Technical White Paper

Chemo-Mechanical Software

Fundamental Properties of Asphalts and Modified Asphalts III Product: FP 06

March 2015

Prepared for
Federal Highway Administration
Contract No. DTFH61-07-D-00005

By
Ronald R Glaser
Appy Beemer
Thomas F. Turner
Western Research Institute
3474 North 3rd Street
Laramie, WY 82072
www.westernresearch.org

WesternResearch
INSTITUTE

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

CHEMO-MECHANICAL SOFTWARE

INTRODUCTION

This technical report describes a software product designed to discover additive combinations of a wide range of independent variables that correlate with a limited set of dependent variables. A specific example would be the search for combinations of infrared spectra changes that correlate with rheological changes in an asphalt binder as it oxidizes. The original incentive for the development of this tool arose from the need to correlate a wide range of chemical measurements of asphaltic materials to the mechanical properties exhibited by those materials. The tool developed is not limited to our current application, and can be extended to many problems where a large number of independent variables are involved in a data set with limited observations of the dependent variable. Examples of possible applications would include, but not be limited to, correlation of spectral data, chromatographic data, or any data set that can be described as a list of x,y or t,y pairs against some other measured property. The crux of these problems is that many measurements are taken that are not related to the property of interest, but finding the relevant combinations is difficult. The problem we have focused on is discovering which changes in the mid-infrared spectra are most closely related to changes in an asphalt binder's rheological response as the material ages. Perhaps combinations of four or five spectral measurements are related to the property changes while the other areas of the spectra are irrelevant. This report describes the computational method and software application to discover the relevant measurements. Application of the method to asphalt problems is described in the respective technical white papers. This is not an experimental report. It is a product description that is essentially mathematical in nature.

BACKGROUND

Problem Definition: Modern analytical techniques often rapidly produce quite large data sets, the most common are those data sets usually described as "spectra". Any set of data that can be formulated as a response as a function of index can be treated as a "spectra".

Consequently, a thorough examination of the relationships between a given collection of spectra (or other arbitrary data matrix) type and an independently measured material property can generally be expressed by the following general relationship:

$$y = f\left(x_0, x_1, x_2, \ldots x_n\right) \tag{1}$$

where y is the dependent variable, e.g., Complex Modulus; and x_i is the independent variable, e.g., IR absorbance at wave number i.

If $f(x)$ is assumed to be algebraically linear, and we have only three spectra representing three asphalts (a, b, and c), along with their asphaltene contents, the equation set is:

$$y_a = k_0 x_{a0} + k_1 x_{a1} + k_2 x_{a2} + \ldots + k_n x_{an} \quad (2)$$

$$y_b = k_0 x_{b0} + k_1 x_{b1} + k_2 x_{b2} + \ldots + k_n x_{bn} \quad (3)$$

$$y_c = k_0 x_{c0} + k_1 x_{c1} + k_2 x_{c2} + \ldots + k_n x_{cn} \quad (4)$$

where *k* is the proportionality constant at each wave number 0 through *n*. Since this is a curve fitting problem, the *x* and *y* pairs are known, and we seek *k*'s that satisfy the equation set. Such a deterministic solution is impossible if *n*+1 exceeds the number of measurements. When the number of measurements is exactly equal to *n*+1, then the fit is perfect, meaning no statistical evaluation of the fit quality is possible. This is analogous to the situation in two dimensions where you are fitting a line to two data points, obtaining a correlation coefficient of 1. To obtain a statistically meaningful test of a multidimensional fit, the observations should exceed the independent variable count by some factor, the larger the better. The actual size of the multiplier depends on the desired confidence in the answer, and the precision of the measurements. Typical mid-infrared spectra will contain nearly 4,000 wave numbers, so the examination of each and every wave number for significance when combined with the others would require 28,000 observations, clearly not practical. This situation is a reoccurring problem with spectral data and other extensive *xy* data sets as well, as the inclusion of all of the data results in an equation system with excessive adjustable parameters, impossible to solve. A number of approaches exist for addressing this problem with a variety of strategies aimed at essentially reducing the number of effective *k*'s (independent variable fit parameters) to be discovered.

Available methods: A few of the methods used to address the problems outlined above are briefly described in the next three sections (Sharaf et al. 1986; Wold 1991; Wold et al. 2001; Barros and Rutledge 2004; Garson 2007; Hasegawa 2006). These methods, particularly the process of correlating spectral data to other process or property variables, have been used successfully in a wide range of applications (Basu et al. 1998; Satya 2005; Satya et al. 2007; Chalmers and Everall 1996; Hasegawa 1999; Karstang et al. 1991; Lachenmeier 2007; Mark and Workman 2007; Sastry et al. 1998; Zagonel et al. 2004), but do not produce a closed form equation in terms of measured quantities, limiting their usefulness in fundamental scientific studies.

Multiple Linear Regression

Multivariable Linear Regression (MLR) is a time-honored technique going back to Pearson (1901). Multivariable regression can establish that a set of independent variables explains a proportion of the variance in a dependent variable at a significant level (through a significance test of R^2), and can establish the relative predictive importance of the independent variables (by comparing beta weights). Variable transformations (most common is the logarithm) can be applied to independent or dependent variables to explore some curvilinear effects. Polynomials can be fit as well by expanding independent variables into a power series.

Multivariable linear regression can solve the matrix Y=MX+B, provided sufficient measurements of Y exist to obtain all of the coefficients in vector M. To be statistically

meaningful, measurements of Y in excess of measurements of X must be available, meaning that a spectra of 3,500 wave numbers would require at least 3,500 measurements of, say, complex modulus. To be statistically reliable, 35,000 would be better. It is generally impossible to apply multivariable linear regression directly to correlation studies involving data rich spectral data. However, the preconditioning of individual data points to related groups (spectral peaks, for example) is helpful to reduce the independent variable count. However, this is usually not sufficient unless a very extensive data set (many observations) is available. A variety of computation approaches have been developed in recent years that address this problem by projecting the data in one way or another into a smaller list of independent variables. These include Principle Component Analysis, Partial least Squares, and others.

Principal Component Analysis and Principal Component Regression

Principal Component Analysis (PCA) techniques are applied to the problem of too many x measurements relative to y measurements by searching for so-called latent variables. The covariance of XX' is examined and parameter space axis rotations are employed to arrive at new coordinates based on eigenvectors of the XX' matrix. In simple terms this means that independent variables that appear to change in a similar fashion are grouped together. The translated *x* variables (often called indicator variables) are projected into a smaller parameter space of latent variables. It is implicitly assumed that these fictitious latent variables somehow describe a truer "latent structure" to the system. Recall that the underlying mathematical model for the entire data set is linear, often patently untrue in chemical systems. This technique results in latent variable data sets with improved variance in the hopes of improving signal to noise ratios. Often, however, irrelevant data included in the translations pass spurious noise to the latent variables.

Principal Components Regression (PCR) is the application of ordinary linear regression methods to the latent variables developed form the principle components analysis. The difficulty with this method is that the complex axis rotations make understanding what the latent variables represent in terms of measurable quantities difficult. Interpretation of the results in terms of chemistry and physics is difficult and requires sensitivity testing by varying the input data. While useful for calibration within the testing range of the data employed, using this method for understanding the underlying science is difficult.

Partial Least Squares

PCR is based on the spectral decomposition of XX' to select latent variables for regression, while the Partial Least Squares method (PLS) is based on the singular value decomposition of X'Y, meaning that the independent variables are compared to the dependant variables. In practice, PLS usually fairs better than PCR since the reduction of parameter space dimensions is accomplished though comparison of the independent variables with the dependent variables. PCR, on the other hand, focuses mainly on what can be thought as the signal strengths of the independent variables alone for parameter space reduction, and is therefore more prone to the introduction of irrelevant signals into the regression. As with PCR, PLS suffers from the difficulty that the complex axis rotations make understanding what the latent variables represent in terms of chemistry and physics is difficult and requires sensitivity testing by varying the input

data. While useful for calibration within the testing range of the data employed, using this method for understanding the underlying science is difficult.

Other Methods

Many other algorithms have been developed in recent years, including neural networks and artificial intelligences. While these “black box methods” can work extremely well over the calibration range used, we still are faced with the difficulty of understanding how the input variables relate directly to dependent variable without sensitivity testing. Because of the difficulty of latent variable methods to demonstrate the correlation in terms of directly measured variables, we developed our own methods to address the issues associated with impossible and/or unfavorable parameter to observation ratios. In the simplest of terms, two strategies can be employed to make the problem tractable; reduce the independent variable count, or increase the number of observations. Once a statistically meaningful correlation can be computed, a method for selecting the most important independent variables must be applied to remove irrelevant signals and find those responses that significantly affect the quality of the fit. When applying this technique to infrared to rheology correlations, the independent variables are spectral wave numbers and represent vibrational modes of functional groups. Consequently, important clues about how chemical changes cause rheological changes can be obtained.

SPECTRELATE SOFTWARE DEVELOPMENT

After reviewing the available methods for multivariate studies of spectral data, we found that understanding the important wave numbers was very difficult with methods utilizing the latent variable concepts. The time honored linear multivariable method does produce closed form correlations, but is only useful if intelligent selection of possible significant wave numbers is known a priori. The number of independent variables to use in the correlations is also limited based upon the number of actual observations used. Monte Carlo approaches would likely work, but the adaptation of the ideas present in PCR to reduce variable counts in a logical manner, along with a defendable method for expanding the observation matrix, is the strategy we chose to employ. The core strategy we employ involves reducing the number of independent variables into groups and increasing observations artificially if needed. In short, we use multivariable linear regression to examine the entire spectrum at one time. This is possible by preconditioning the data set by combining independent variable single measurements (that is, a single wave-number) that represent the same information to reduce the independent variable list and by increasing the observation matrix size by generating synthetic replicates from knowledge of measurement precision. The method precision provides an envelope for the application of random variations in the existing measurements to generate a probable collection of additional synthetic observations.

Consider a set of asphalt spectra aged at different conditions. We also have rheological measurements of these materials corresponding to the each spectrum measurement. For the time being we are not interested in the dependent variable (the rheology) but only the spectra. There will be variations in the spectra depending on the aging conditions. If we take a data set that contains infrared spectra of an asphalt that has been oxidized over a range of times, all other

factors held constant (pressure, temperature, concentrations) and then plot a single wave number against all the others, we find that for some wave numbers the plot is nearly a perfect line, and for others, the plot is quite scattered. Let's illustrate this idea with some real data from WRI's ambient atmosphere thin film aging study.

If we plot the 1035 wave number against 3200, for AAB-1, we get the plot shown in figure 1.

Each point in the plot represents those absorbances for that amount of oxidation. So, an increase in the 3200 absorbance is always accompanied by an increase in the 1035 absorbance by a factor of roughly five. For this binder, it appears that the 1035 response and the 3200 response are very closely related (correlation coefficient of 0.99) and describe either the same functional group in different vibrational modes, or possibly describe reaction products produced by the same mechanism. Other wave numbers have no correlation at all (correlation coefficient of 0.10), as this comparison of 1035 with 3600 demonstrates in figure 2.

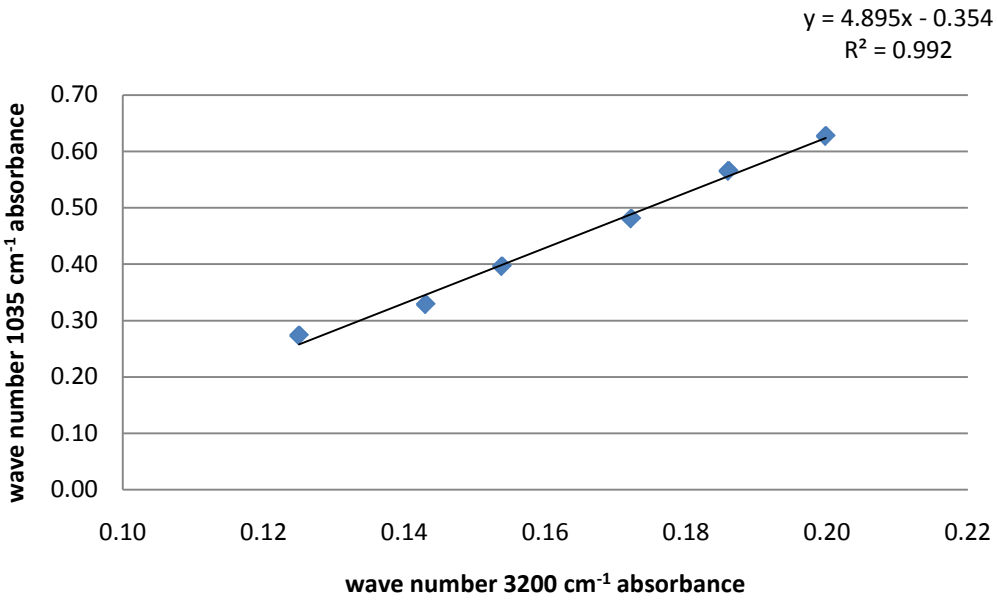

Figure 1. Graph. The correlation between wave numbers 3200 and 1035 for oxidized AAD-1 asphalt binder.

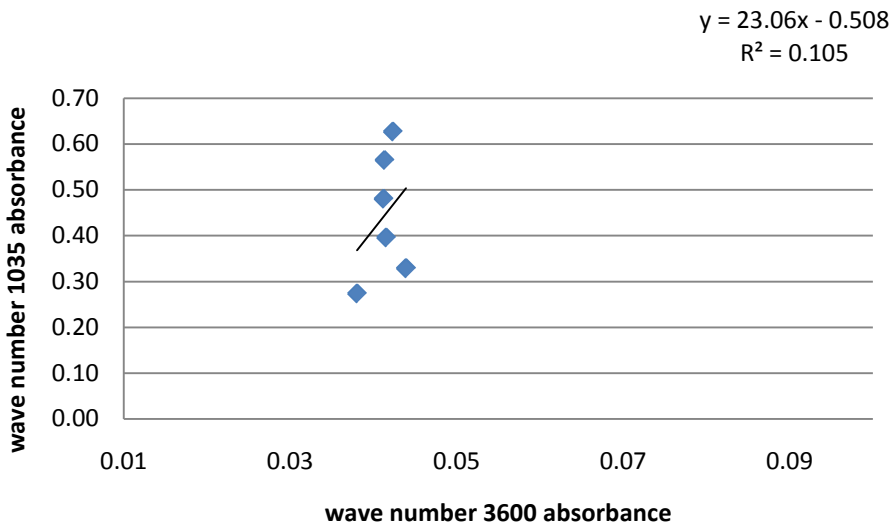

Figure 2. Graph. The correlation between wave numbers 3600 and 1035 for oxidized AAD-1 asphalt binder.

If we perform this exercise for each and every wave number, and record the correlation coefficient, we can construct “regression spectra” to investigate which wave numbers seem to contain the same information as the asphalt oxidized (these independent variable change in proportion to each other, and are not co-linear, but co-related). A co-linear relationship is on the same line. The slopes and intercepts are identical. For the binder AAD-1, we get the plot shown in figure 3 when the regression spectra are plotted.

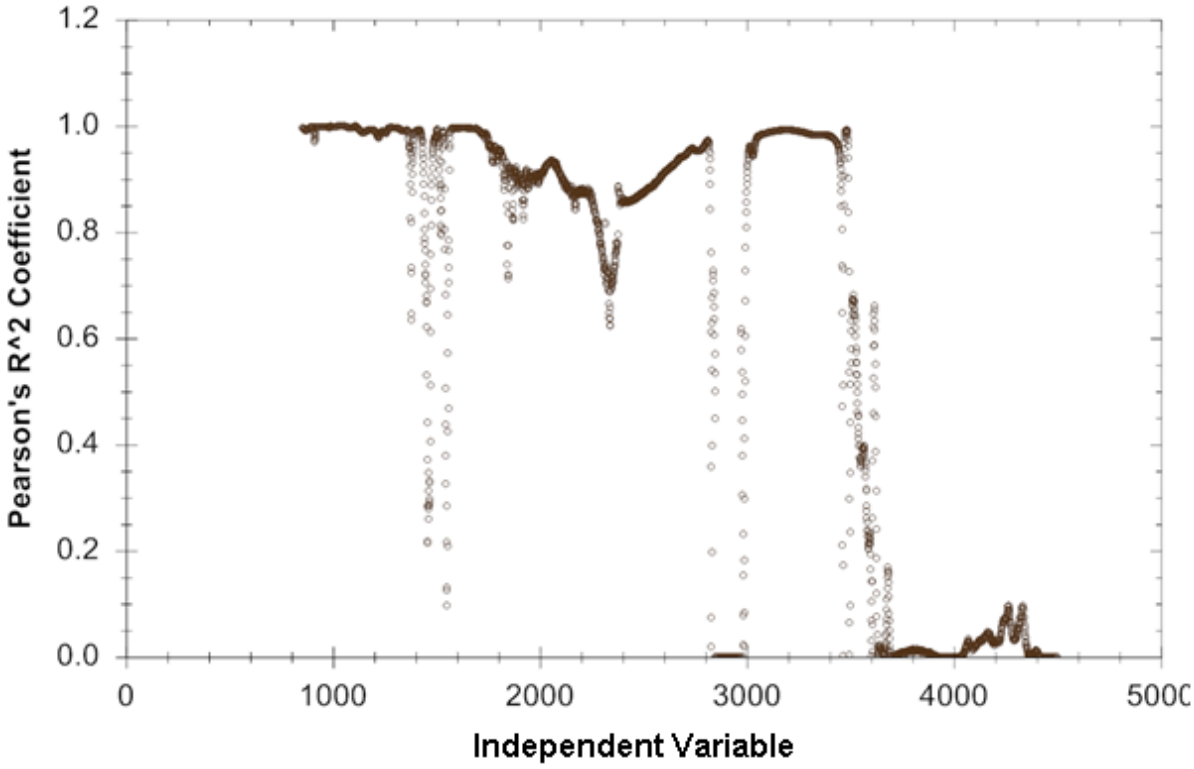

Figure 3. Screen capture. The regression spectrum for wave number 1035 for oxidized AAD-1 asphalt binder.

For this particular binder, if we were to choose a cross correlation threshold of 0.95, then most of the spectrum would fall into one group, greatly simplifying any regression efforts. However, when studying a collection of spectra for a variety of binders, we find that groupings vary from binder to binder. Compare a regression spectrum of AAB-1 (figure 4) with AAD-1 above (figure 3).

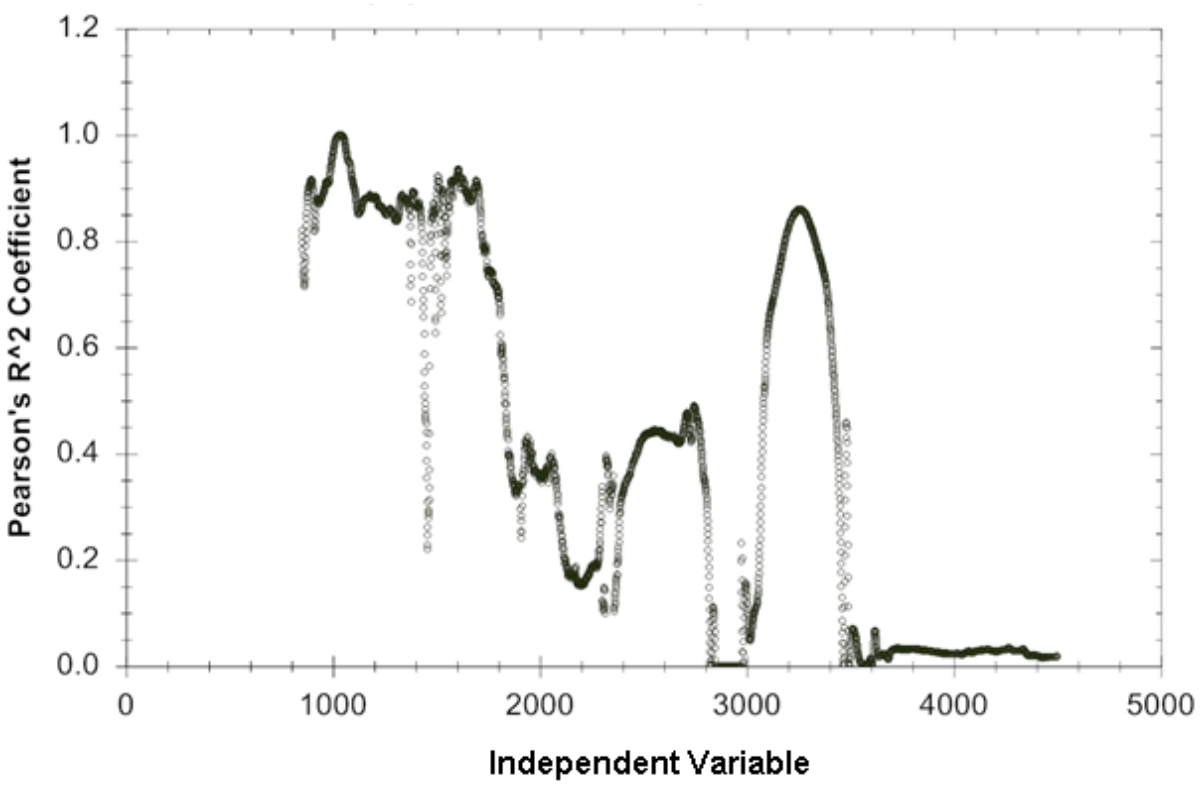

Figure 4. Screen capture. The regression spectrum for wave number 1035 for oxidized AAB-1 asphalt binder.

If we further extend the exercise and produce an “overall” grouping schema for many binders by creating a regression for this wave number (1035) we get the regression spectrum shown in figure 5.

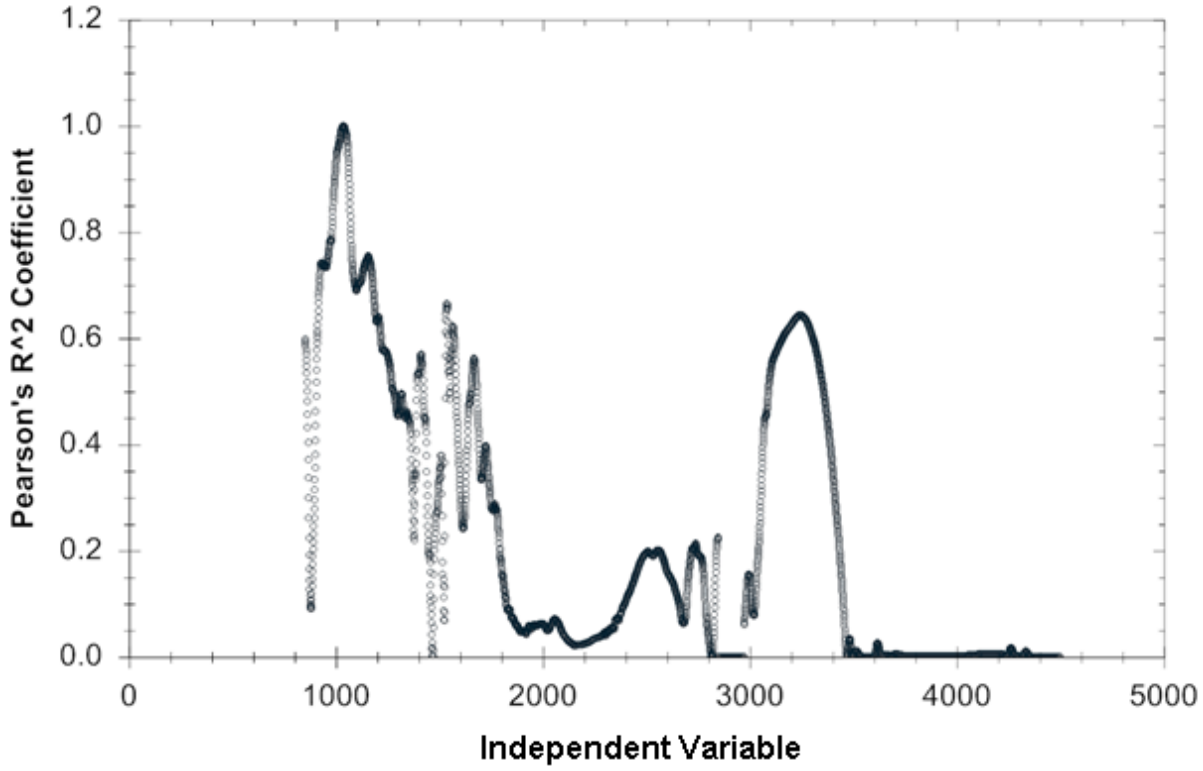

Figure 5. Screen capture. The regression spectrum for wave number 1035 for oxidized asphalt binder data set.

This plot suggests that nothing really cross correlates at any wave number with 1035 except adjacent wave numbers making up the primary peak. However, if we take our pre conditioning a step farther, and consider only the changes in the spectra by subtracting the RTFO spectra (time 0) from the remaining spectra aged at times up to 12 weeks, a very different pattern emerges. We are now only considering the changes that occur with oxidation.

The raw spectral data is shown in figure 6 with nonlinear and solvent affected areas set to zero:

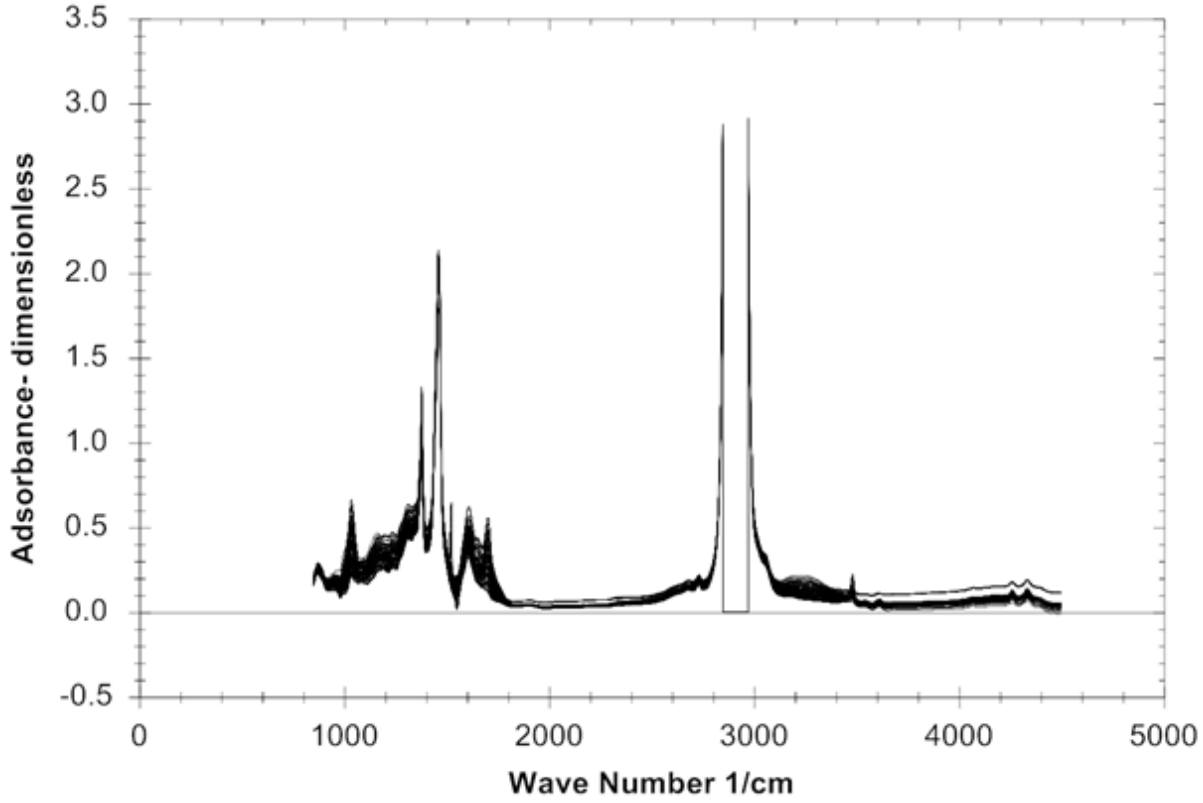

Figure 6. Screen capture. Mid infrared spectra with non-linear response areas removed.

The RTFO (zero oxidation time) subtracted spectra is shown in figure 7 (effectively eliminating the y axis intercept in the regressions and focusing on the changes).

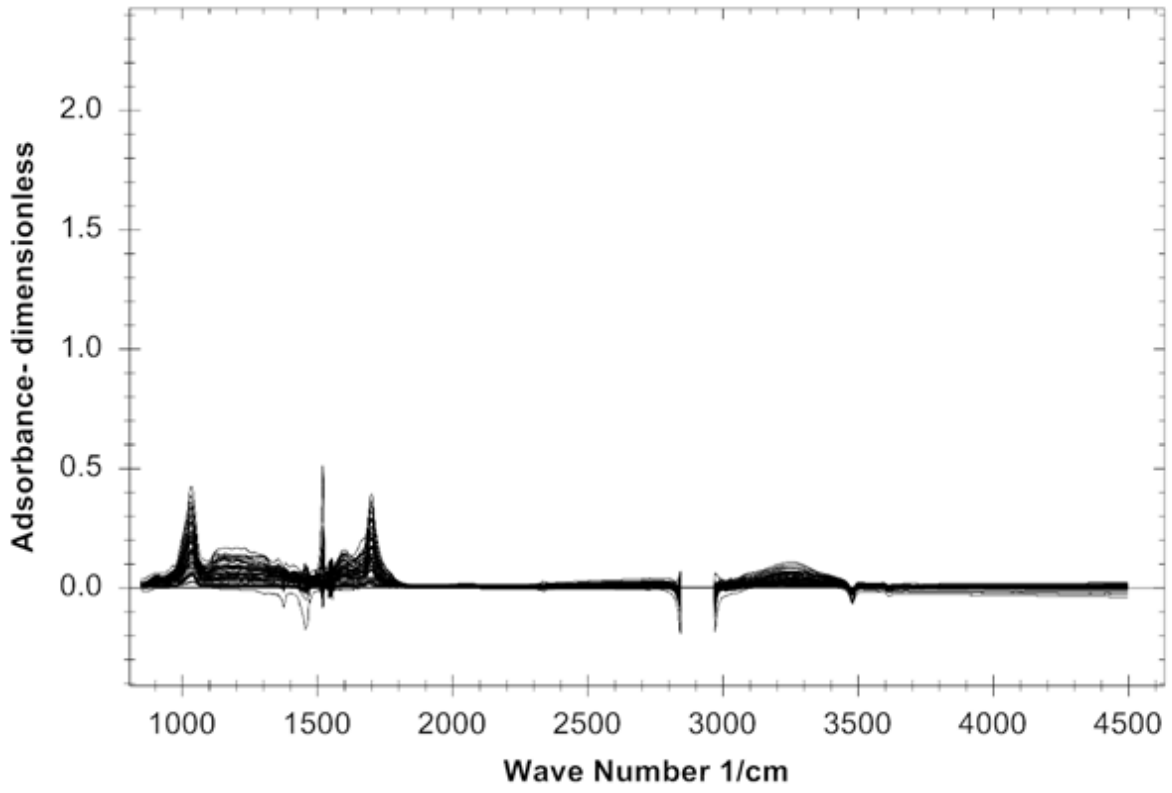

Figure 7. Screen capture. Mid infrared change spectra with non-linear response areas removed.

When 12 different binders are correlated, we get a the regression spectrum shown in figure 8.

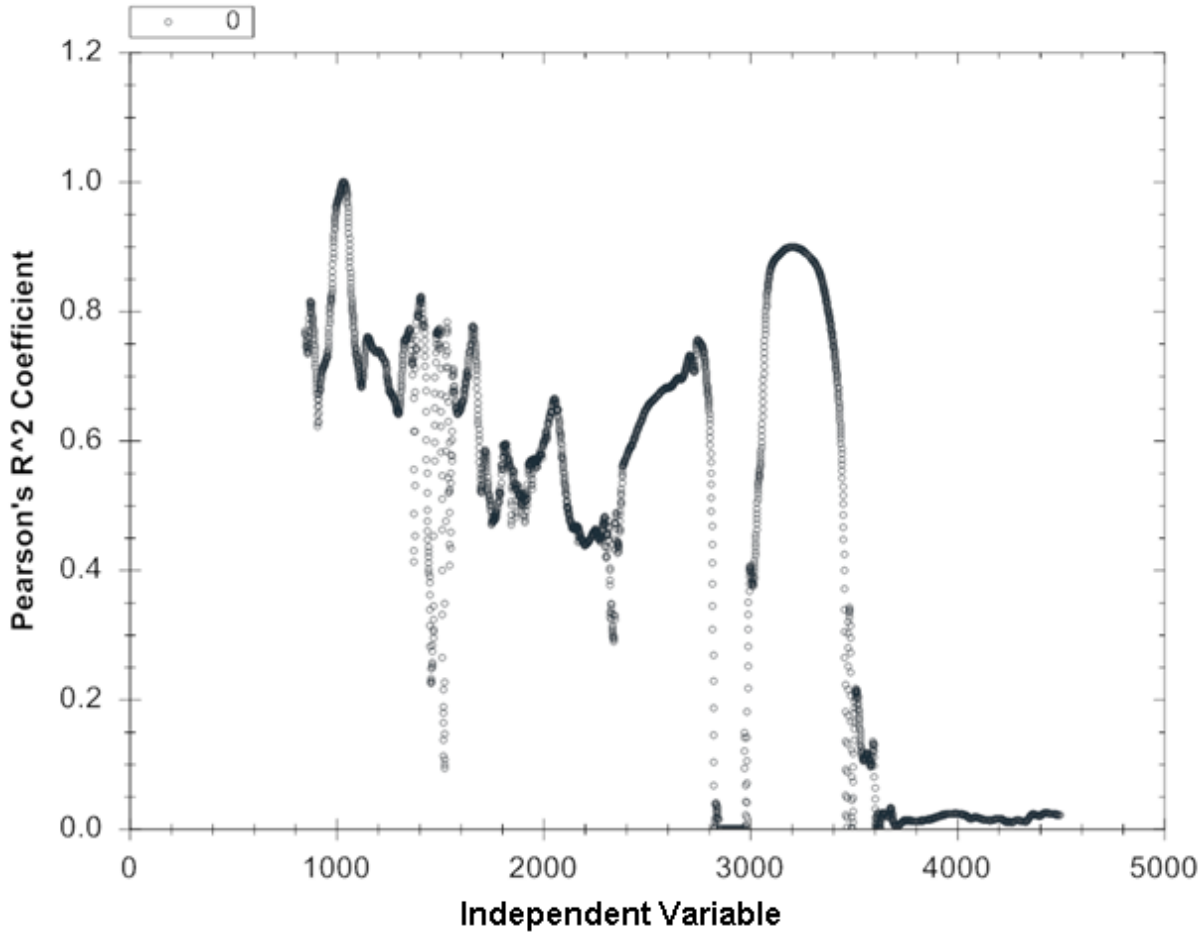

Figure 8. Screen capture. The regression spectrum for wave number 1035 for oxidized asphalt binder data set change spectra.

In figure 8 above, as compared to figure 5, the increase in sulfoxide (wave number 1035) and the increase in hydrogen bonding (wave number 3200) appears to be related. They both grow proportionally as oxidation progresses, providing clues to the underlying chemical mechanism.

When the program groups the related wave numbers, the user must specify a correlation threshold as criteria for group selection. If a regression threshold of r^2=0.8 is selected, then this group would contain all the wave numbers inside the grayed out area (figure 9), reducing the number of independent variables significantly. A higher threshold produces more groups, with fewer members in each group. A threshold above r^2=0.95 or so would not include any of the 3200 peak in the group. For example, if try to correlate each wave number in the collection of spectra with every other wave number with an excessively strict regression threshold of 1, we would have 3200 groups(spectral range of 800 to 4000) with a single member in each group. As we reduce the criteria for assuming the wave number represents the same information as others in the spectra, then those wave numbers near the wave number of interest begin to be included, describing the region of the spectral band that “moves” with the wave number of interest in the data set. If the threshold for acceptance into the group becomes low enough, we begin to see

other regions in the spectrum some distance from the immediate peak or band that either describes other vibrational modes of the same functional group, or in the case of a reaction series like oxidation, these distance related responses might include other functional groups participating in similar reaction paths. The decision on how many groups is reasonable is essentially a matter of judgment, perhaps best tempered with experimenting with different threshold values in the regression analyses.

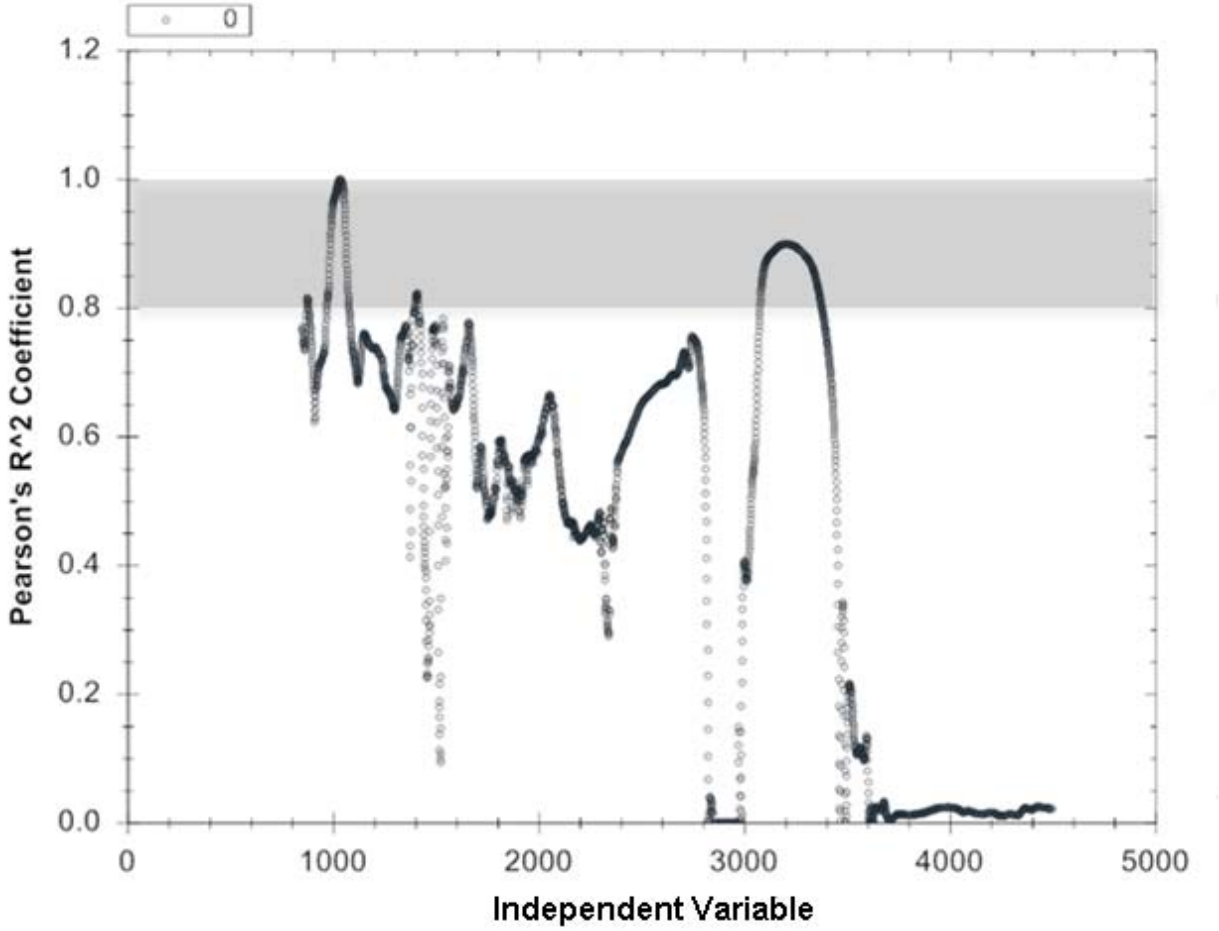

Figure 9. Screen capture. The regression spectrum for wave number 1035 for oxidized asphalt binder data set with a grouping range shown.

Using this same data set, which is 70 oxidation spectra obtained from 12 binders, the number of groups identified with this technique is shown in table 1.

Table 1. Independent variable counts at different grouping threshold values.

Net Change Spectra (12 Binders-70 Spectra)	
R^2 Threshold	Number of Groups
0.99	284
0.95	101
0.9	51
0.85	35
0.8	28
0.75	22
0.7	20
0.65	16
0.6	12
0.55	11
0.5	9
0.45	8
0.4	7
0.35	6
0.3	5

The question then arises as to how groups are selected to properly characterize the data without including differentiation caused by method noise. A defendable theory is not currently known to the authors, but a semi-log plot of the table above does provide some guidance with regard to a point of diminishing return appearing to exist near the threshold of r^2=0.9 (figure 10).

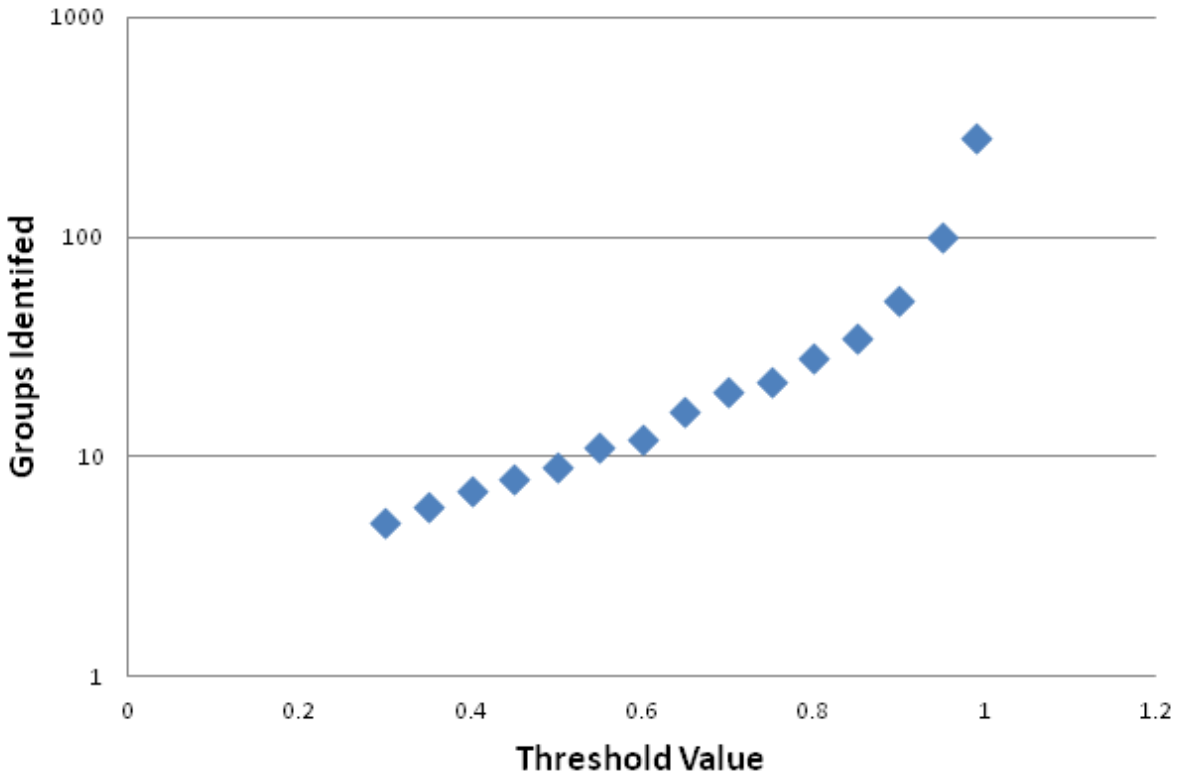

Figure 10. Graph. Groups produced at different threshold values.

Providing the method above produces a reasonable guide to identify meaningful signals from the data, we have 51 independent variables to examine in combination for significance with only 70 observations available when using a grouping threshold of 0.9. The observation ratio is only slightly greater than 1, so any fitting procedure would a 51 dimensional analogue to drawing a line through two points in 2 dimensions. If we had 357 observations, (O/P, observations/parameters = 7) then we could begin to believe the statistical validity of our correlation attempts. We accomplish this by generating synthetic independent data and dependent data (in this example, spectra and rheological measurements) by randomly varying the measurements within the method precision envelope. The idea here is that if it were practical to create such a huge data base, the data would look very much like the synthetically generated set. Once the regression is performed using a commercially available multivariable linear regression package, the significance of the individual variables can be ranked based upon the F test for that variable (calculated by the statistics package). The F test is a numerical representation of the quality of a fit with and without the independent variable of interest. Those variables that test with a higher value are more significant than those that test with a lower value. The formula use to calculate F test is

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1} \right)}{\left(\frac{RSS_2}{n - p_2} \right)} \quad (5)$$

RSS_1 and RSS_2 are residual sum of squares of model 1 and model 2 (residual sum of squares is the squared difference between measured and model values, and in our case the two models are those with and without the parameter in question). The numbers of parameters used for each model are p_1 and p_2, and n is the number of observations. So, in simple terms this is a ratio of goodness of fit with and without the parameter in question. The F-test value provides a means of ranking the significance of the parameters in the models proposed, which can be used as a rejection criteria.

Our software performs the correlations by first grouping the independent variables, then creating, if needed, sufficient observations by synthetic replication to make the matrix not only tractable, but statistically significant. The idea of using synthetic replication is rather simple, but difficult to describe concisely in multidimensional problems. Fortunately, the principles at work can be shown in simple two dimensional linear fitting problems with only two adjustable parameters, the line slope and the intercept. Clearly, at least two points are required to obtain some idea of the relationship exiting between the measurements, if any. Figure 11 is a plot of two points generated from the relationship y=0.25x+1.

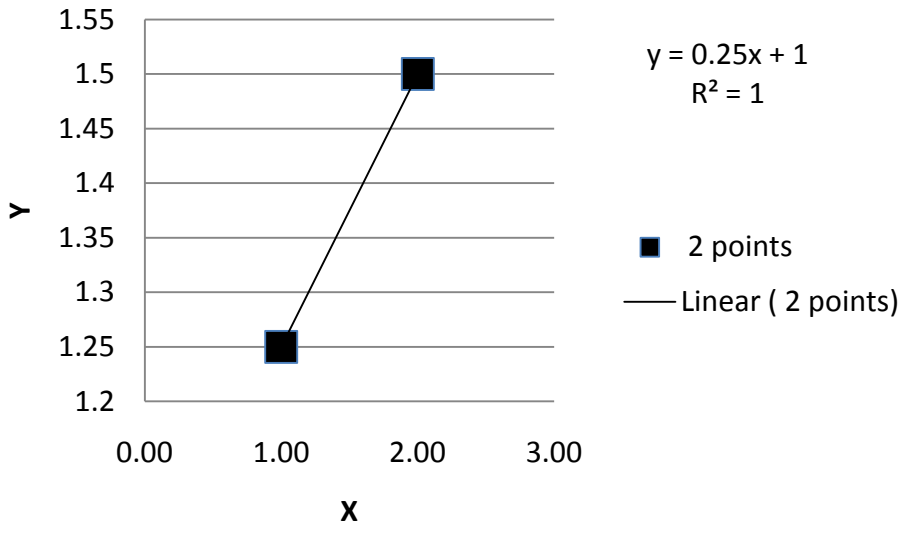

Figure 11. Graph. Simple linear plot with insufficient measures to test confidence.

If the measurements are perfect (as they are in figure 11) the relationship between the variables is clear. Notice that Pearson's regression coefficient, r^2, is perfect in this case and that this metric gives us no indication of confidence in the data. Suppose we have a very low signal to ratio for these measurements of only 0.5, which is a proportional precision of +/- 200%. If we repeat our measurements 7 times for each point, we get the following plot:

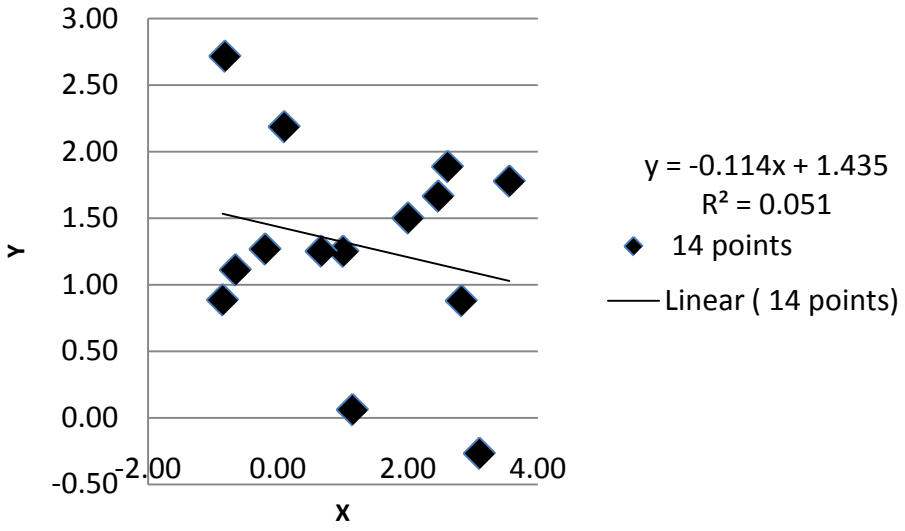

Figure 12. Graph. Simple linear plot with 7 replicates per measurement, signal to noise ratio 0.5.

In this case, the correlation coefficient is quite small, indicating very little confidence in the data. If this data were part of a multidimensional fit, the F test, which is also based upon the sum of squared residuals, would low. The important point to realize is even though there is a correlation in this set, we cannot get an accurate fit without running many, many replicates. The other point to keep in mind, is that synthetic replicates assumes the actual measured value is near the mean, but we have no way of knowing without real replicates. The actual fit is poor in this case as well. This rarely matters in our software implementation, since we are checking several thousand dimensions (independent variables) in the spectra to see if they are statistically believable. If not, they are thrown out. The situation improves significantly with better precision and/or areas of the spectra with a strong response. Keep in mind that the proportional precision in spectra varies with measurement strength. Low absorbance signals have huge error bounds, while strong signals have small proportional error bounds. The absolute error is approximately constant and the software provides that option for specifying precision. Consider a signal noise ratio of 5 (20% precision) as shown in figure 13, which would apply to a stronger absorbance measurement than seen in figure 12:

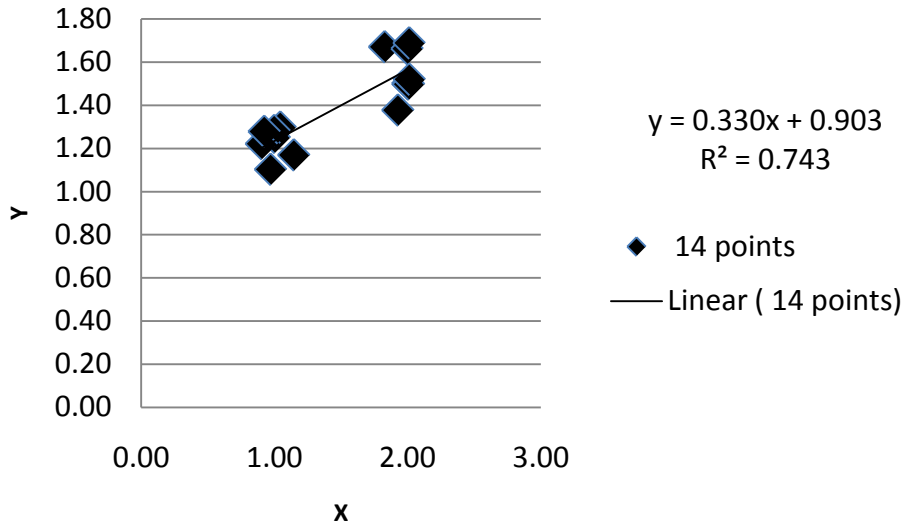

Figure 13. Graph. Simple linear plot with 7 replicates per measurement, signal to noise ratio 5.

Further improvement in the signal to noise ratio to the value of 20 (5% precision) leads to a high confidence correlation:

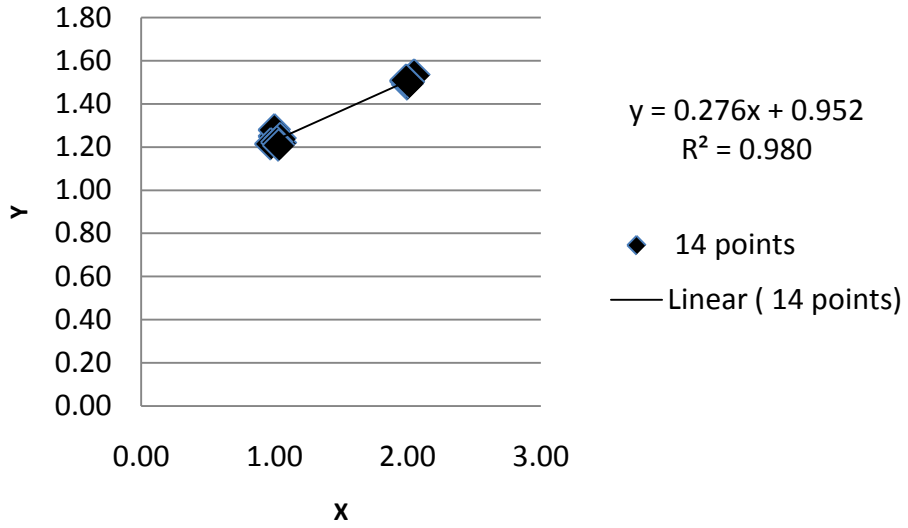

Figure 14. Graph. Simple linear plot with 7 replicates per measurement, signal to noise ratio 5.

For the hypothetical, simple data set presented here, we have plotted in figure 15 the relationship between the number of replicates, and regression coefficient for measurement precisions ranging from 5% to 200%. We typically begin to have a reasonable estimate of the correlation coefficient at when the number of replicates exceeds 7. This behavior carries over to multivariable correlations as well.

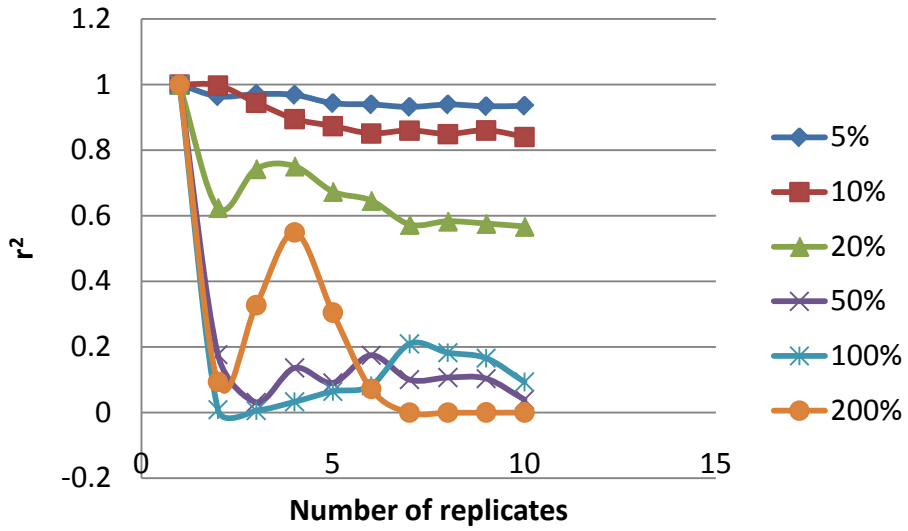

Figure 15. Graph. The regression coefficient dependence on the number of replications and the measurement precision.

To recap, the Spectrelate program reduces the number of independent variables by grouping measurements that correlate with each other into a single value. In the event the ratio of observations to parameters (or observations to independent variables +1) is unfavorable for matrix solution or more commonly meaningful regression coefficient or F test calculations, synthetic replication can be used to increase the O/P ratio in order to judge which variables have insufficient precision and signal strength to justify their use. These are eliminated from consideration. The initial multivariable correlation is performed assuming all of the independent variables are needed in the model. The F-test ranking is then used to discard the variables one by one, performing repeated correlations until there is only one independent variable remaining. The Pearson's correlation coefficient is plotted against the number of model parameters, and a judgment made, guided by the shape of this curve, of how many independent variables are needed to explain the variance within the precision of the data. If properly used, the resulting selected independent variables will produce the proper correlation without grouping or synthetic replication provided the O/P ratio for the final model is large enough.

PROGRAM IMPLEMENTATION

Data Input

As with any computational software, experimentally obtained measurements must be entered to perform the calculations. The data is entered in the form of two files, one containing the independent variable measurements, and a second file containing the dependent file measurements.

In the example files provided, the independent variables are the measured absorbances at each frequency, expressed as wave number. The data is read as a comma delimited file, which most

commercially available spectrometers can produce. Any data that can be placed into a Microsoft Excel spreadsheet can be saved as a comma delimited file. These files can also be opened in Excel for editing or perform alternative calculations such as variable transformations. The first row is arbitrary and can contain any information of the user desires. The second row serves as a label for the data in each column, and for the example datasets this will be the name of the infrared file. This text entry will appear in the observation list in the software. The first column is an exception as it contains the wave number list. Each spectrum must be the same length and have the same spacing between wave number readings.

The dependent variable file is of similar form to the independent variable file. The first row can have arbitrary entries (not used by the software). The second row identifies the samples from which the measurements (or measurements) were taken that are found in the column below. Any number of measurements can exist in the column below, but there must be at least one. The second rows should indicate the same sample was used for both the infrared measurements in the columns below it and also the same physical sample was used to obtain the dependent variable measurement (in our case, rheological measurements). The first column is the labels of each dependent variable measurement. If user wishes to investigate variable transformations, such as the logarithm of a measurement, it is easy to do here by entering the measured values for each sample in row 1, and the transformed value in row 2. The label for these entries (each row) will be displayed in the dependent variable selection list. This selection must be made before any kind of regression can be attempted. The structure of the independent and dependent variable files are easily understood by examination. They can be opened in excel, or a text editor.

The process of reading the data and placing it into numerical arrays in the program is initiated by first selecting the “Load Independent Data” button (figure 16).

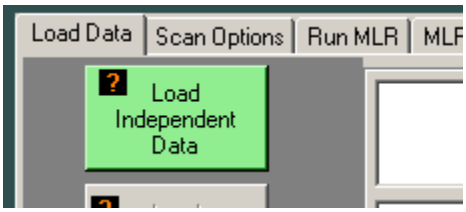

Figure 16. Screen capture. Load data button.

An open file dialog box appears (figure 17), and the user can navigate the computer file structure to find the desired file.

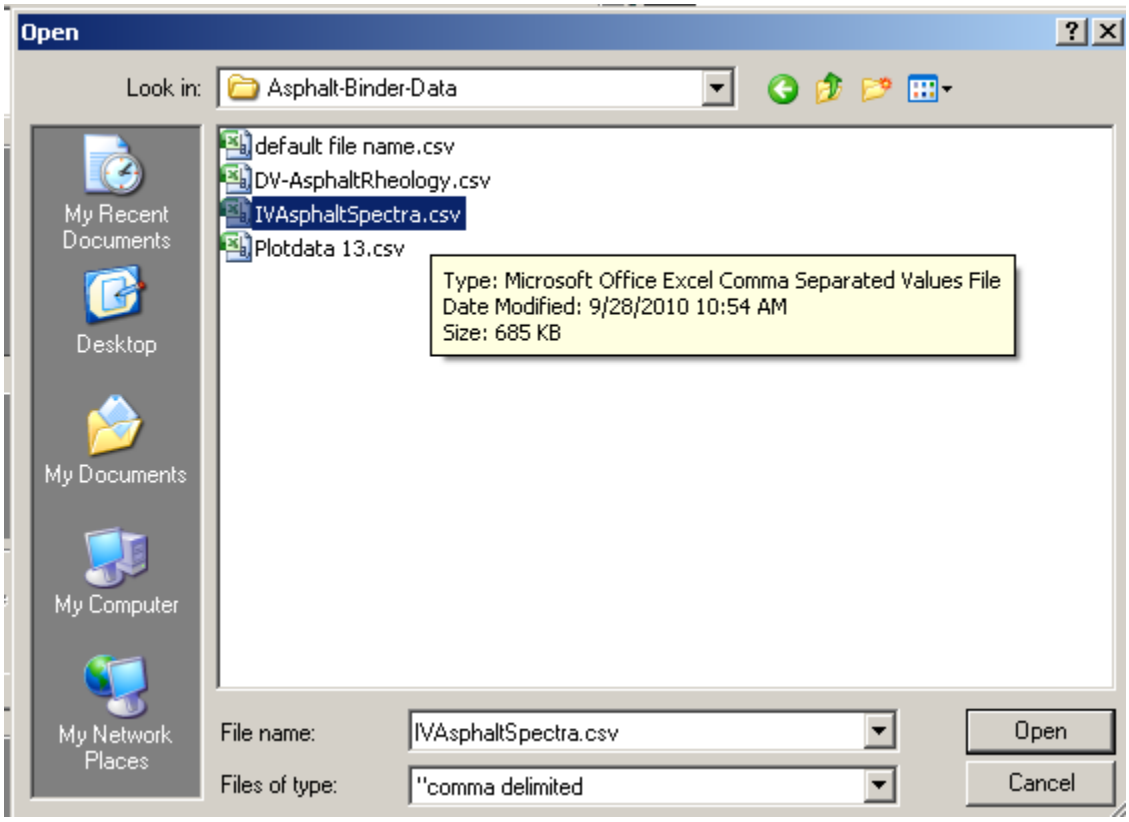

Figure 17. Screen capture. Data file selection.

Once the data has been successfully read, the independent variable sets will be plotted in two of the plots to the right (raw data and reduced data, see figure 18). At this point, these are identical, but will change if cross correlation grouping is used.

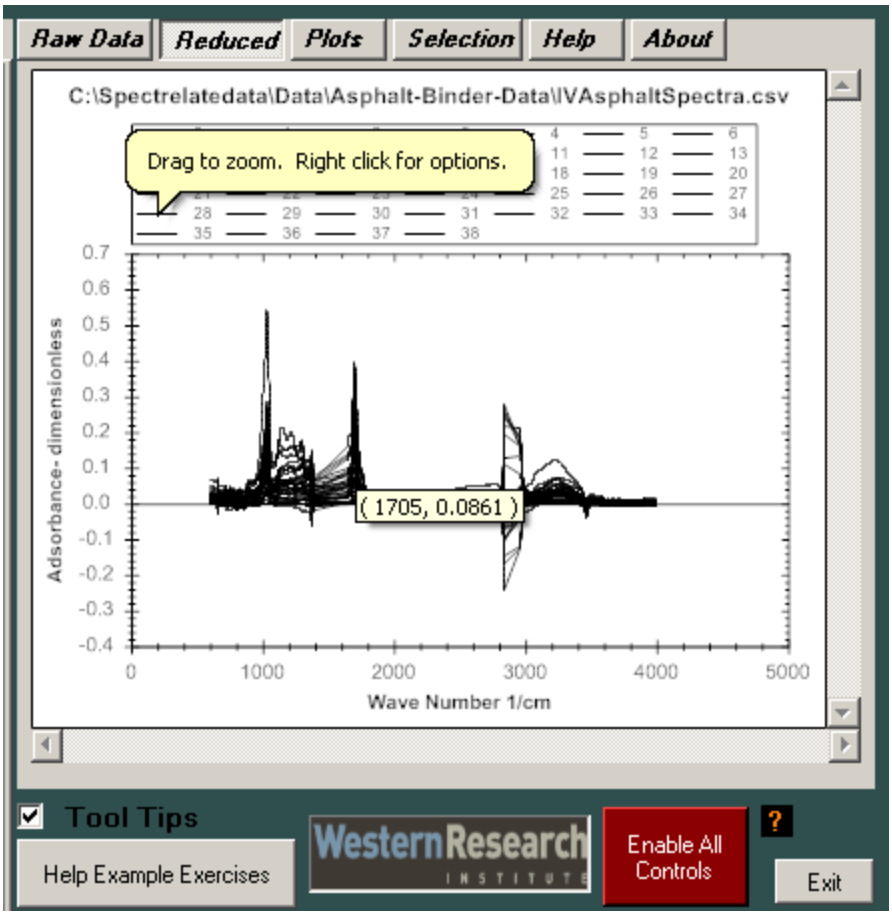

Figure 18. Screen capture. Data display.

A similar procedure for reading the dependent data file is employed. It is started with a click on the “Load Dependent File” button (figure 19).

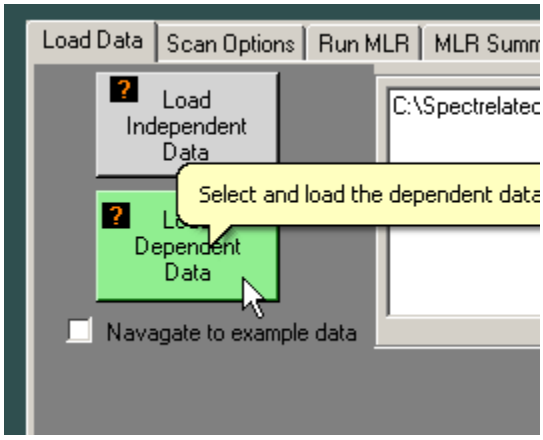

Figure 19. Screen capture. Dependent data button.

The dependent data are not plotted. However, the dependent variable selection list box will be enabled after the program verifies that sample counts match for both independent and dependent files. If you have several candidate dependent variables in your file, then there will be a list here. At least one item must be selected to proceed with any regression attempts (figure 20).

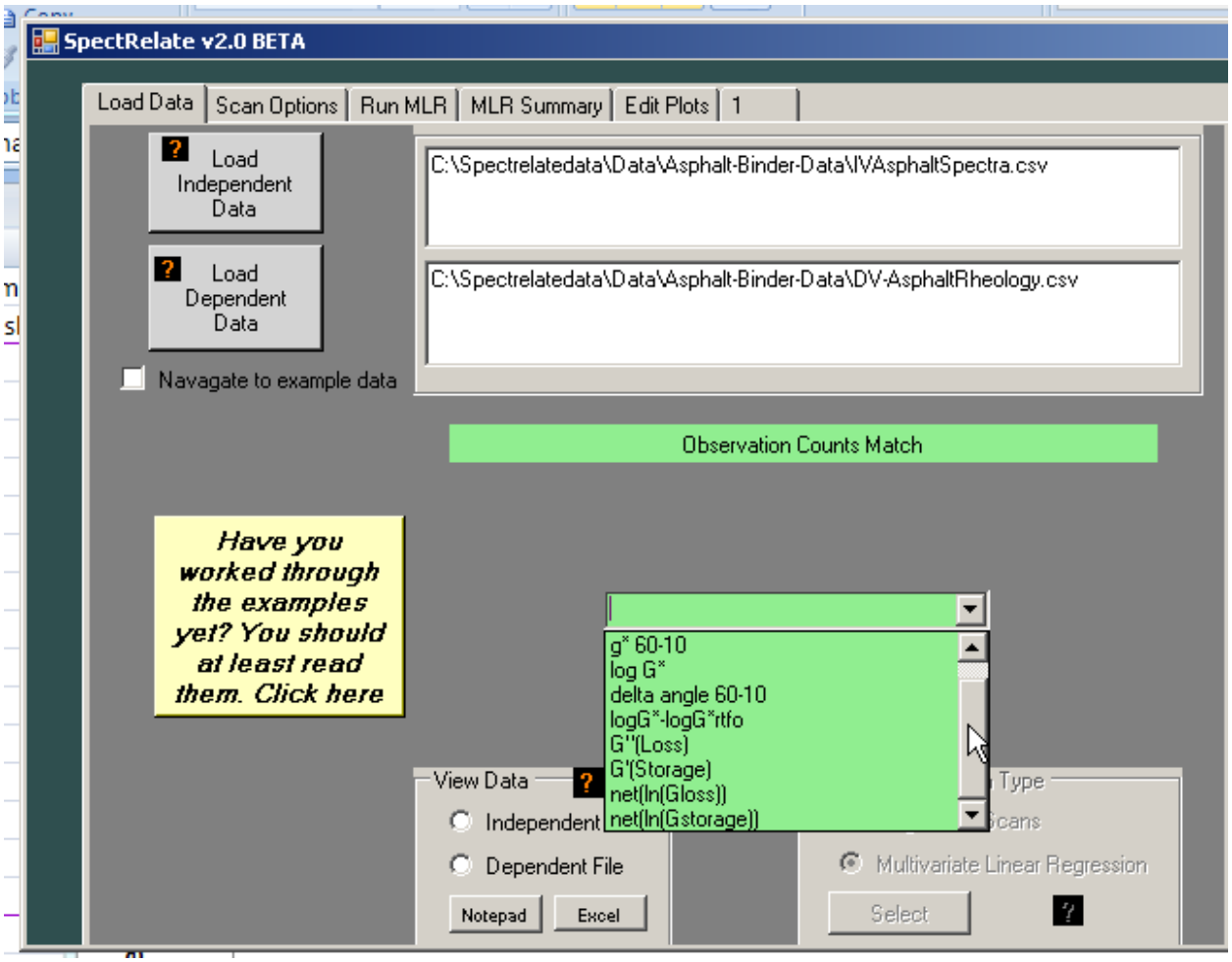

Figure 20. Screen capture. Dependent variable selection.

Dependent Variable Selection

A dependent variable must be selected to enable additional controls. Here we select the change in the logarithm of the complex modulus.

Regression Type

Two options exist regarding regression type. The regression scan goes through the entire list of independent variables and does a single variable regression on each one of them (figure 21).

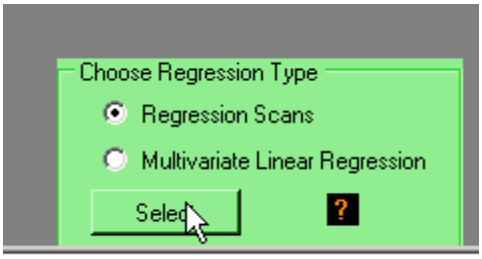

Figure 21. Screen capture. Regression type selection.

Upon selection, the tab control shifts to the regression scan page, and a scan is performed on the current data set using one variable at a time. The regression coefficient, the intercept, or the slope of the line for each one of these linear correlations is plotted on the y axis under the plots tab. The example below in figure 22 shows the regression coefficient.

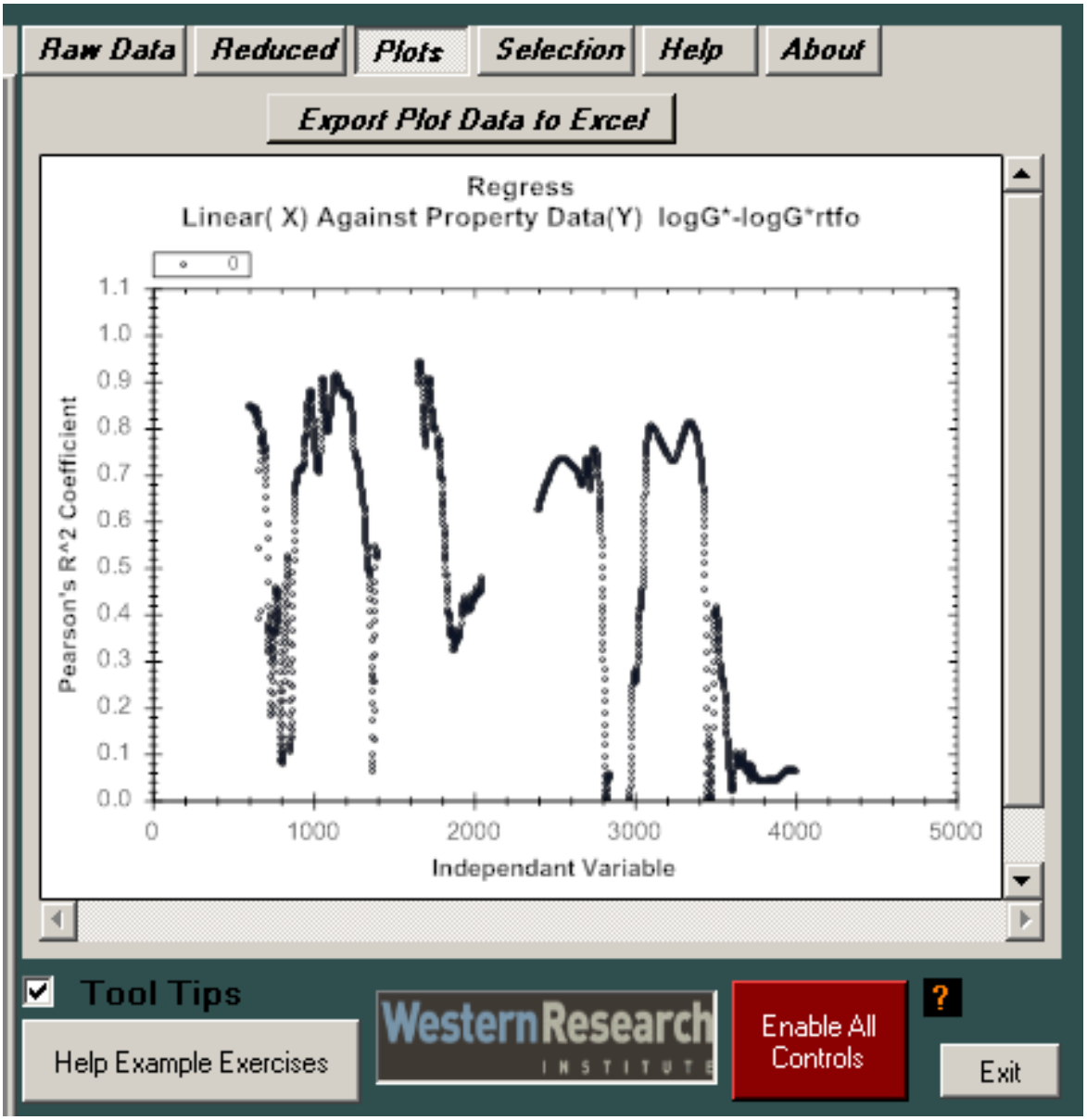

Figure 22. Screen capture. Regression spectra example.

The regression coefficients for individual independent variable can be queried from the plot with a mouse hover over the point. This is quite helpful if you want to examine the actual regression using the individual plots feature (figure 23).

Figure 23. Screen capture. Regression spectra example (magnified).

In this example, we select the wave number 1660 and its value and the associated regression correlation coefficient is shown. To obtain the actual regression plot, the independent variable is selected from the drop down list for individual plots (figure 24).

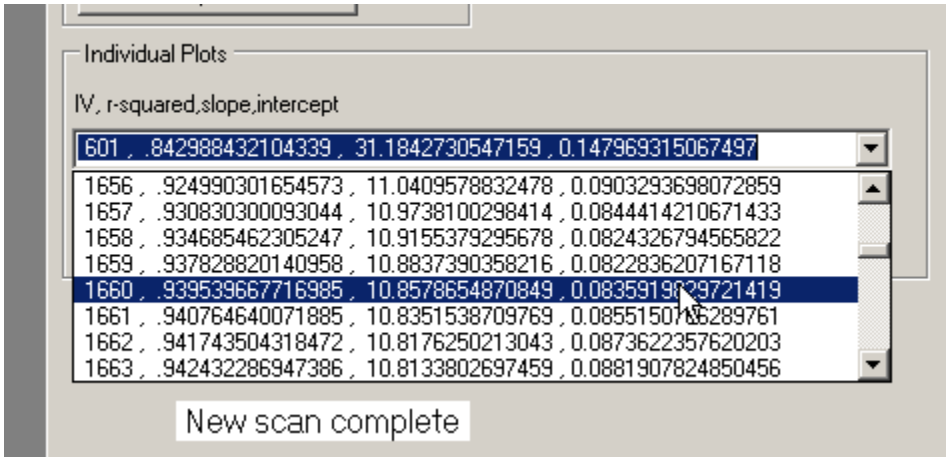

Figure 24. Screen capture. Regression plot selection.

A click of the single iv regression plot button produces the plot for examination (figure 25). The legend lists the measured data points and the regression line.

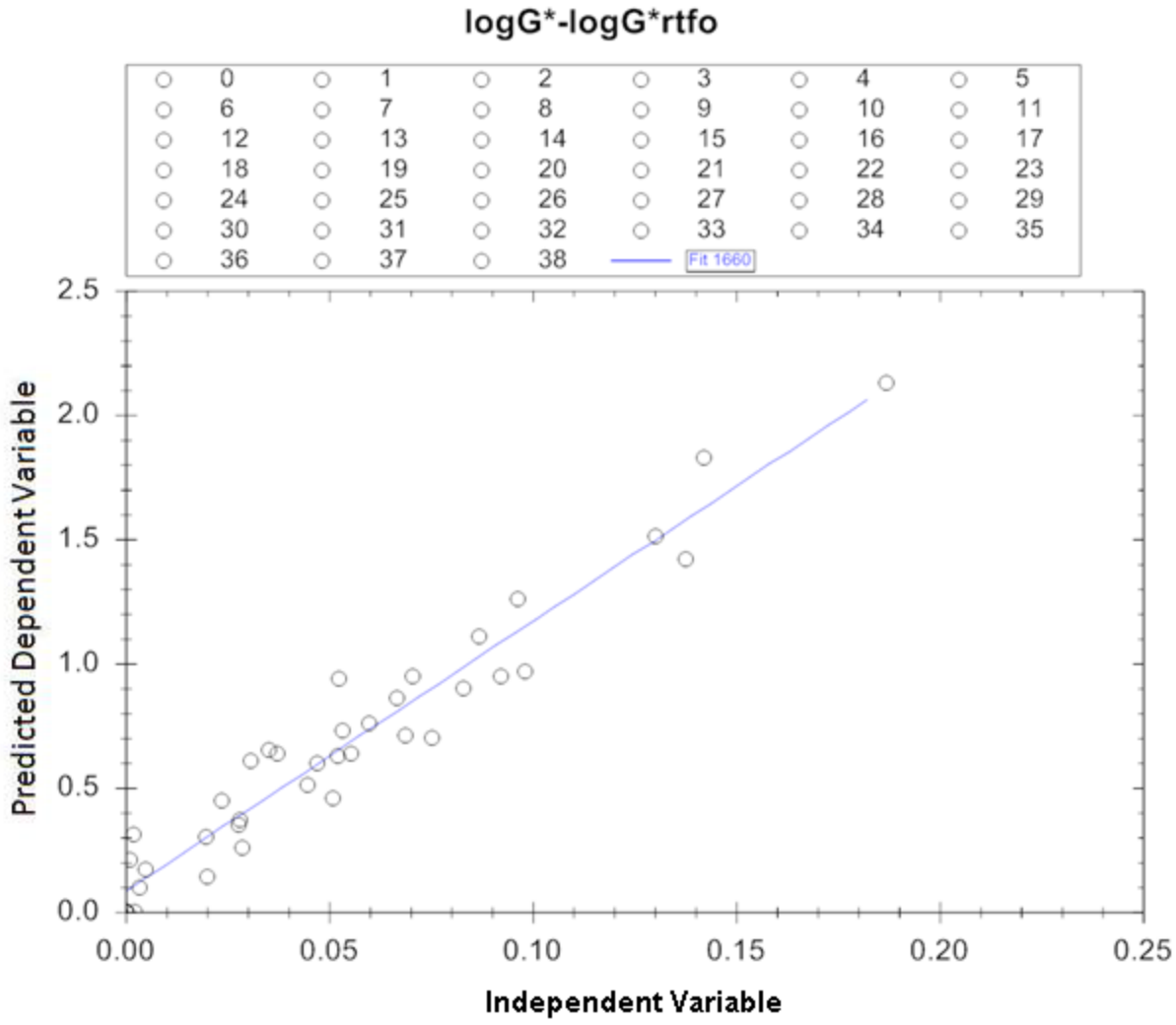

Figure 25. Screen capture. Regression plot example.

The poorly correlating wave number 1878 produces the following plot (figure 26). The examination of the individual plots is helpful to investigate the possibility of a relationship that is not linear and therefore appears to correlate poorly when a variable transformation to linearize the relationship may provide a better indication of the relationship. Obviously, such a situation does not exist in figure 26.

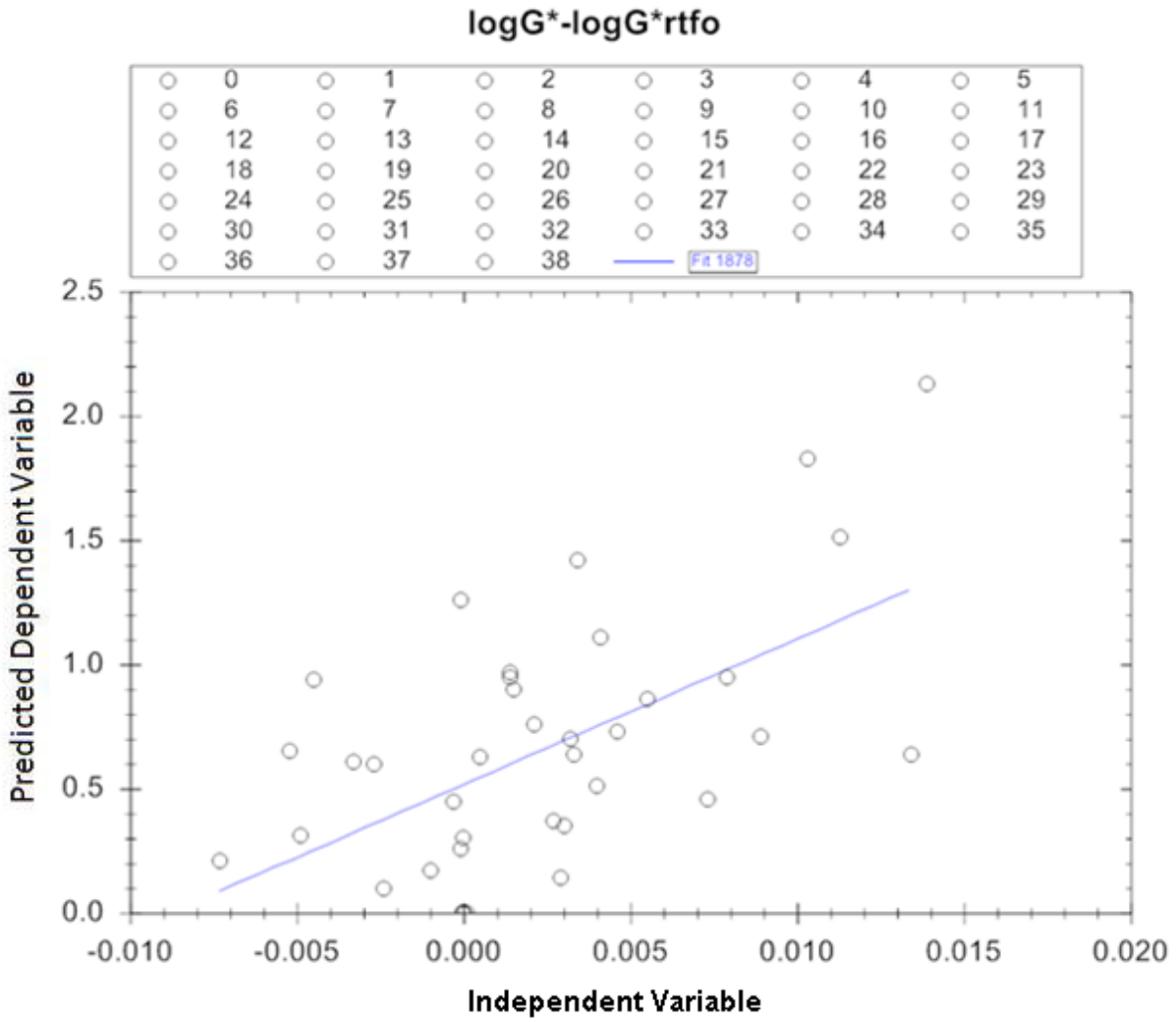

Figure 26. Screen capture. Regression plot example 2.

The discussion to this point described the capabilities of using the regression scan option to investigate single independent variable correlations of the form for each and every independent variable in the set:

$$y = ax + b \tag{6}$$

The other option, found on the Load Data tab, is multivariate linear regression, which is used to investigate additive correlations of more than one independent variable.

$$y = b + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n \tag{7}$$

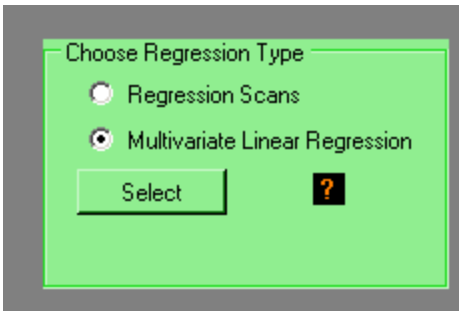

Figure 27. Screen capture. Multivariate regression selection.

When this option is selected (figure 27), the next step is to reduce the number of independent variables through some kind of selection process. The automated grouping process has been discussed in some detail earlier, but in some cases, other criteria may be more useful in choosing the list of independent variables likely to explain the system variance. A conceptual model may be available or highly correlating wave numbers from the scan study may produces better correlations when combined. In the example we have been showing, we notice several wave numbers correlate fairly well, and it might worth trying some of those. So, in the iv reduction selection list, we would select manual (figure 28).

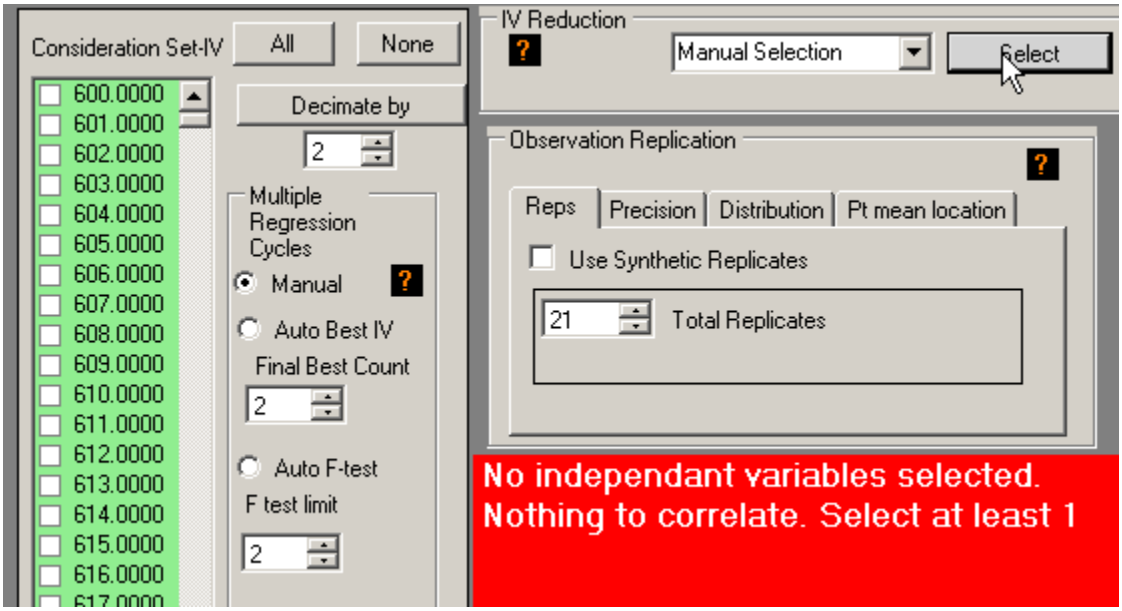

Figure 28. Screen capture. Independent variable reduction method selection.

The consideration set on the left shows every independent variable in the data set, in this case IR wave numbers spaced apart by 1 unit. Notice that the Observation replication function is disabled, and the status window informs that nothing is selected to correlate. If we select all of the wave numbers, then we discover the correlation cannot possibly be done (figure 29).

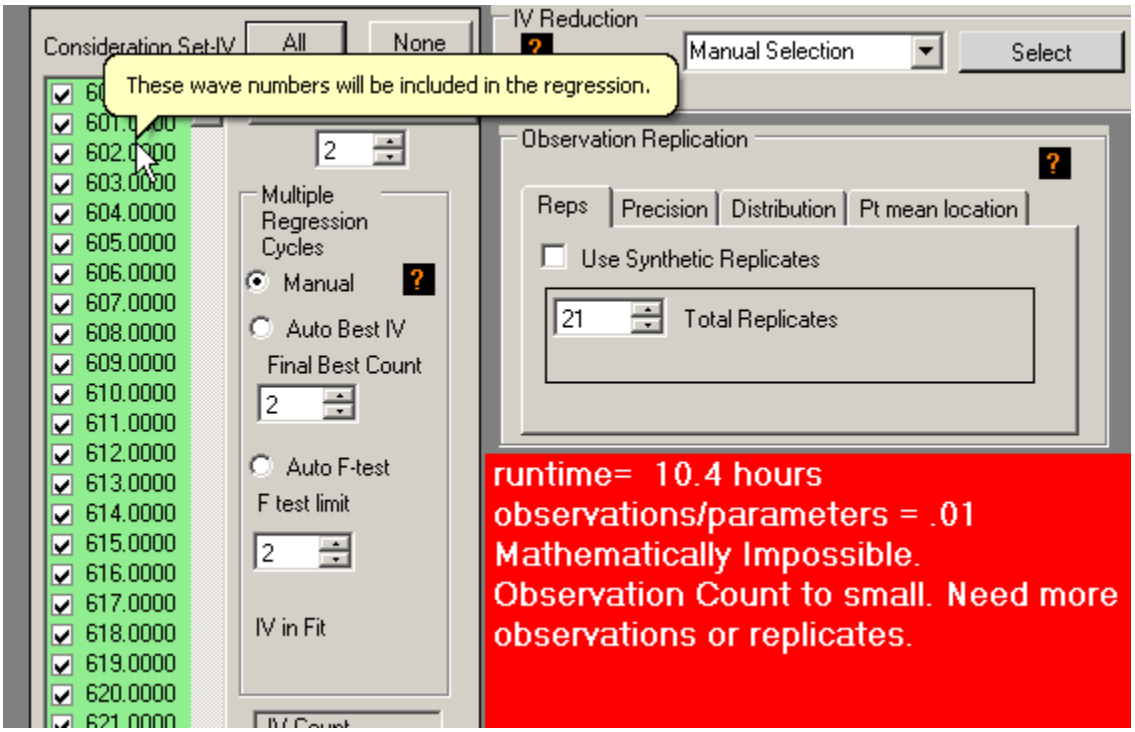

Figure 29. Screen capture. Independent variable selection example 1.

If we deselect all of the wave numbers, and then select just 1663, we get this screen (figure 30).

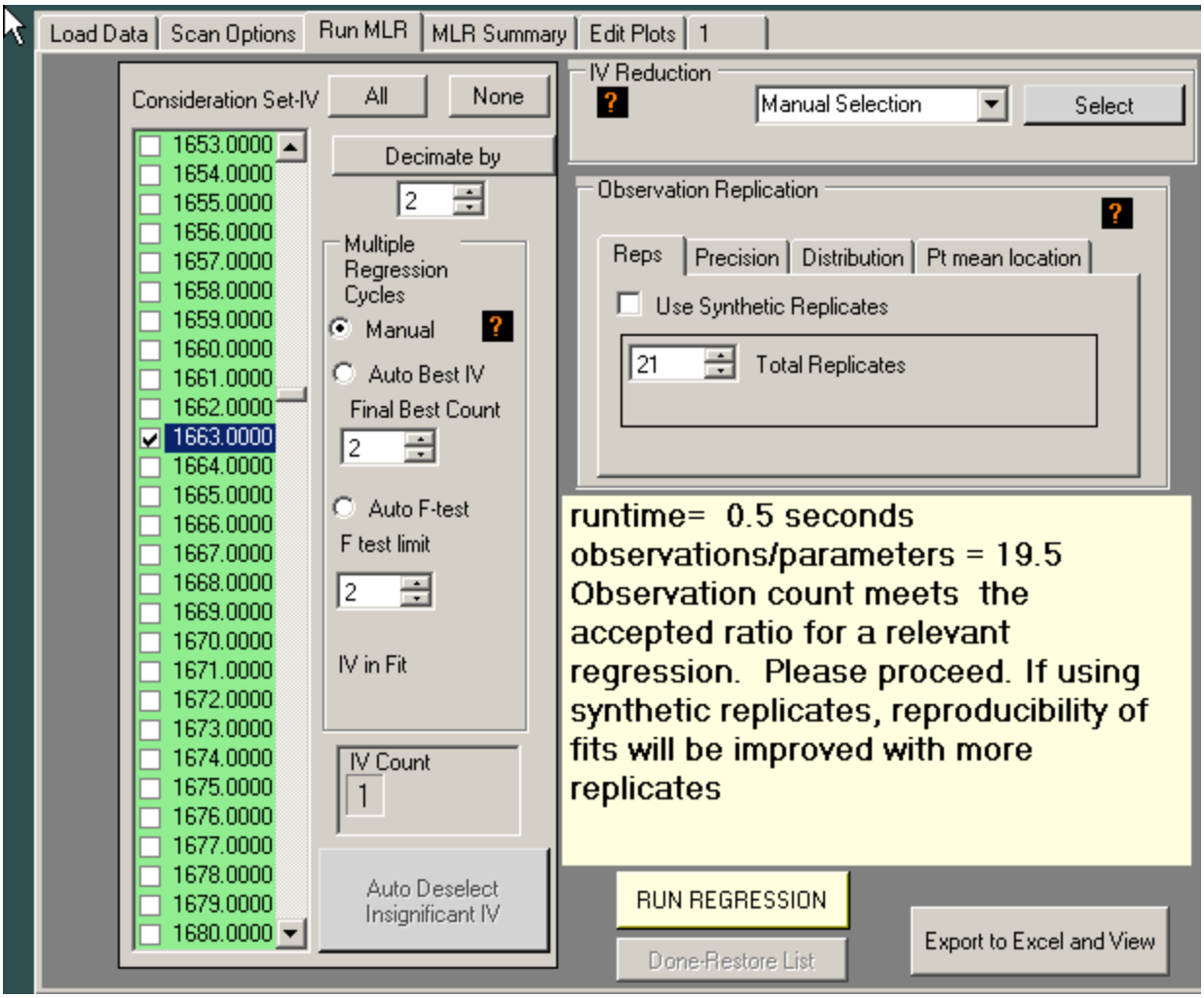

Figure 30. Screen capture. Independent variable selection example 2.

It is now possible to calculate a regression, the result being a simple linear one variable regression shown in figure 31.

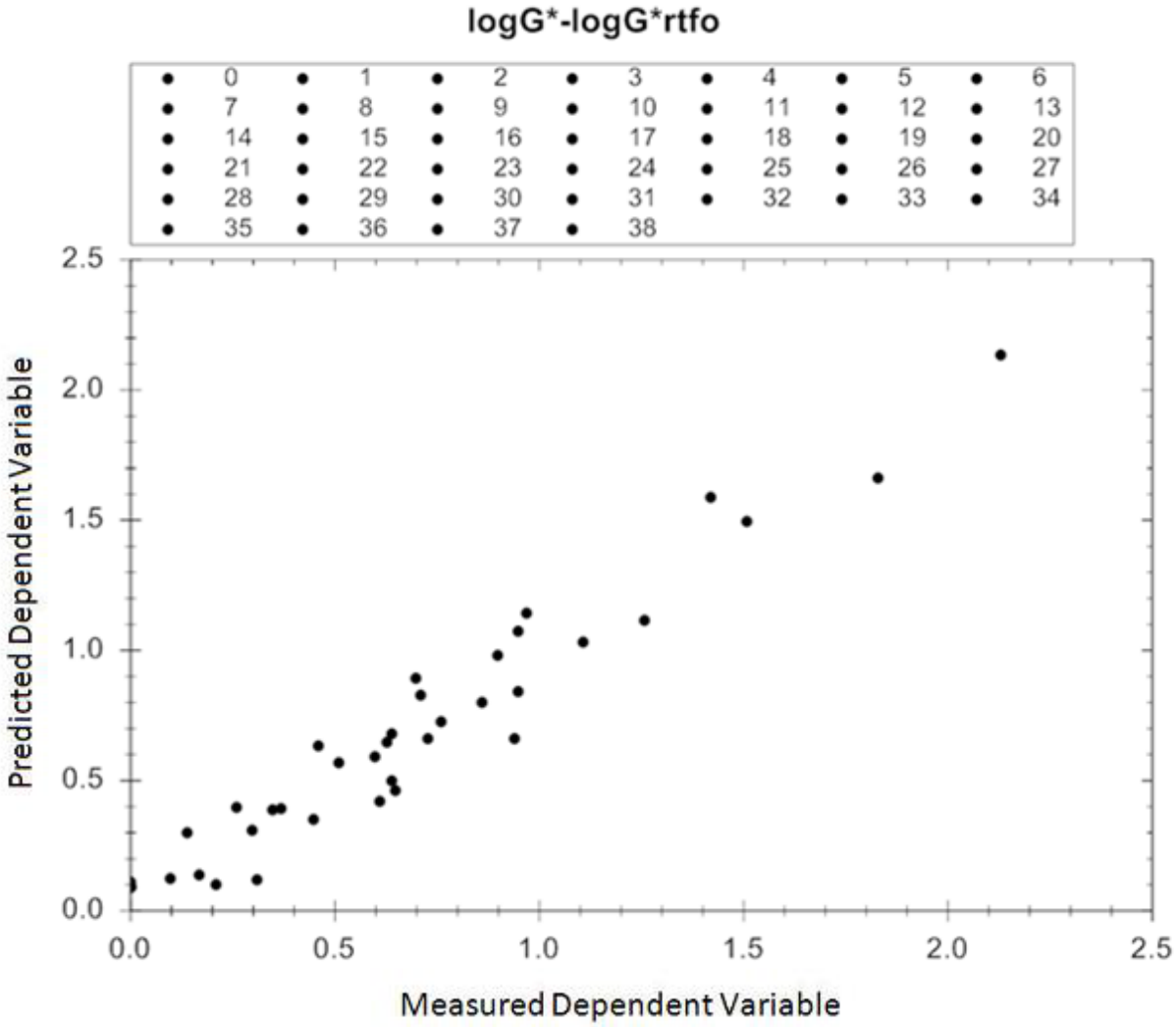

Figure 31. Screen capture. Regression results plot of predicted and measured dependent variable values.

If we select 6 independent variables, the O/P ratio now falls to 5.57 (above 7 would statistically be better), and the regression can proceed (figure 32).

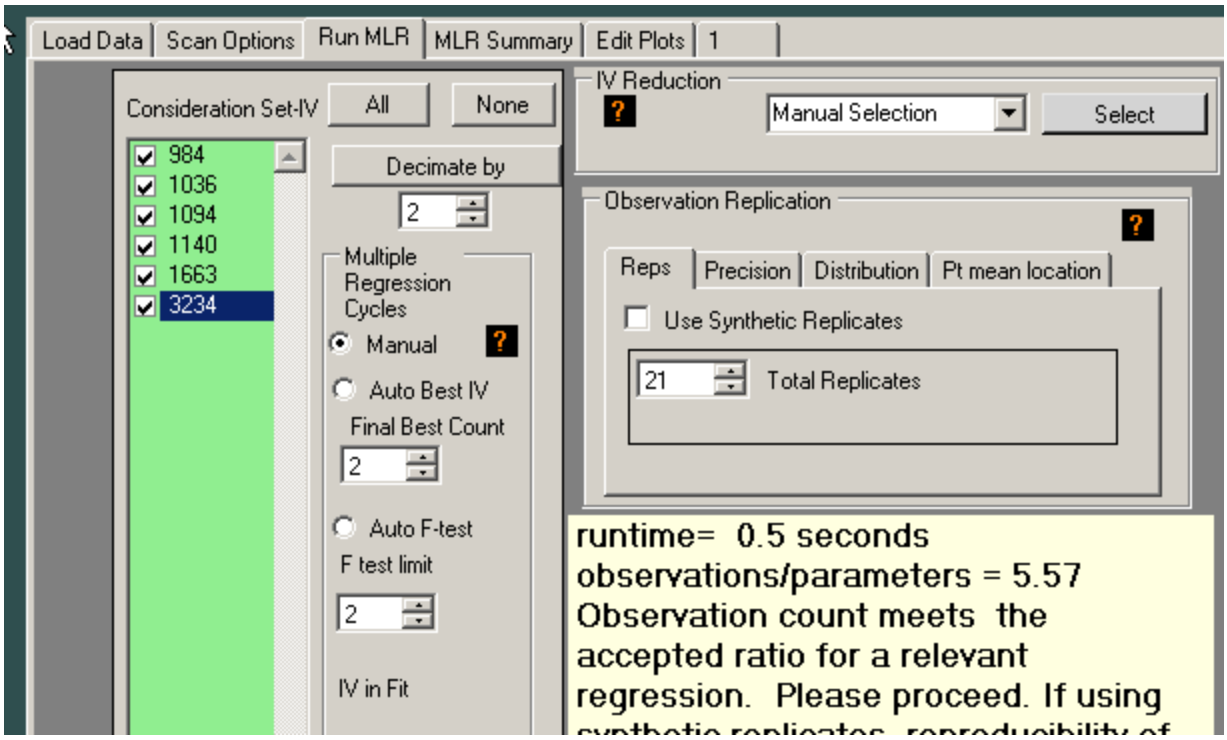

Figure 32. Screen capture. Software notification of favorable observation to parameter ratio.

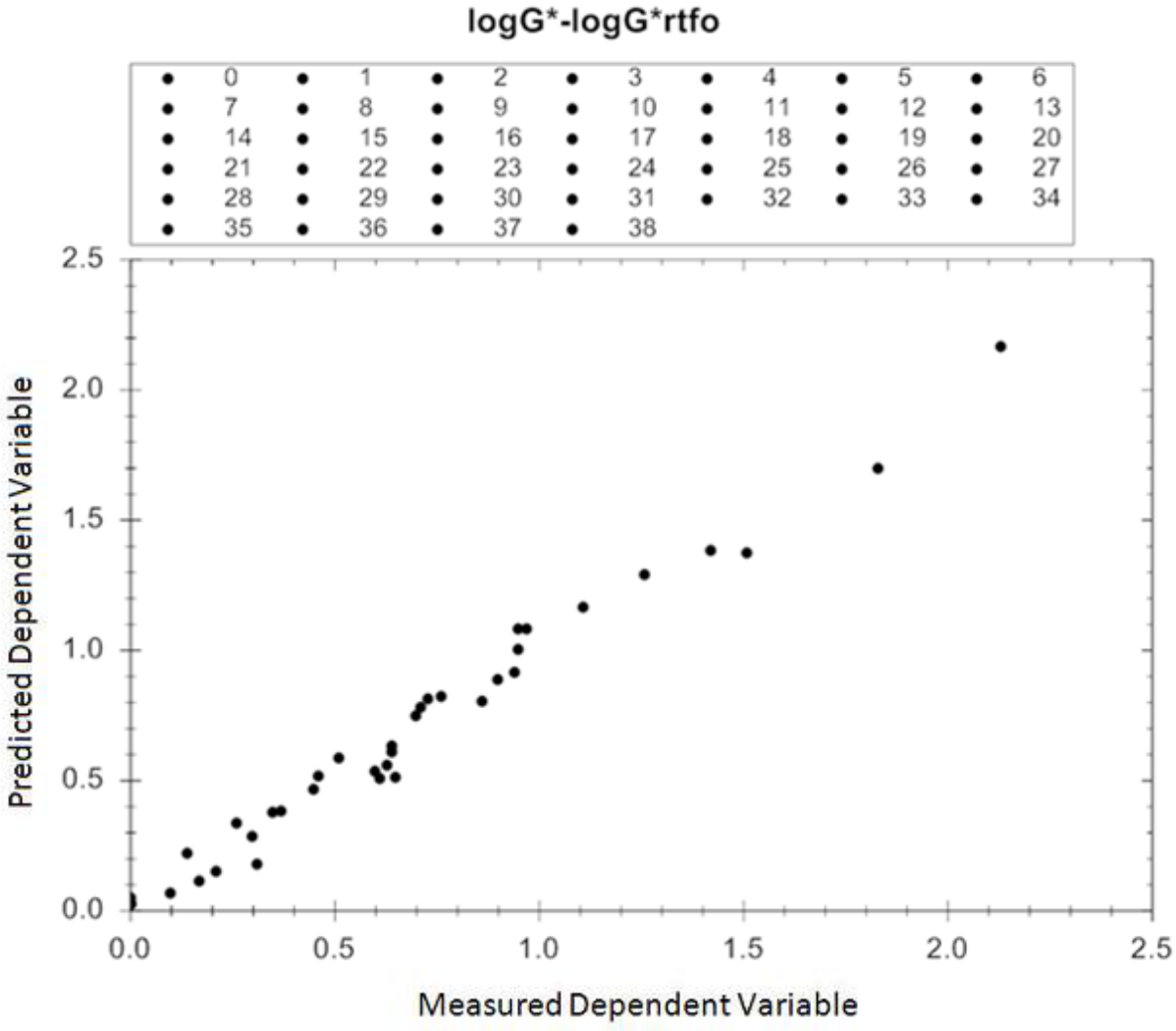

Figure 33. Screen capture. Regression results plot, example 2.

That worked quite well (figure 33), so perhaps more wave numbers would be better. When the count exceeds 38, the screen turns red, indicating the computation is impossible (figure 34).

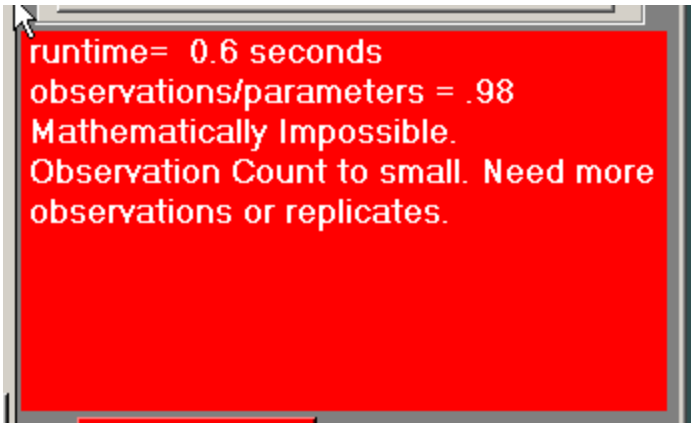

Figure 34. Screen capture. Impossible observation to parameters ratio notification.

If we deselect a couple of wave numbers, we get a yellow screen (figure 35).

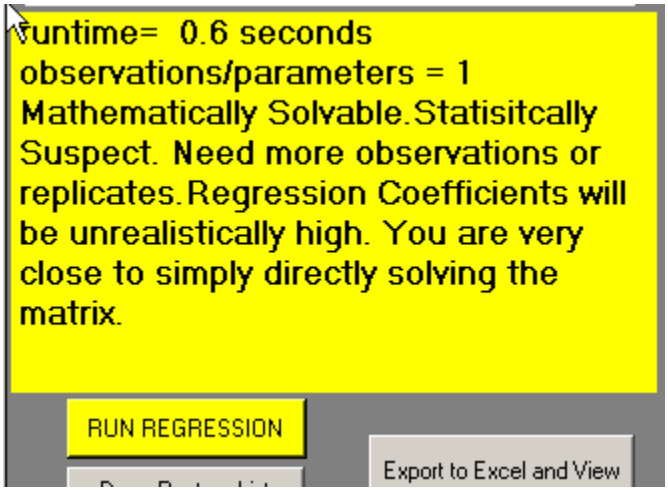

Figure 35. Screen capture. Possible observation to parameters ratio notification, but statistically suspect.

An attempt to run the correlation failed in the dynamic linked library as the matrix still could not be solved (in theory, this would have been soluble if there were no collinear variables). This situation raises an error message that is displayed to the user in figure 36. When you get this message, you must either choose fewer independent variables, group the independent variables to reduce the count, run more experiments to get more observations, or synthetically produce more observations.

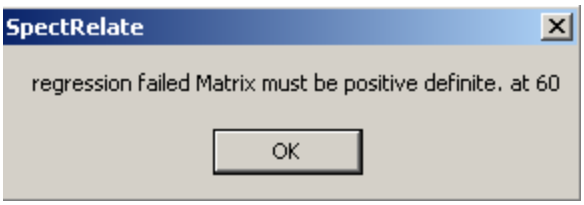

Figure 36. Screen capture. Computation failure error message.

We deselect a few more wave numbers. With 33 independent variables, the O/P ratio is still yellow at 1.15, but maybe we can solve the matrix anyway. The program runs, and the user is warned (figure 37).

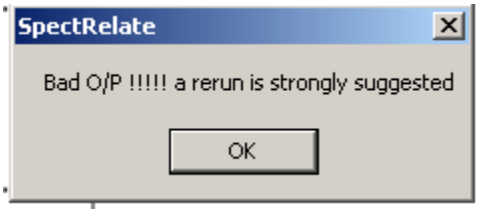

Figure 37. Screen capture. Successful computation, with suspect results warning.

The results are simply just stunning! (figure 38).

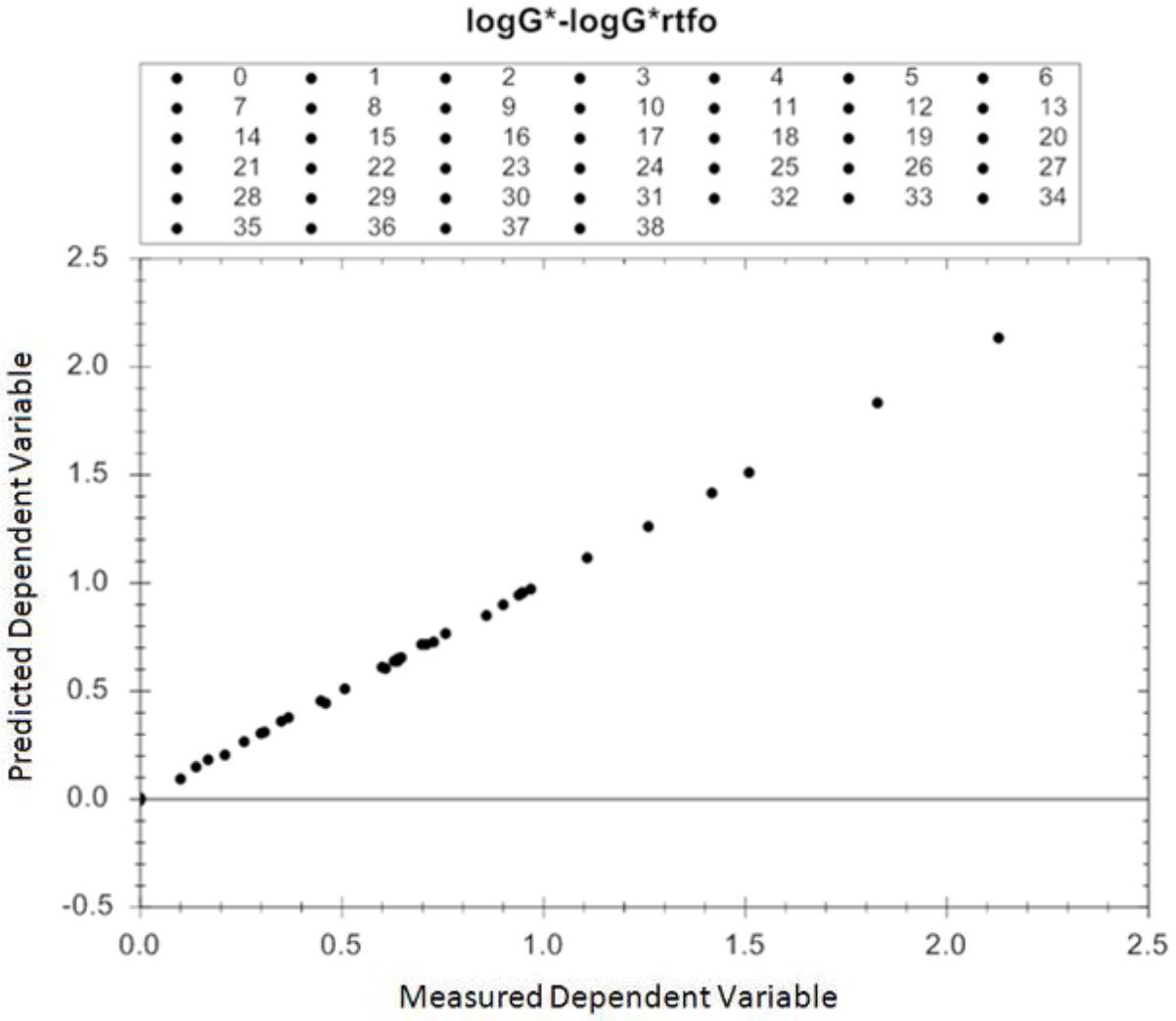

Figure 38. Screen capture. Regression results plot example 3.

The regression coefficient is .9998. With such a great correlation, why is the user warned? (This phenomenon, is often the reason PLS, CPA, and other chemometric methods produce such high correlation coefficients.) The easiest way to explain the problem here is to consider a simple linear regression with just two points in the data set. Any fit of a straight line to two points will yield a perfect correlation coefficient. In multidimensional space, the O/P ratio can be thought of as the number of points *in each dimension*. In this example, we have essentially drawn a straight line through two points in multiple dimensions. This results in a nearly perfect correlation! This also happens with data sets consisting of randomly generated data. We have two options at this point. We can examine the F test values for each variable and discard those that are insignificant. This is most easily accomplished using the auto deselect button. We now have 20 variables and 21 parameters in our model. The O/P ratio is now 1.86, a bit better than the 1.15 of the previous, but still far from being statistically believable. We still get a near perfect regression. The F test limit can be adjusted from the default value of 2. Let's try an F test limit of 3. The variable count

drops to 18, O/P still only 2. An F test limit of 10 gives us an O/P of only 2.44. At an F test limit of 20, we have an O/P of 6.5, with only the 5 best fitting variables. The results are summarized below in figure 39.

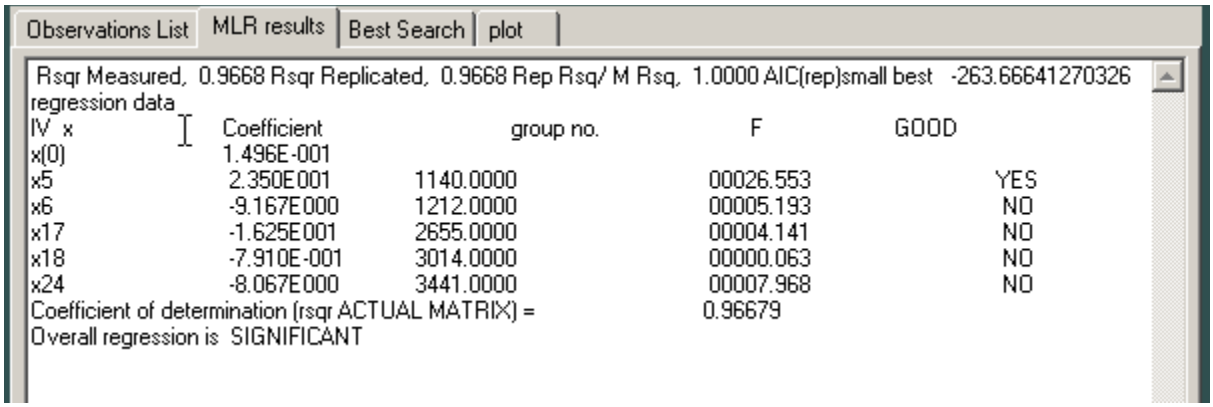

Figure 39. Screen capture. Regression results summary tab example 1.

The other option would be to use synthetic replication. Up to this point we have only used classic multivariable correlation methods and they remain at the heart of our method. Out of over 3000 available measurements in the spectra, classical methods limit us to examining just a few at a time. In this example, we can only examine additive relationships using about 1% of the data in order to get a calculation to be possible. The added constraint of statistical validity lowers that number even further to a fraction of a percent, so, the chance of missing an important combination is quite high.

Returning to the original variable list, We enable the synthetic replication and adjust the total replicates to 7, resulting in an O/P of 7.18. The correlation coefficient is now 0.9749, slightly higher than the 5 variable results we obtained before. But now we have 37 variables, most of them insignificant (figures 40 and 41).

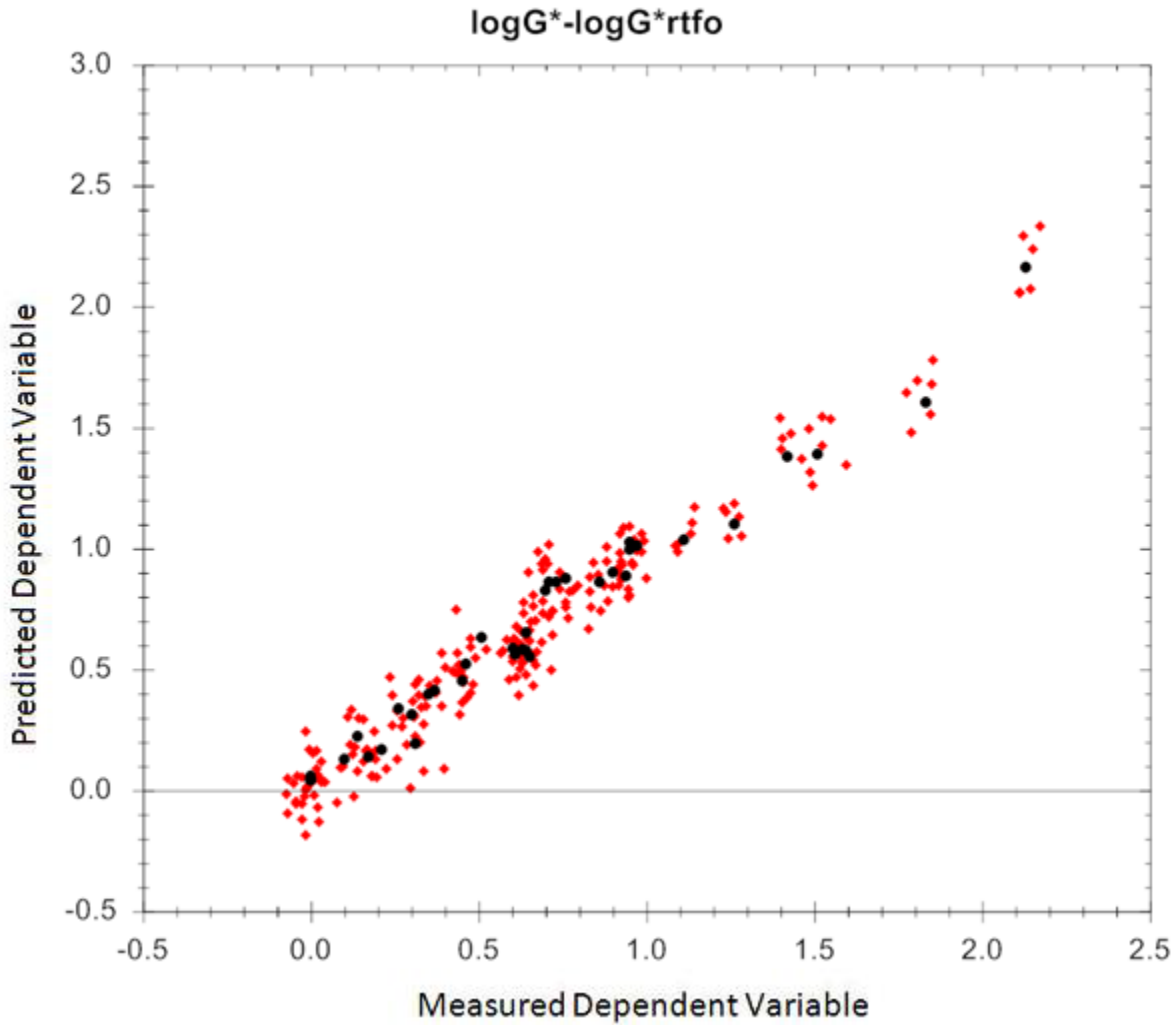

Figure 40. Screen capture. Regression results plot example 4.

Observations List | MLR results | Best Search | plot

Rsqr Measured, 0.9749 Rsqr Replicated, 0.9505 Rep Rsq/ M Rsq, 0.9750 AIC(rep)small best -2275.78771472376

regression data

IV x	Coefficient	group no.	F	GOOD
x(0)	4.465E-002			
x1	-7.265E-002	0984.0000	00000.078	NO
x2	1.696E000	1036.0000	00130.507	YES
x3	-1.178E-001	1094.0000	00000.248	NO
x4	5.895E-001	1135.0000	00006.364	NO
x5	9.084E-001	1140.0000	00014.804	NO
x6	4.913E-001	1212.0000	00004.229	NO
x7	1.084E-001	1298.0000	00000.213	NO
x8	-2.222E-001	1353.0000	00000.889	NO
x9	8.022E-001	1663.0000	00011.338	NO
x10	1.829E000	1701.0000	00118.286	YES

Figure 41. Screen capture. Regression results summary tab example 2.

We can now use the Auto Best IV feature to reduce the model to any desired number of variables, and for comparison with the previous results, we will go down to 5 (figure 42).

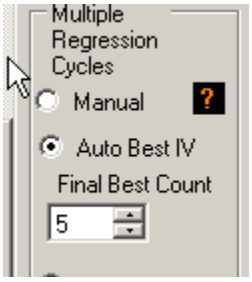

Figure 42. Screen capture. The Auto best independent variable list run option.

The results are shown figure 43.

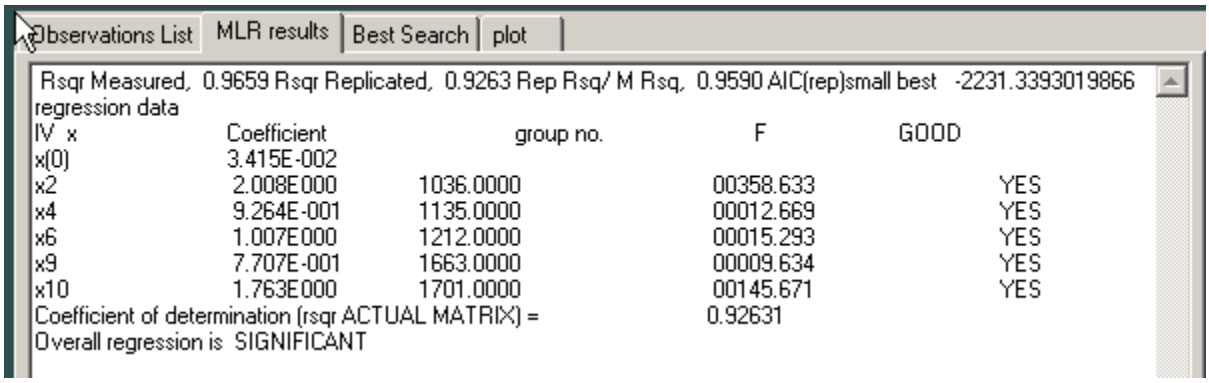

Figure 43. Screen capture. Regression results summary tab example 3.

The regression correlation coefficient we obtained is nearly identical using classic multivariable regression, and there is some agreement concerning the wave numbers of importance. Some of the wave numbers appearing in the two results are likely correlated, and describe the same measurement. We have not used grouping before the correlation, and that minimizes, limiting complexity. In theory, we could apply synthetic replication to this data set without grouping first. With some data sets, that contain few independent variables, this makes sense, but with this set, the size of the matrix that must be solved would be excessive, and very long execution times would result. In addition to reducing the size of the matrix for solution, grouping also eliminates the problem of several areas of the spectra representing the same phenomena.

To reduce the independent variable list, we chose one of the grouping methods in the IV reduction selection list (figure 44).

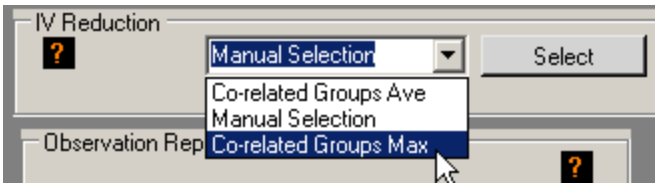

Figure 44. Screen capture. Independent variable grouping options.

We have examined the use of the Manual selection option for classic multivariable regression solutions, and the possibility to use synthetic replication to advantage without grouping first. The two grouping options work the same, as described earlier with the only exception of how the group measurements are reduced to a single number. The two methods either average the values of the wave numbers in that group, or find the largest average measurement (largest wave number over the entire range of data measured in that grouping. This can vary in individual spectrum. The idea here is to obtain the best signal to noise ratio) and use the value for that single wave number. For our studies using IR independent variable, we have found slightly better correlations using the highest peak method, but one could imagine situation where an average may better suit the physics and/or chemistry behind the relationships sought.

When a method is selected, the user is prompted for a regression threshold. For reasons discussed previously (see figure 10), this data set should probably use a value no larger than 0.9 (figure 45). The user may use higher values for more groups and selectivity.

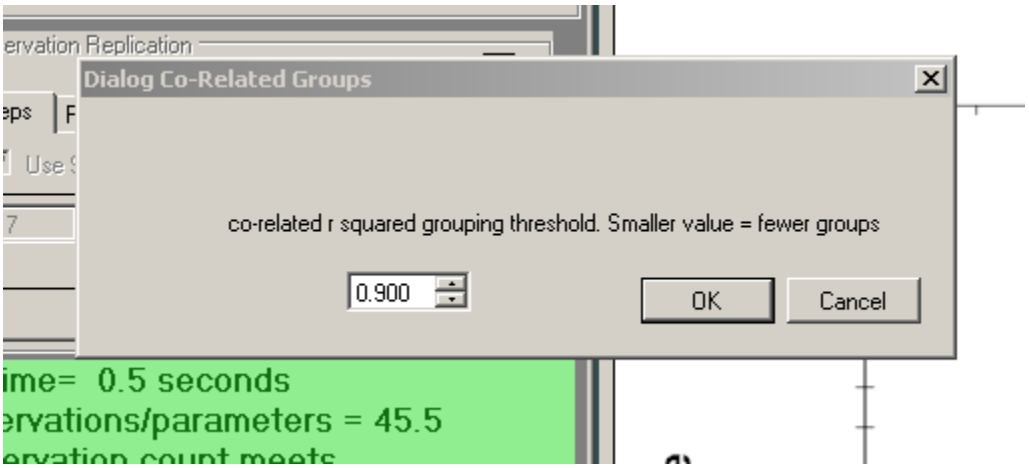

Figure 45. Screen capture. Independent variable grouping threshold setting.

When the prompt is dismissed by clicking “OK” the variables are grouped. The list of groups appears in the consideration set, with 36 groups identified. The ratio is only 1.05, and this might actually run, but would be statistically suspect, so enabling synthetic replicate at a value of 7 sets us up for a meaningful regression. Using all of the groups, we get a correlation coefficient of 0.9675. Deselecting the F test values less than 2 we reduce the number to 15 independent

variables. Running the regression on this smaller model actually produces a slightly better fit with a correlation coefficient of .9691. Eliminating F-tests less that 2, we have 8 variables to fit, resulting in a correlation coefficient of 0.9688, essentially as good as fit as 15. With F test elimination still set at 2, we reduce the number of variables in the model to 6 with essentially as good a fit. Continuing the process of discarding the variables one at a time based upon the lowest F-test, we get these results for 5 variables in figure 46. The "Rsqr Measured" and reported by the software, is the value obtained without replication. The "Rsqr Replicated" includes the replicated data. Next is the ratio of the two. AIC is a statistical measure of model size we found to be either useless or miscalculated. "rsqr ACTUAL MATRIX" is identical to "Rsqr Measured" with the exception of being computed by the regression programming object while the "Rsqr Measured" is calculated by code written by the software authors.

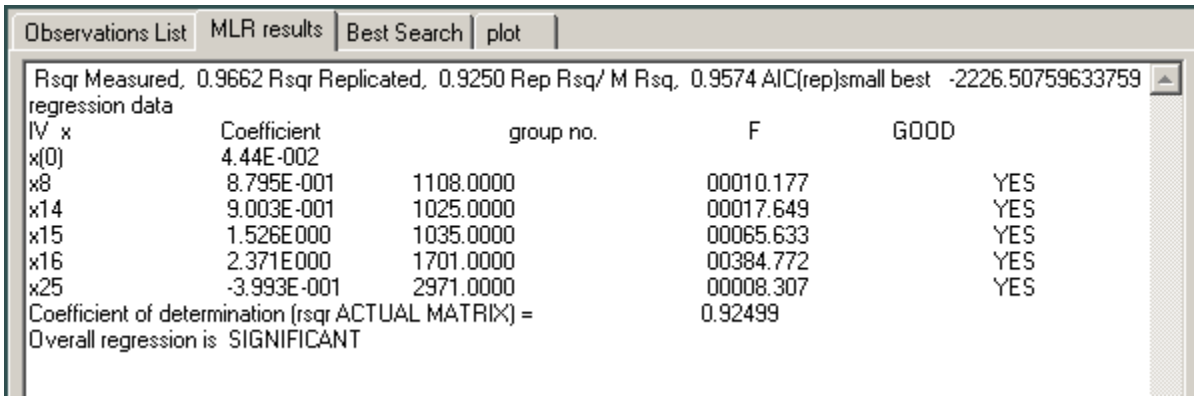

Figure 46. Screen capture. Regression results summary tab example 4.

And for 4 variables there is little reduction in the quality of the fit (figure 47).

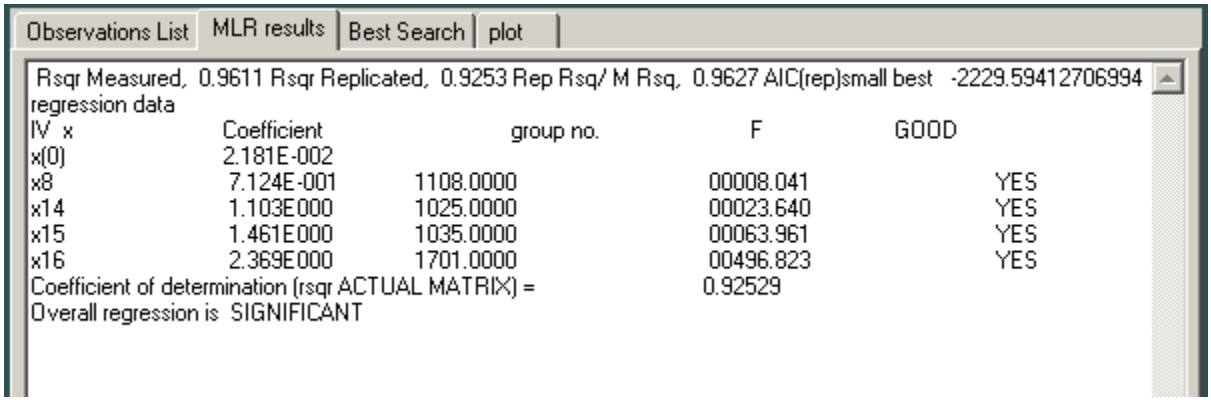

Figure 47. Screen capture. Regression results summary tab example 5.

The 3 most significant wave numbers are (figure 48):

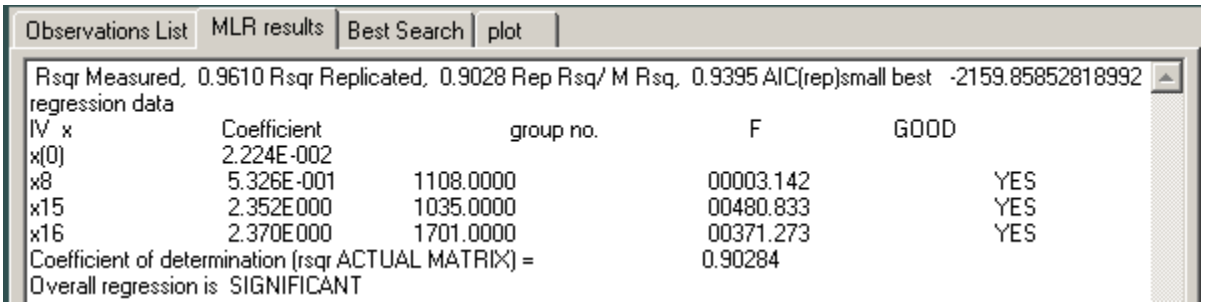

Figure 48. Screen capture. Regression results summary tab example 6.

The best 2 variables are (figure 49):

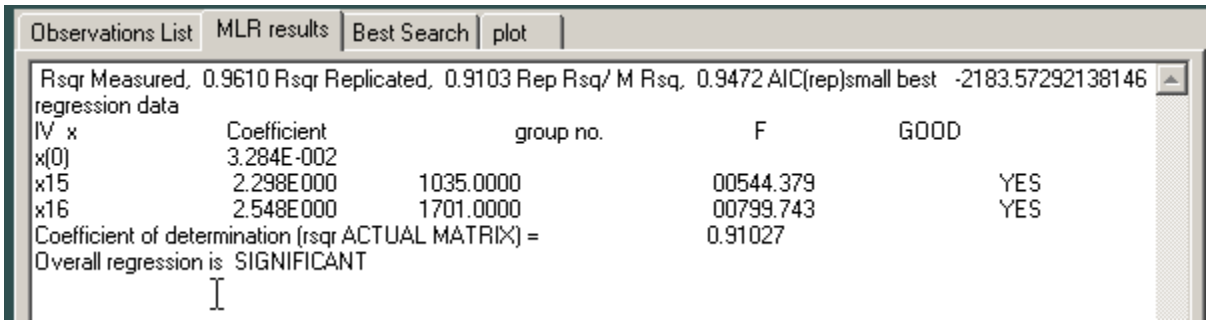

Figure 49. Screen capture. Regression results summary tab example 7.

These two wave numbers are generally accepted to represent carbonyl and sulfoxide responses in asphalt binder oxidation. The fit coefficients are nearly the same for these two wave numbers. A simple sum of these absorbencies should work well as oxidation extent of reaction indicator.

Results and Output

When the user is satisfied with the regression results, the software provides a number of options to export the data for use in spreadsheets and publications.

A right mouse click in many areas of the program interface produces a context menu for cut and paste operations of highlighted text (in blue). The following screen capture illustrates a copy operation from the “MLR results” tab (figure 50).

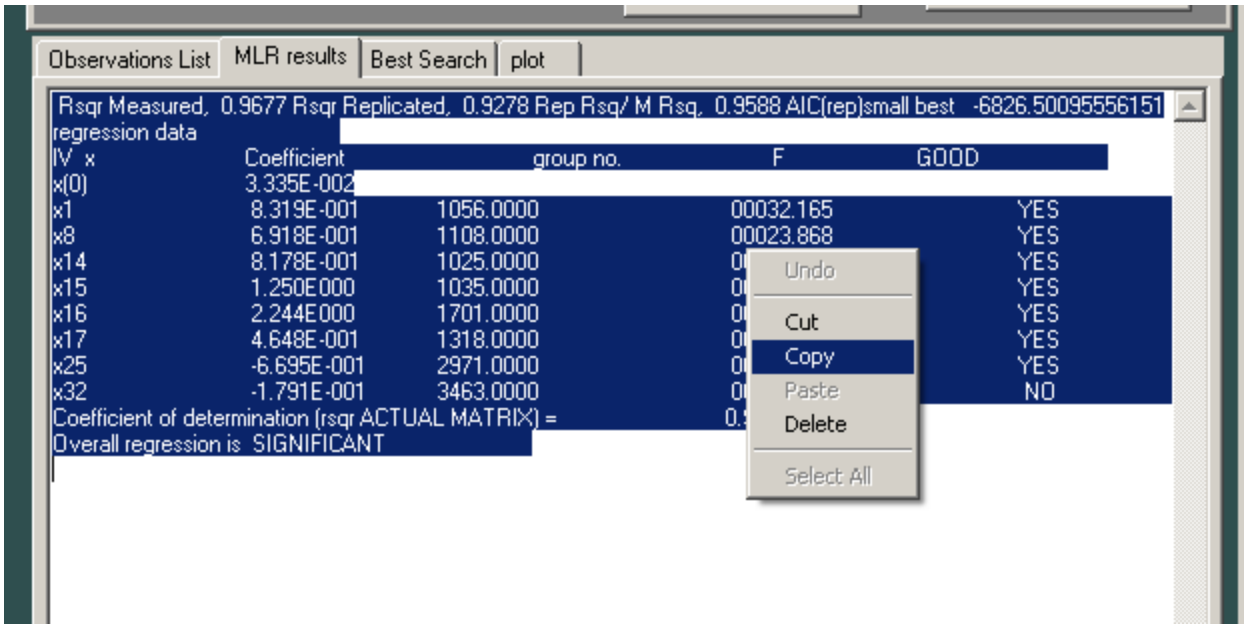

Figure 50. Screen capture. Regression results summary tab copy and paste example.

For a more detailed document for archiving purposes, the MLR Summary tab in the upper left corner fully documents the run with data, files used, and an ANOVA table. This information (must be highlighted) can be copied through the right click context menu (figure 51).

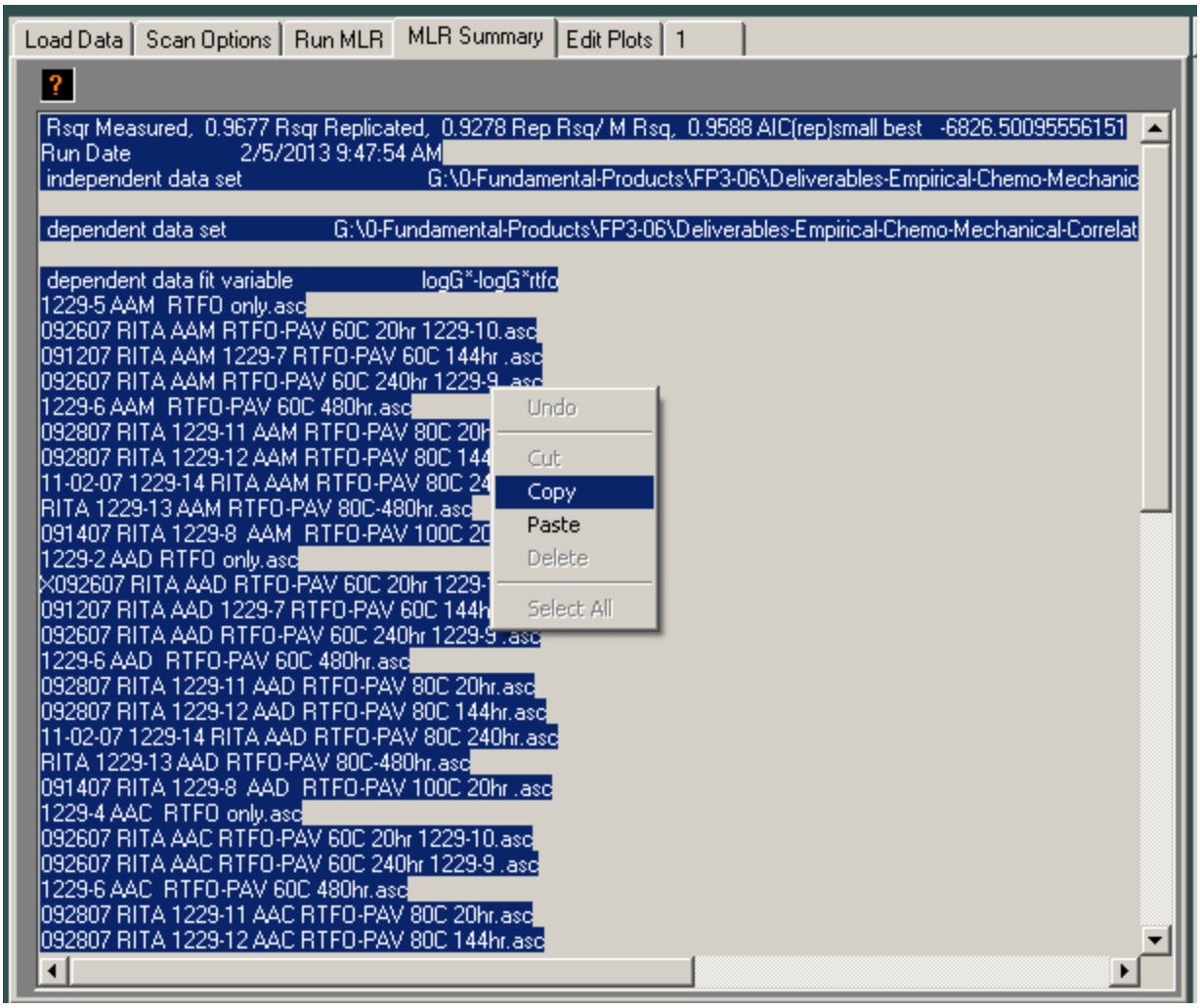

Figure 51. Screen capture. Complete run summary results copy and paste example.

The comparison plot of the regression predicted dependent variable and the actual measured values can be exported as an image, or the data can be exported to a spreadsheet for re-plotting with other software (figure 52).

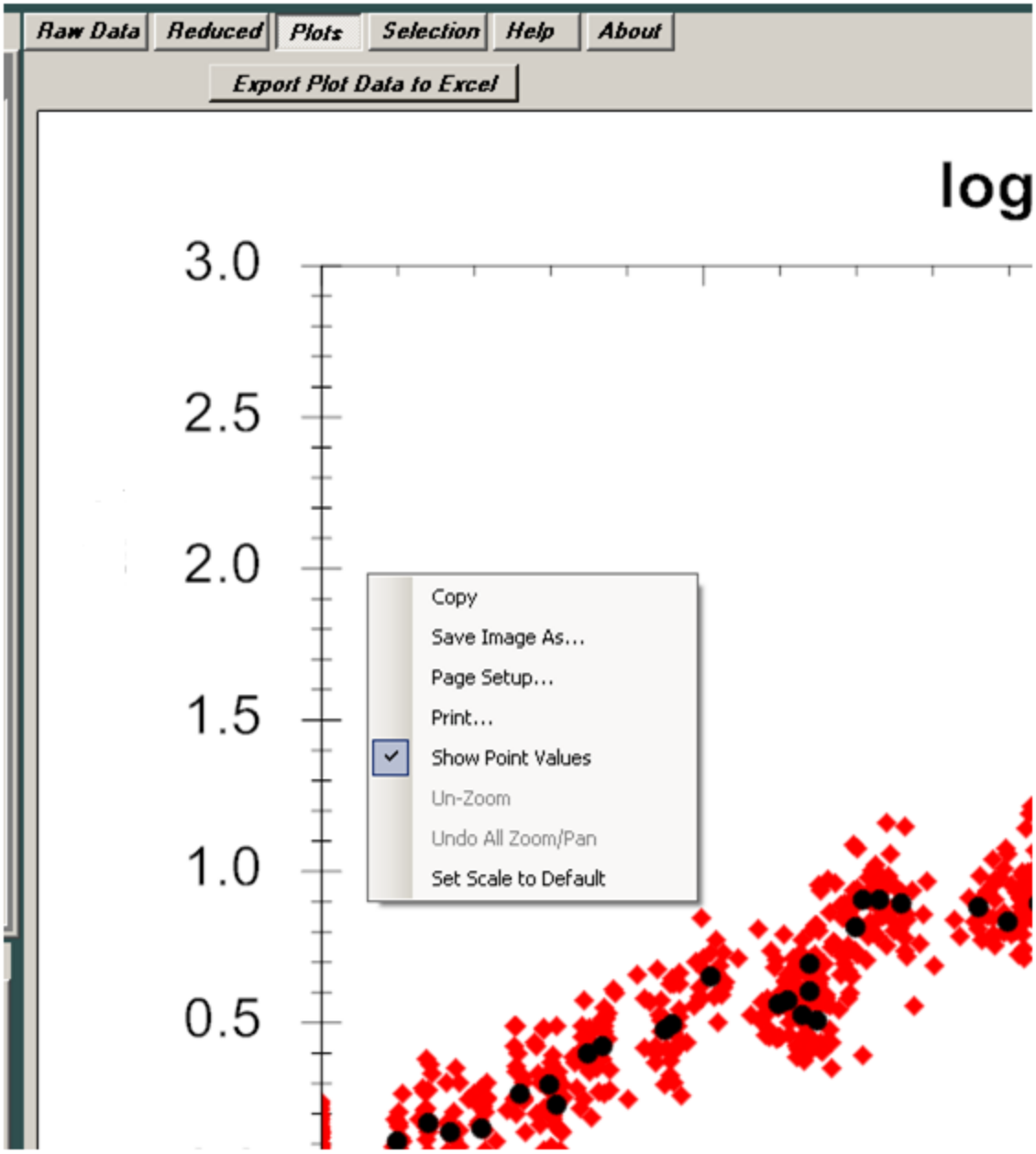

Figure 52. Screen capture. Right click context menu for plots.

The Edit Plots tab is where the plot appearance can be edited to suit the user's needs (figure 53).

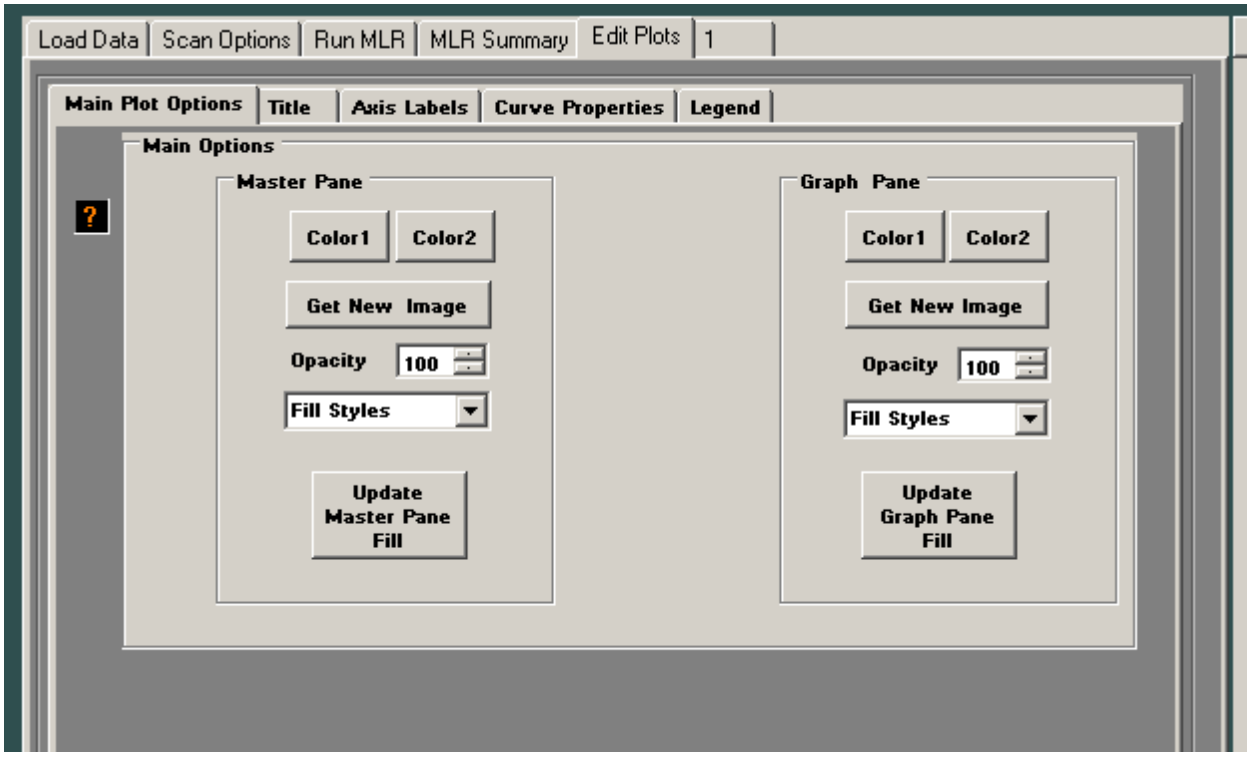

Figure 53. Screen capture. Plot attribute editing area.

RESULTS AND DISCUSSION

The use of time honored multivariable linear regression with provisions to deal with excessive independent variables based upon measurement method precision has been developed. This has proven to be extremely useful in our inquiries into asphalt binder oxidation chemistry. A software package, called Spectrelate has been developed that guides the user with regard to avoiding common errors often encountered by users of multivariable methods without the usual training normally required to understand statistical validity.

CONCLUSIONS

Multivariable empirical approaches to data analysis are a valuable tool for discovering significant variables involved in chemo-physical processes. This software package has proven quite useful in the study of rheological changes due to asphalt binder oxidation, and the details of these significant advances can be found in other fundamental properties product technical white papers. The product has many other potential applications in other areas of scientific investigation and provides a number of advantages to other often employed methods. These advantages include the production of an easily understood relationship in terms of measured variables using a time honored method with a rich history of statistical validity study.

RECOMMENDATIONS

The methods employed here could be expanded to non-linear regression problems.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Federal Highway Administration, U.S. Department of Transportation, for financial support of this project under contract no. DTFH61-07D-00005.

DISCLAIMER

REFERENCES

Barros, A. S., and D. N. Rutledge, 2004, Principle Components transform-partial least squares: a novel method to accelerate cross-validation in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 73 (2): 245-255.

Basu, B., D. Saxena, V. Kavl, M. I. S. Sastry, and R. T. Mookken, 1998, Prediction of Oxidation Stability of Inhibited Base Oils from Chemical Composition using an Artificial Neural Network (ANN). *Lubrication Science*, 10: 121-134.

Chalmers, J. M., and N. J. Everall, 1996, FTIR, FT-Ramen and chemometrics: applications to the analysis and characterization of polymers. *Trends in analytical chemistry*, 15 (1): 18-24.

Garson, G. David (n.d.). "Popular algorithms," from Statnotes: Topics in Multivariate Analysis. Retrieved 11/21/2007 from http://www2.chass.ncsu.edu/garson/PA765/statnote.htm .

Hasegawa, T., 2006, Spectral Simulation Study on the Influence of the Principle Component Analysis Step on Principle Component Regression. *Applied Spectroscopy*, 60 (1): 95-98.

Hasegawa, T., 1999, Detection of Minute Chemical Species by Principal-Component Analysis. *Anal. Chem.*, 71 (15): 3085-3091.

Karstang, T. V., A. A. Christy, B. Dahl, and O. M. Kvaheim, 1991, Diffuse reflectance Fourier-transformed infrared spectroscopy in petroleum exploration: a multivariate approach to maturity determination. *Journal of Geochemical Exploration*, 41 (1-2): 213-226.

Lachenmeier, D. W., 2007, Rapid Quality control of spirit drinks and beer using multivariate data analysis of Fourier transform infrared spectra. *Food Chemistry*, 100 (2): 825-832.

Mark, H., and J. Workman Jr., 2007, Chemometrics in Spectroscopy What Can NIR Predict? *Spectroscopy*, 22(6): 20-26, www.spectrocopyonline.com, Nov. 21, 2007.

Pearson, K., 1901, On lines and planes of closest fit to systems of points in space. *Phil. Mag. Ser. 6*, 2: 559-72.

Sastry, M. I. S., A. Chopra, A. S. Sarpal, S. K. Jain, S. P. Srivastava, and A. K. Bhatnagar, 1998, Determination of Physiochemical Properties and Carbon-Type Analysis of Base Oils Using Mid-IR Spectroscopy and Partial Least-Squares Regression Analysis. *Energy & Fuels*, 12 (2): 304-311.

Satya, S., 2005, *Chemometrics: A Tool to Predict Crude Oil Properties*, PhD. Dissertation, Dept. of Chemical Engineering, University of Utah.

Satya, S., R. M. Roehner, M. D. Deo, and F. V. Hanson, 2007, Estimation of Properties of Crude Oil Residual Fractions Using Chemometrics. *Energy & Fuels* , 21 (2): 998-1005.

Sharaf, M. A., D. L. Illman, and B. R. Kowalski, 1986, *Chemometrics*, Volume 82 in Chemical Analysis, Elving, P. J., and J. D. Winefordner, eds., John Wiley & Sons, New York.

Wold, S., 1991, Chemometrics, why, what and where to next. *Journal of Pharmaceutical & Biomedical Analysis*, 9 (8): 589-596.

Wold, S., H, Trygg, A. Berglund, and H. Antti, 2001, Some Recent Developments in PLS Modeling. *Chemometrics and Intelligent Laboratory Systems*, 58 (2): 131-150.

Zagonel, G. F., P. Peralta-Zamora, and L. P. Ramos, 2004, Multivariate monitoring of soybean oil ethanolysis by FTIR. *Talanta*, 63 (4): 1021-1025.