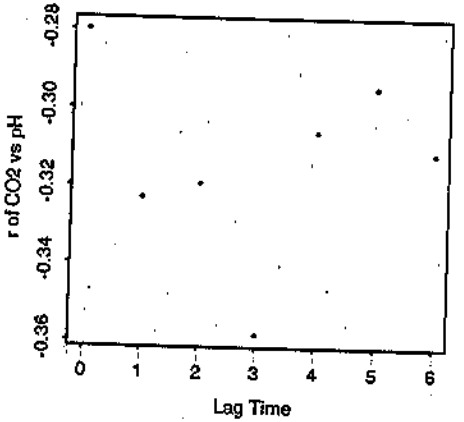


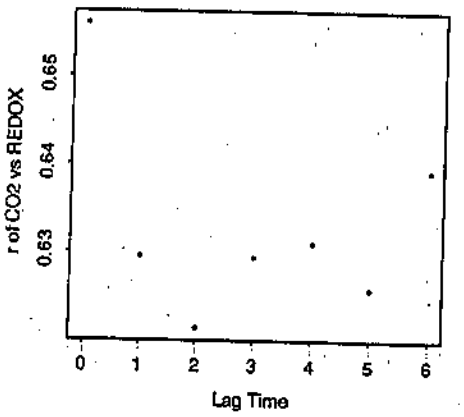












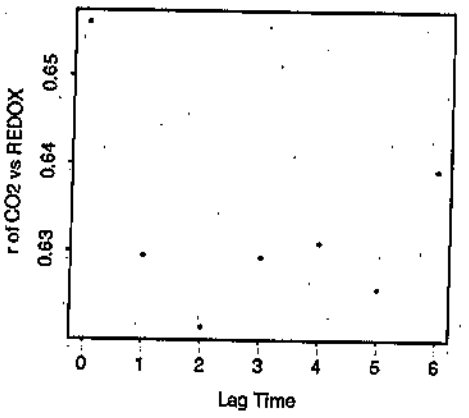


Figure 7a: Adjusted R^2 for Different Model Sizes
Benzene Degradation

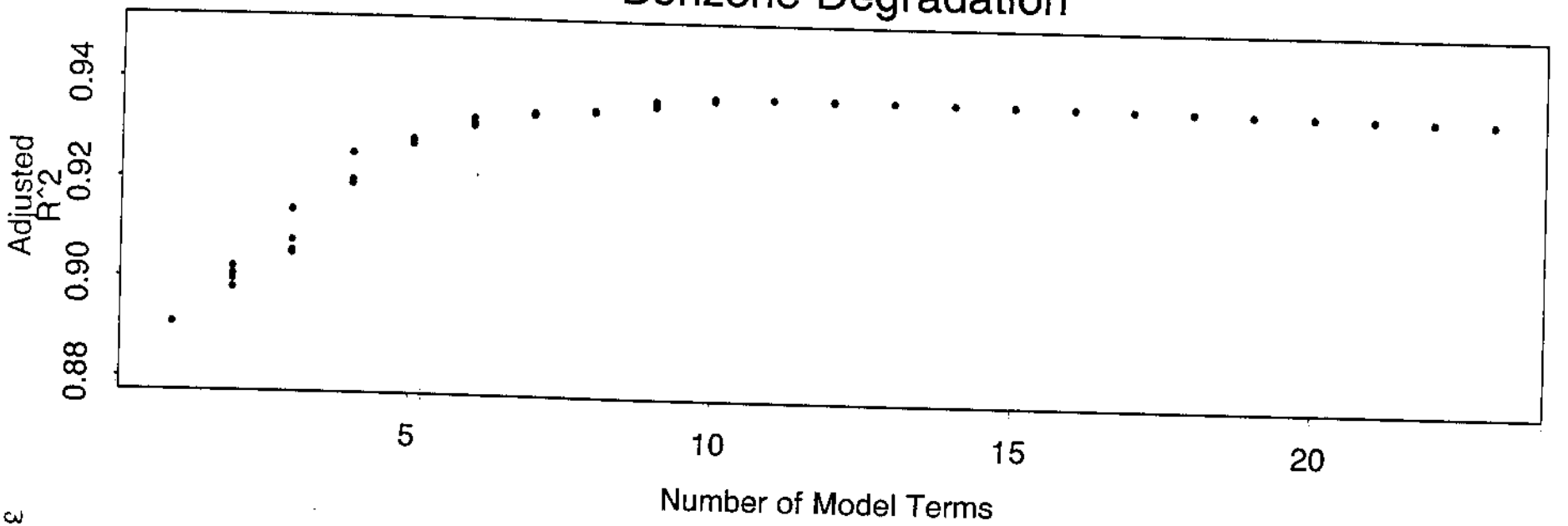


Figure 7b: Mallow's Cp for Different Model Sizes
Benzene Degradation

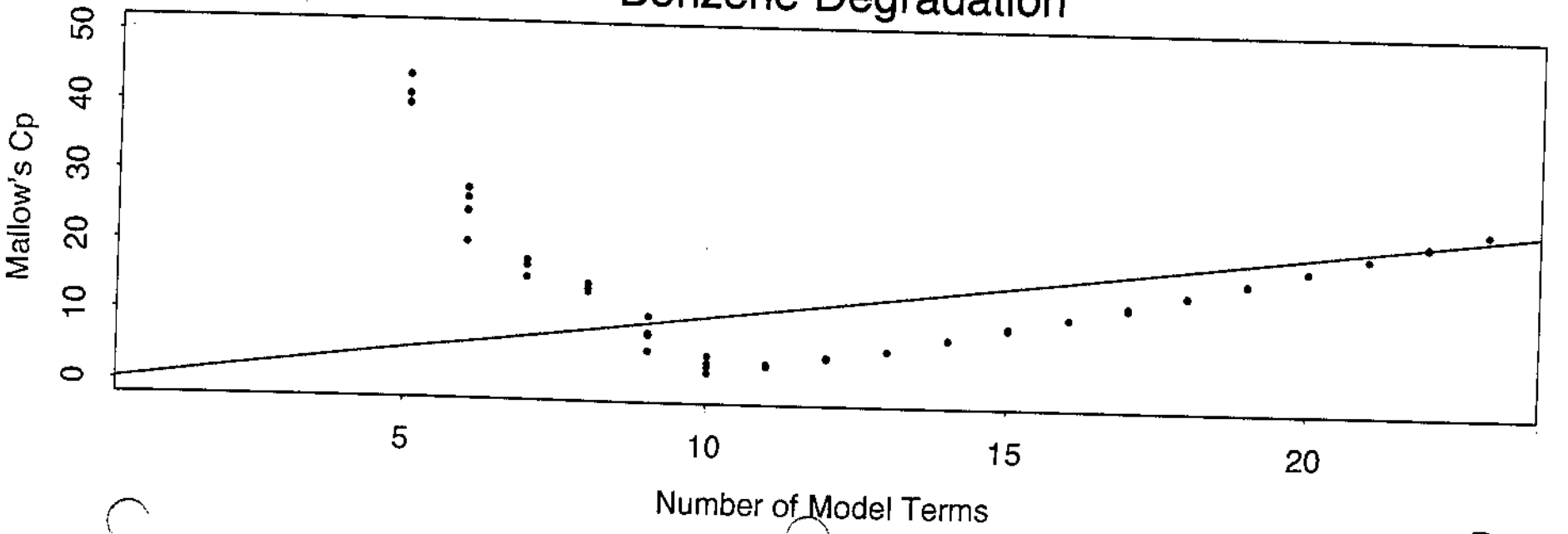
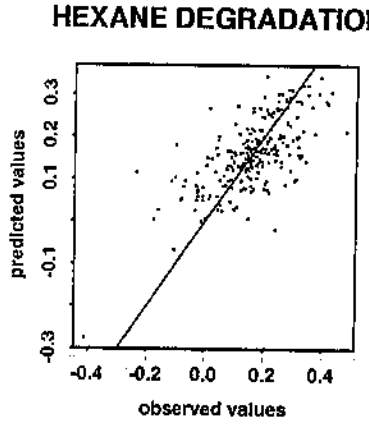
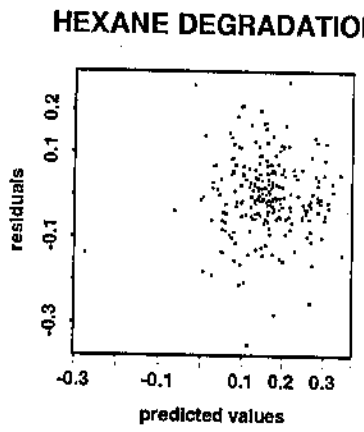
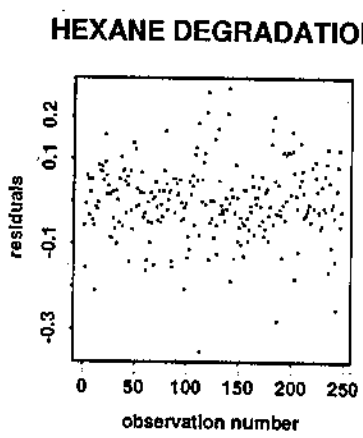
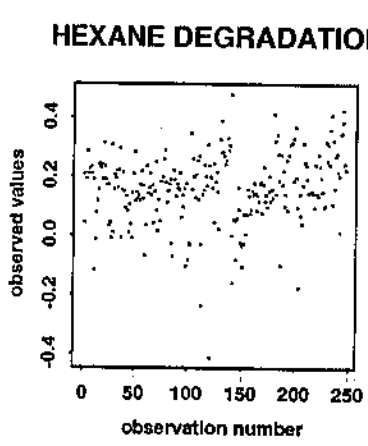


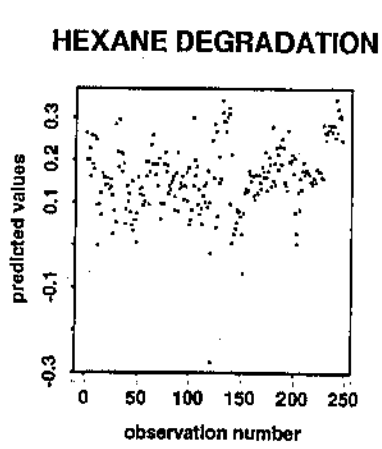
Figure 8a: Residual Plots of Hydrocarbon Models

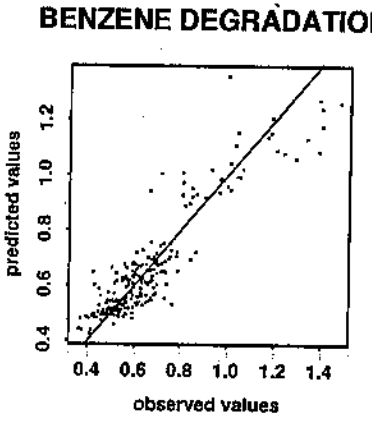


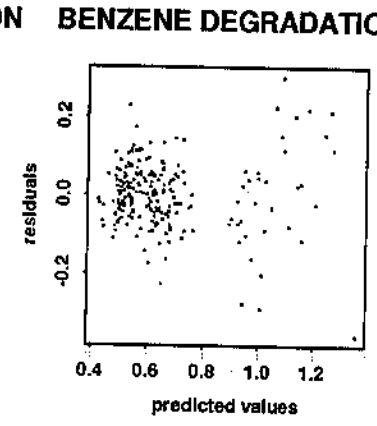


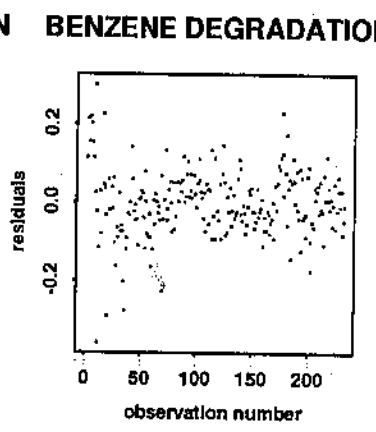


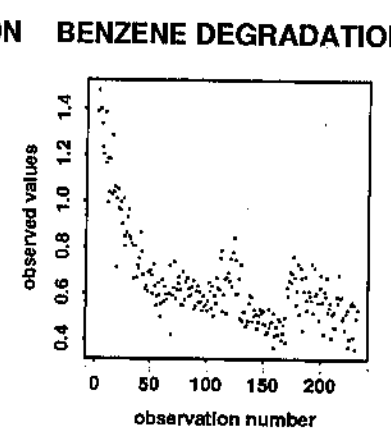


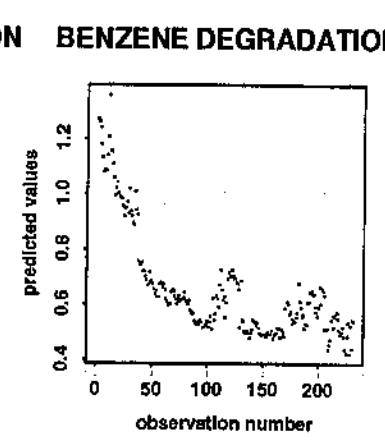


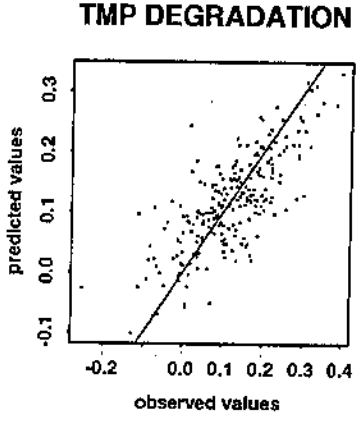


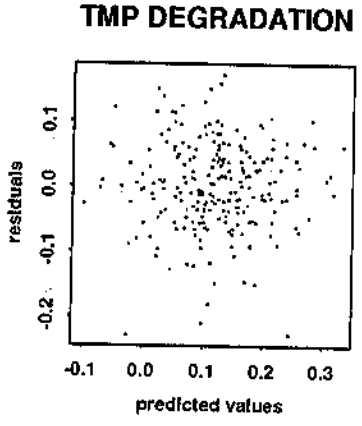


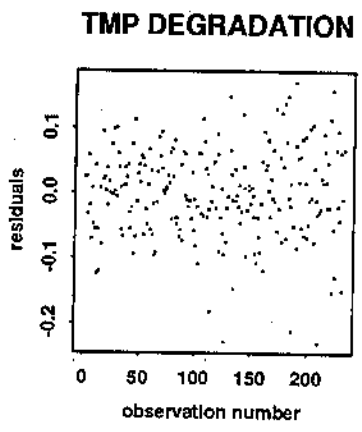


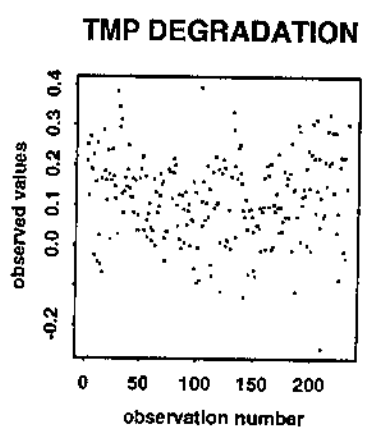












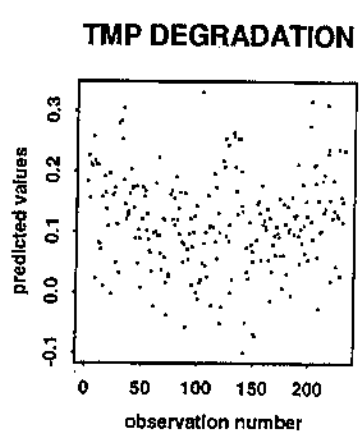


Figure 8b: Residual Plots of Hydrocarbon Models

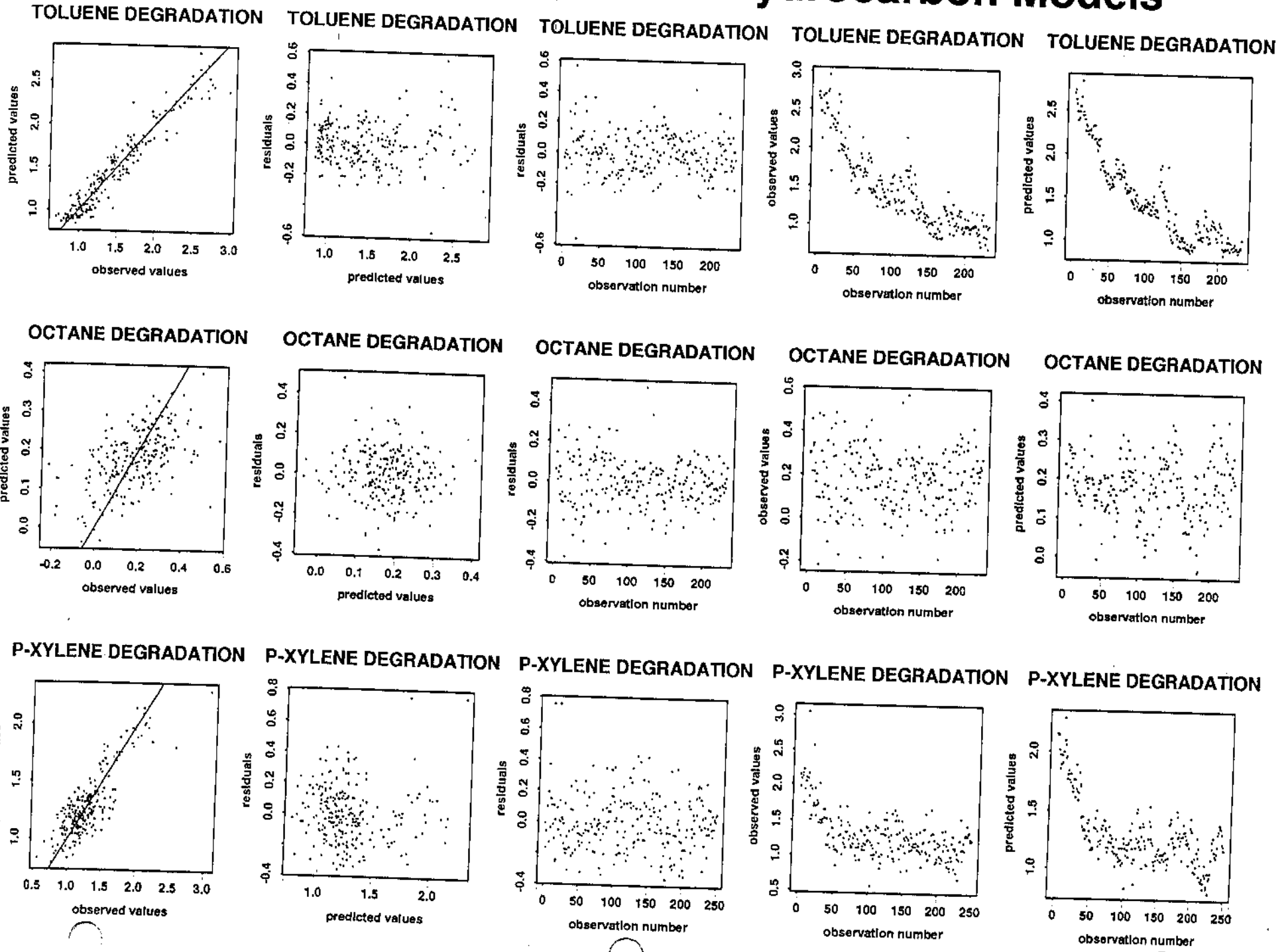


Figure 9: Normal probability plots for the six hydrocarbons

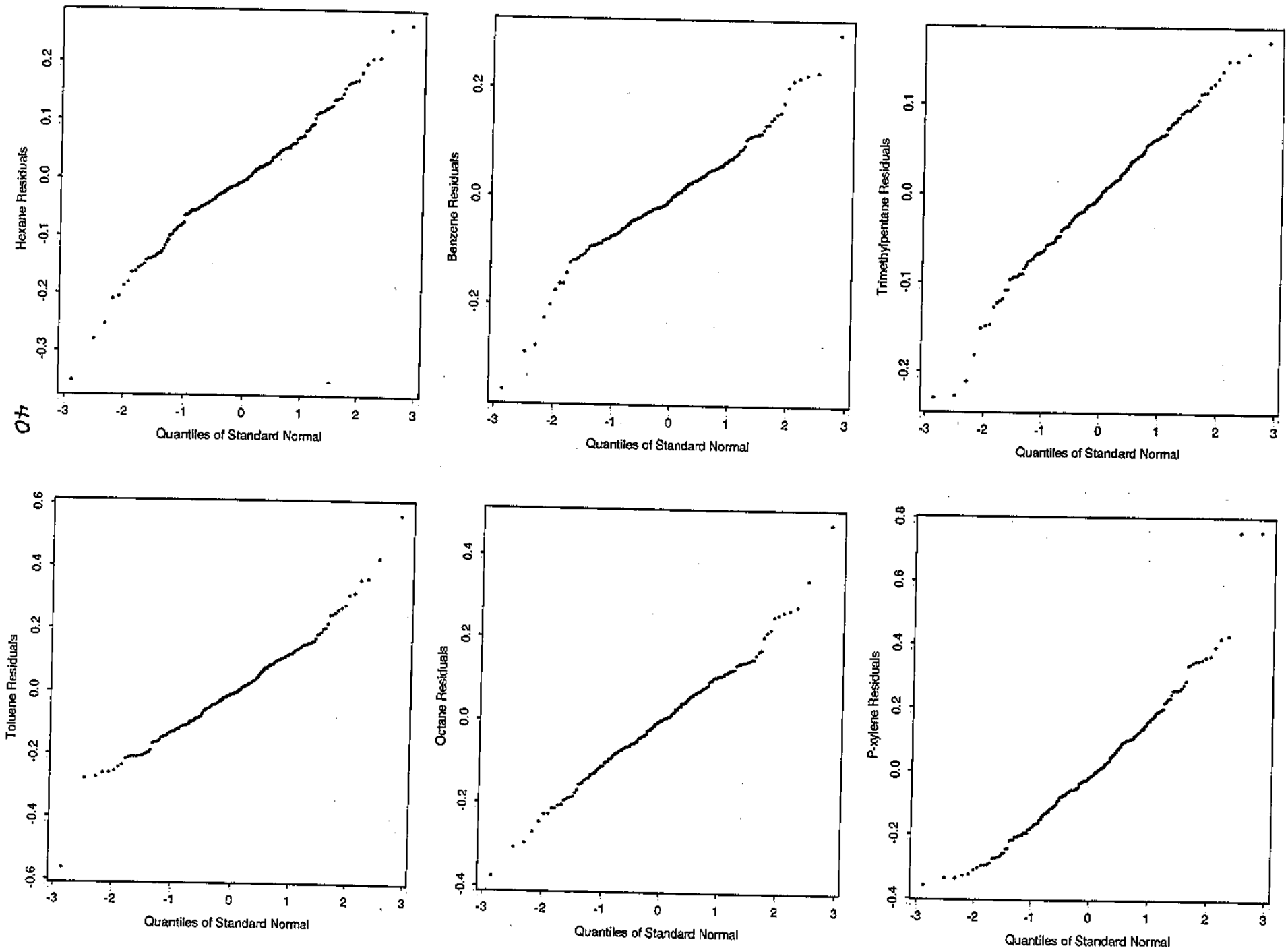


Figure 10: The Box-Behnken Design

This is a Box-Behnken design for six hydrocarbons. The three numbers represent three levels of loading rates: -1 refers to the lowest loading rate level, 0 refers to the middle loading rate level, and 1 represents the highest loading rate level. For example, the first case would have the first, second, and third hydrocarbons set at their lowest rate and the third, fifth and sixth hydrocarbons set at their middle rate. There are 54 total cases to run to obtain all the necessary information for the analysis. The design is orthogonal so no multicollinearity would exist.

Box-Behnken Design for Six Hydrocarbons

Case	Hydrocarbon					
	1	2	3	4	5	6
1	-1	-1	0	-1	0	0
2	1	-1	0	-1	0	0
3	-1	1	0	-1	0	0
4	1	1	0	-1	0	0
5	-1	-1	0	1	0	0
6	1	-1	0	1	0	0
7	-1	1	0	1	0	0
8	1	1	0	1	0	0
9	0	-1	-1	0	-1	0
10	0	1	-1	0	-1	0
11	0	-1	1	0	-1	0
12	0	1	1	0	-1	0
13	0	-1	-1	0	1	0
14	0	1	-1	0	1	0
15	0	-1	1	0	1	0
16	0	1	1	0	1	0
17	0	0	-1	-1	0	-1
18	0	0	1	-1	0	-1
19	0	0	-1	1	0	-1
20	0	0	1	1	0	-1
21	0	0	-1	-1	0	1
22	0	0	1	-1	0	1
23	0	0	-1	1	0	1
24	0	0	1	1	0	1
25	-1	0	0	-1	-1	0
26	1	0	0	-1	-1	0
27	-1	0	0	1	-1	0

Case	Hydrocarbon					
	1	2	3	4	5	6
28	1	0	0	1	-1	0
29	-1	0	0	-1	1	0
30	1	0	0	-1	1	0
31	-1	0	0	1	1	0
32	1	0	0	1	1	0
33	0	-1	0	0	-1	-1
34	0	1	0	0	-1	-1
35	0	-1	0	0	1	-1
36	0	1	0	0	1	-1
37	0	-1	0	0	-1	1
38	0	1	0	0	-1	1
39	0	-1	0	0	1	1
40	0	1	0	0	1	1
41	-1	0	-1	0	0	-1
42	1	0	-1	0	0	-1
43	-1	0	1	0	0	-1
44	1	0	1	0	0	-1
45	-1	0	-1	0	0	1
46	1	0	-1	0	0	1
47	-1	0	1	0	0	1
48	1	0	1	0	0	1
49	0	0	0	0	0	0
50	0	0	0	0	0	0
51	0	0	0	0	0	0
52	0	0	0	0	0	0
53	0	0	0	0	0	0
54	0	0	0	0	0	0

REFERENCES

Box, George E. P. and Draper, Norman R., 1987, Empirical Model-Building and Response Surfaces, John Wiley and Sons, Inc., New York

Freund, Rudolf J. and Littell, Ramon C., 1991, SAS System for Regression, 2nd edition, SAS Institute Inc. Cary, NC

Littell, Ramon C., Freund, Rudolf J., and Spector, Phililp C., 1991, SAS System for Linear Models 3rd edition, SAS Institute, Inc., Cary, NC

Myers, Raymond H., 1990, Classical and Modern Regression with Applications, Wadsworth Publishing Company, Belmont, CA

Moore, David S. and McCabe, George P., 1992, Introduction to the Practice of Statistics, 2nd edition, W.H. Freeman and Co., New York