

Tel: (814) 863-0281

Fax: (814) 863-7114

The Pennsylvania State University

323 Thomas Building

University Park, Pa 16802

Statistical Consulting Center

Department of Statistics

Eberly College of Science

**SCCID**: N/A

**Client**: Penn State Statistics Department

**Consultant**:

**Date**:

**Research field:**  Environmental pollutants and Health Risk

**Project Title:** Application of SOM Artificial Neural Networks, PCA and GIS Technologies for the Characterization of Human Health Risk to Petrochemical Pollutants in the Niger Delta

**Project description:**

This study aims to determine levels of pollutants due to petroleum related activities in the Niger Delta. It also aims to categorize areas of the Niger Delta based on the levels of pollutants. The study looks at 4 different soil sites, 5 different bottom sediment sites, and 4 different water sites. Some specific goals of the research and study are to determine the concentration of heavy metals and other organic compounds in soil, sediments, and water

**Research Questions**

1. Which areas of the Niger Delta are most polluted due to petroleum activity?
2. Can the different areas be categorized by concentration of pollutants?

**Statistical Questions**

1. Could some of the pollution be attributed to situations other than those related to petroleum activities?
2. What are the relationships between the different types of pollutants?
3. Are there interactions between the different types of pollutants?
4. What sorts of properties explain the variability in soil content?

**Variables of Interest:**

Explanatory Variables: soil, soil PAH, bottom sediment, water, PBT, IRA, and background level.

Response variables: pollutant levels, concentration levels

**Study diagram**

A possible study design would include the following measurements.

**Measurements taken from water**

**4 sites**

**Measurements taken from soil**

**4 sites**

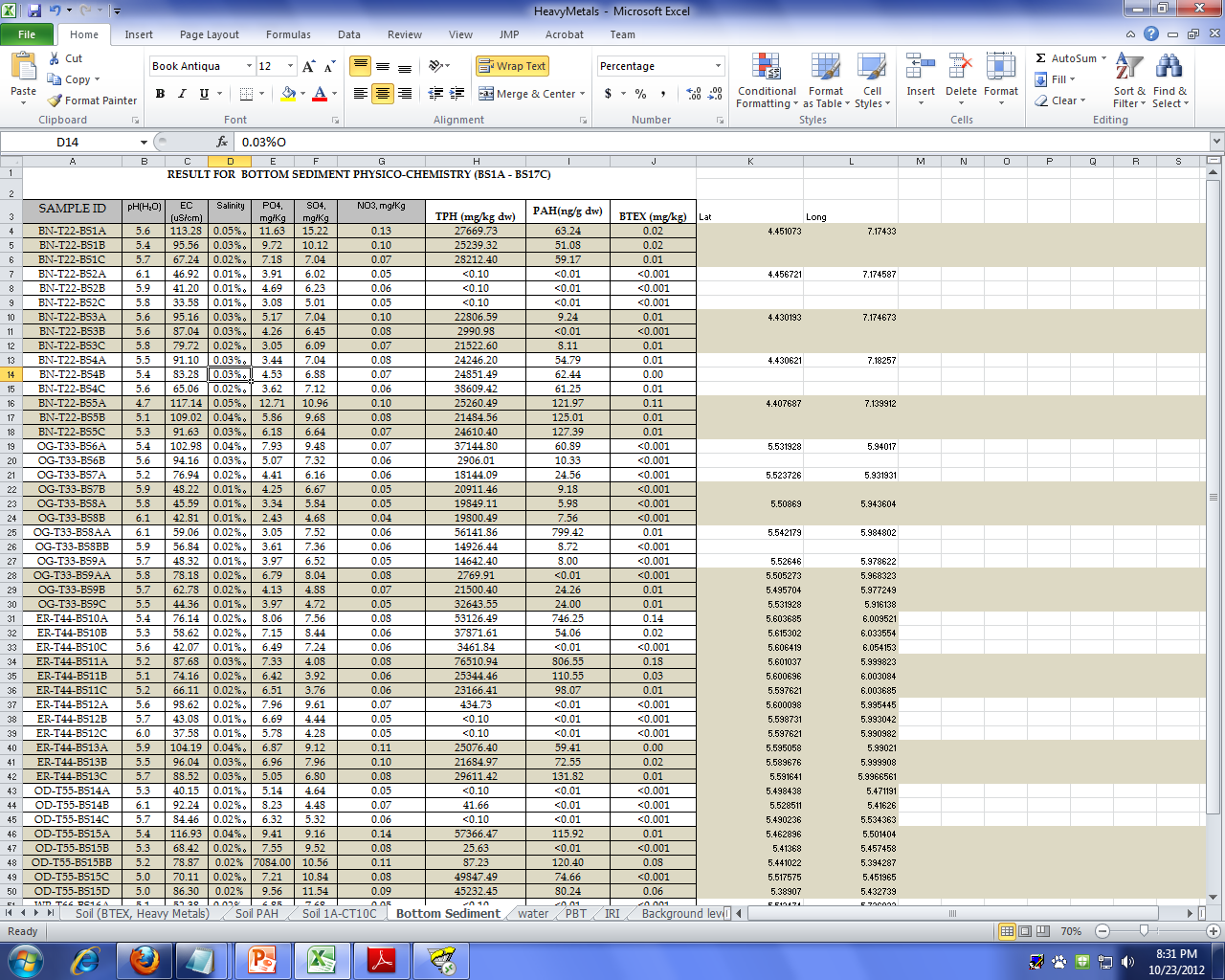
**Measurements taken from bottom sediment**

**5 sites**

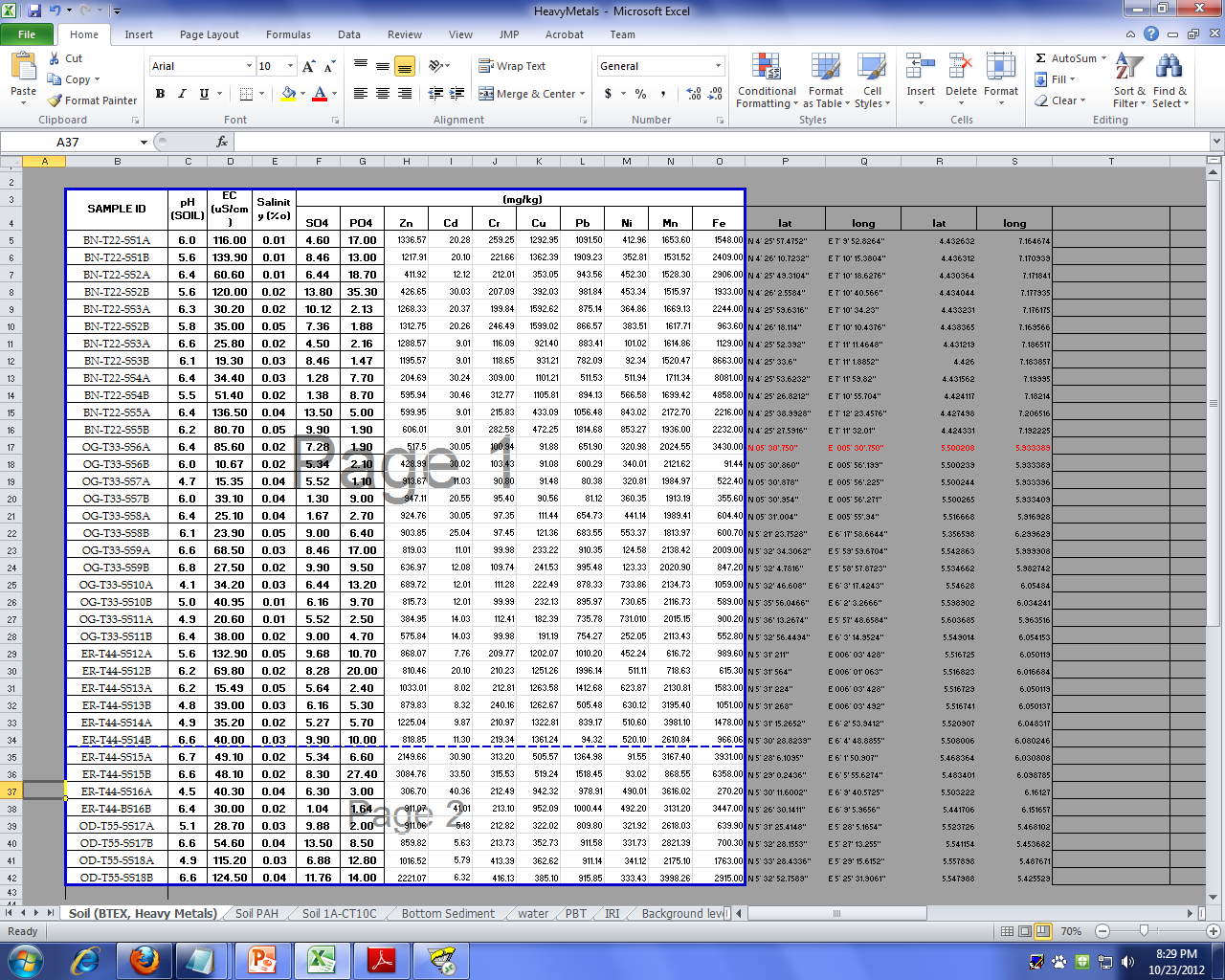
***Exploratory Data Analysis***

*Data Cleaning:*

Exploratory data analysis, or EDA, is an essential first step in understanding any set of data. The first step we took was to clean the data so that SAS could recognize the inputted data. The data cleaning consisted of removing all carrots and percent signs from the sediment data. An “un-cleaned” version of the data sheet can be seen below.



Upon attempting to run some scatterplots of the soil data set, we found that there was a typo in one of the observations, which created an additional location or ID that did not exist for the soil data.



We changed this observation ID so it read SS, not BS so that it was correct in the soil data set.

*Number Summaries:*

After cleaning the data, we began our data analysis by running some basic 5 number summaries. In SAS, the code to obtain these 5 number summaries is:

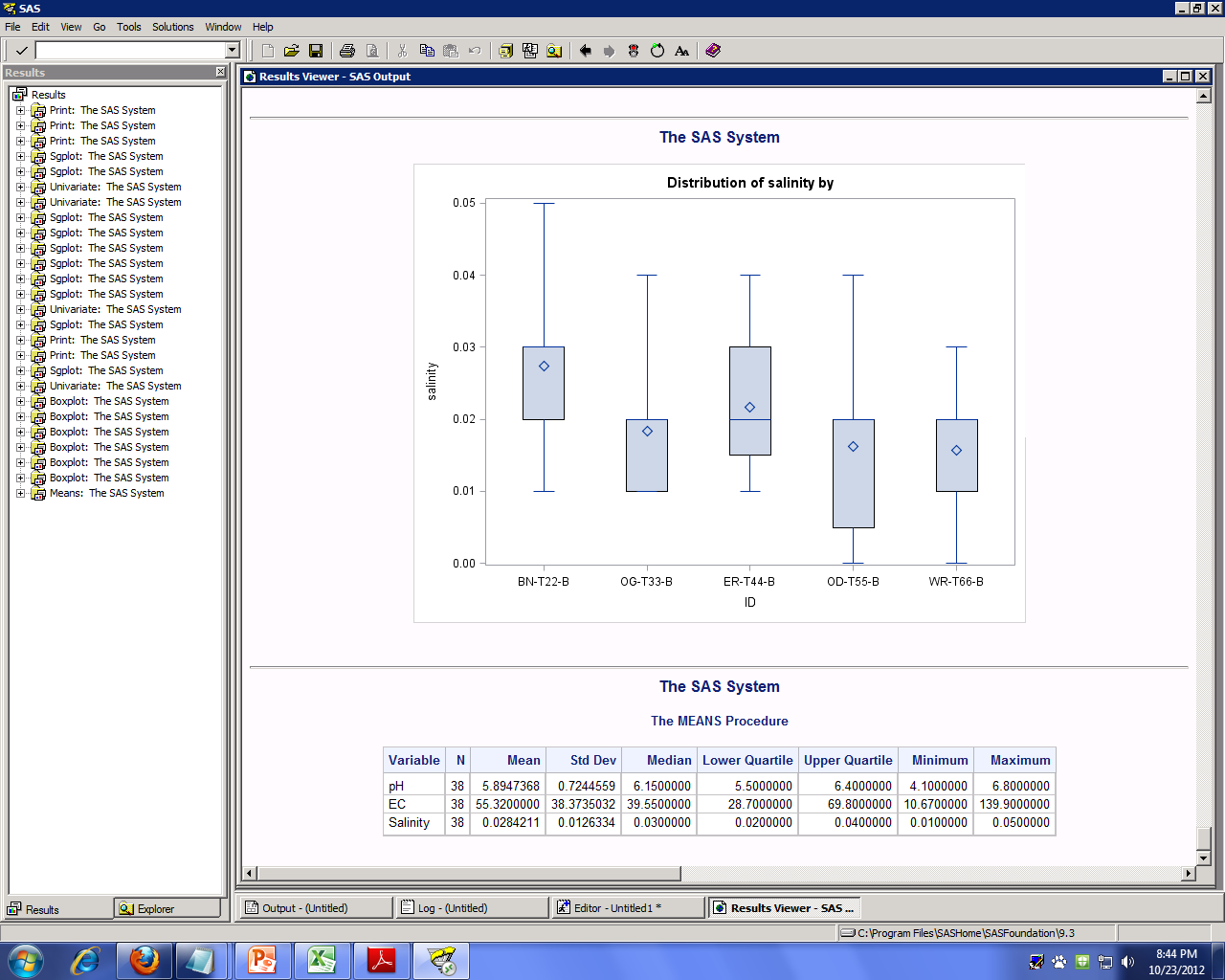
**PROC MEANS** data=soil n mean std median q1 q3 min max;

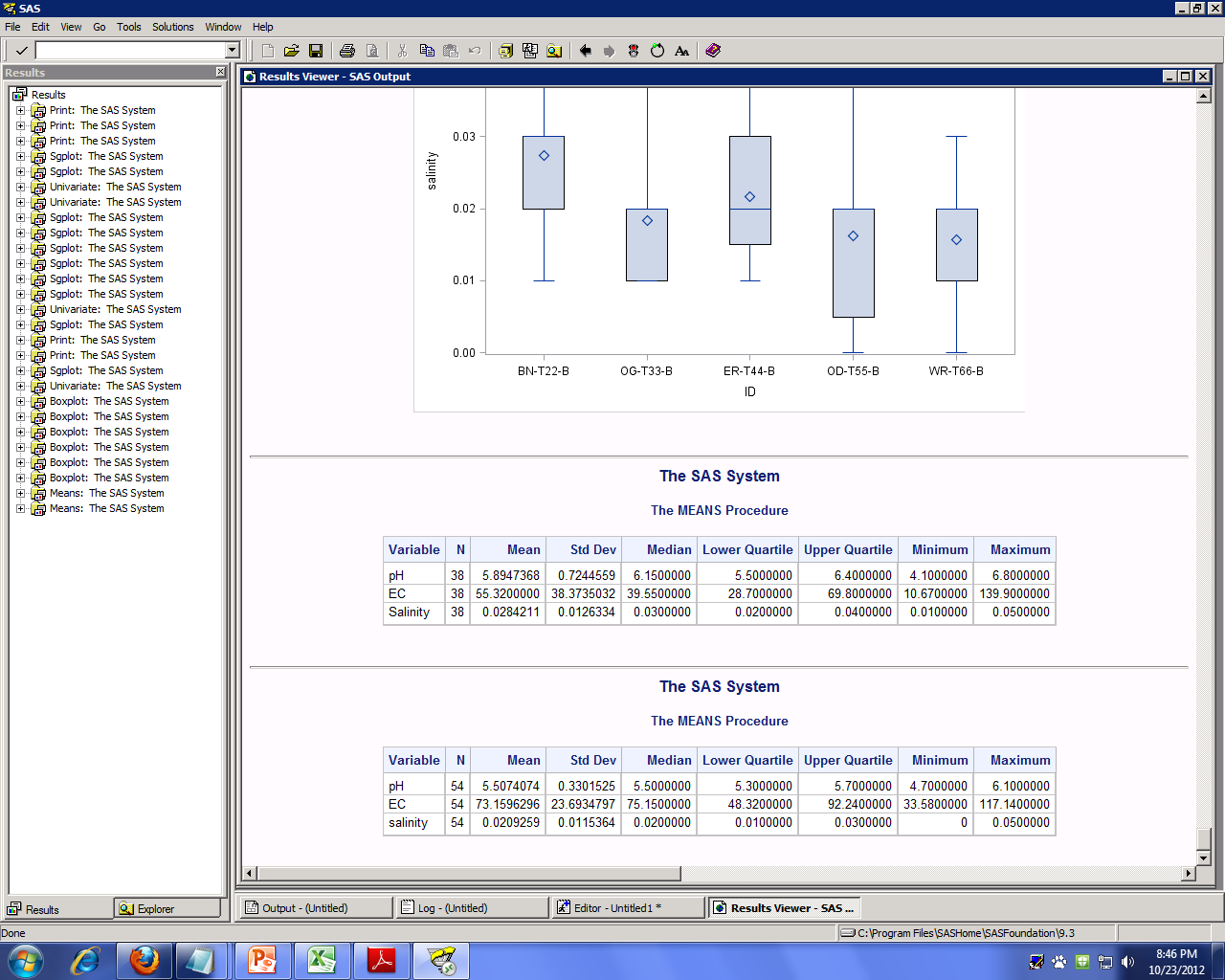
var ph ec salinity;

**RUN**;

We ran 5 number summaries for both the soil data set and the sediment data set. To obtain the sediment 5 number summary, change the “data=soil” statement in the first line of code above to “data=sediment”.

The number summaries are as follows: (the soil summary is first, then the sediment summary)





The summaries above do not give us a whole lot of information since they do not take into account the different locations for soil and sediment, which is what we are really interested in. We can see that there is a possibility that the EC variable for the soil data is not normal since the mean and median are not equal and are not very close to equal. Normality for the other variables and the other data set seem good since the means and medians are very close to equal.

*Scatterplots:*

After running some 5 number summaries, we ran scatterplots of each variable by location for each data set. The code in SAS to obtain these scatterplots is:

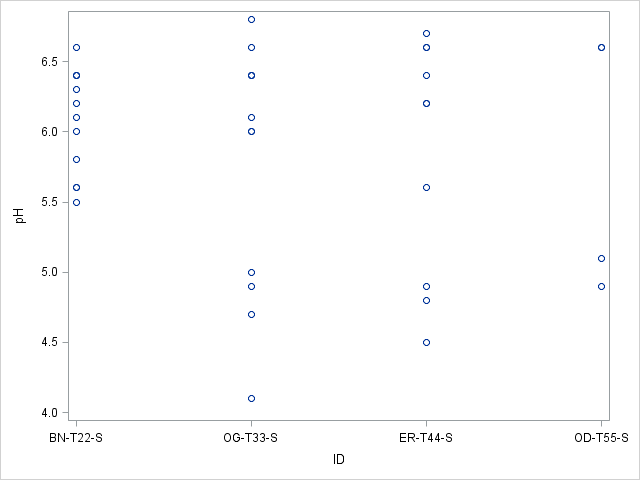
**PROC SGPLOT** data=soil;

scatter y=pH x=ID;

**RUN**;

The ID should be on the x-axis for every plot. The y-axis will change between pH, EC, and salinity. The “data=” statement will change from “data=soil” to “data=sediment”. Overall, 6 plots should be created.

The plot for pH vs. ID for the soil data can be seen below.



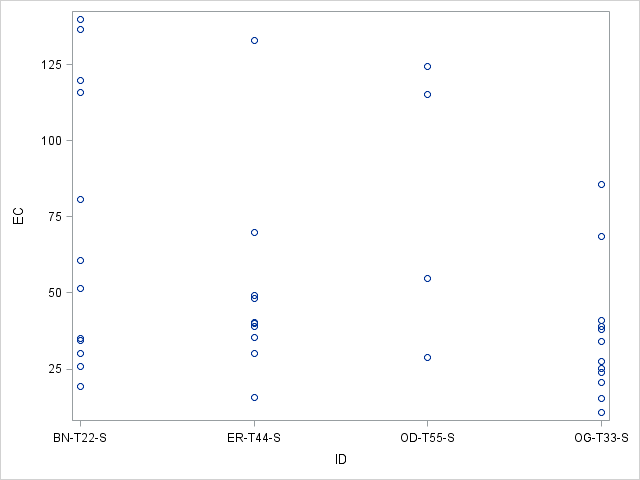
This plot shows that the pH variable for the soil data is pretty normal when compared at the different locations. The boxplots presented later in this report show a better representation of this scatterplot. Overall’ this plot shows that there isn’t really anything weird going on with the pH soil data.

The following plots all appeared very similar to the one above:

Soil: pH vs. ID, salinity vs. ID

Sediment: pH vs. ID, EC vs. ID, salinity vs. ID

The one scatterplot that looked different than the others was EC vs. ID for the soil data. This plot looked like this:



For the ER-T44-S ID, you can see that there is one value that is significantly further from the rest of the data points. This point has the arrow facing it. This point looks to be an outlier. Later, we will run our statistical tests with the outlier in the model and again with it removed, as to determine whether the point is influential or not.

*Boxplots:*

In order to make the scatterplots easier to understand, we ran boxplots for all the same combinations as above, i.e. pH vs. ID, EC vs.ID, etc…

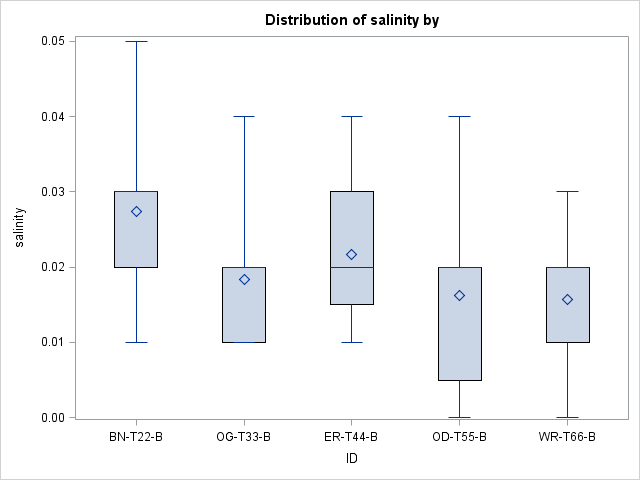
The SAS code to obtain the boxplots is:

**PROC BOXPLOT** data=sediment;

plot salinity\*id;

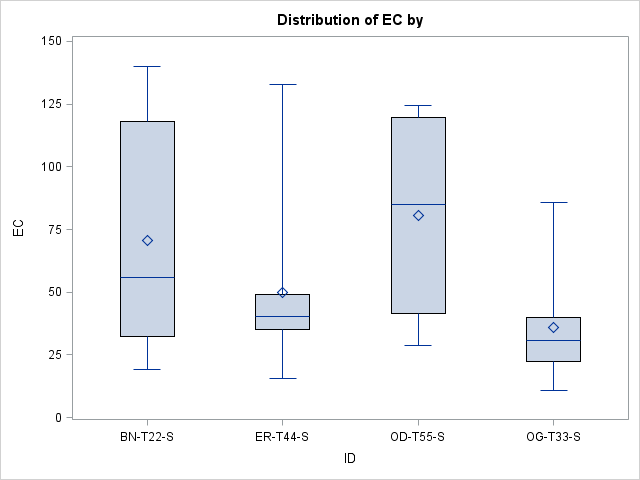
**RUN**;

This specific code yields a boxplot for the salinity variable organized by the ID for the sediment data. The boxplot obtained from this code is:



This boxplot reveals to us that the means are fairly consistent at all five locations of the sediment data. There does not seem to be any outliers or oddities about the plot.

The boxplots for all the variables looked very similar, besides the one for EC vs. ID for the soil data. This plot looked like this:



From the scatterplot for this distribution, we saw that there was a possible outlier in the ER-T44-S ID location. Also, from the 5 number summaries we saw that the data may not be normal for this variable. This boxplot agrees with what we hypothesized in the previous sections.

*Correlations:*

After running all the boxplots, we ran a correlation procedure to see how each of the chemicals in the data sets were correlated with pH, EC, and salinity. The SAS code to run the correlation procedure is:

**PROC CORR** data=sediment;

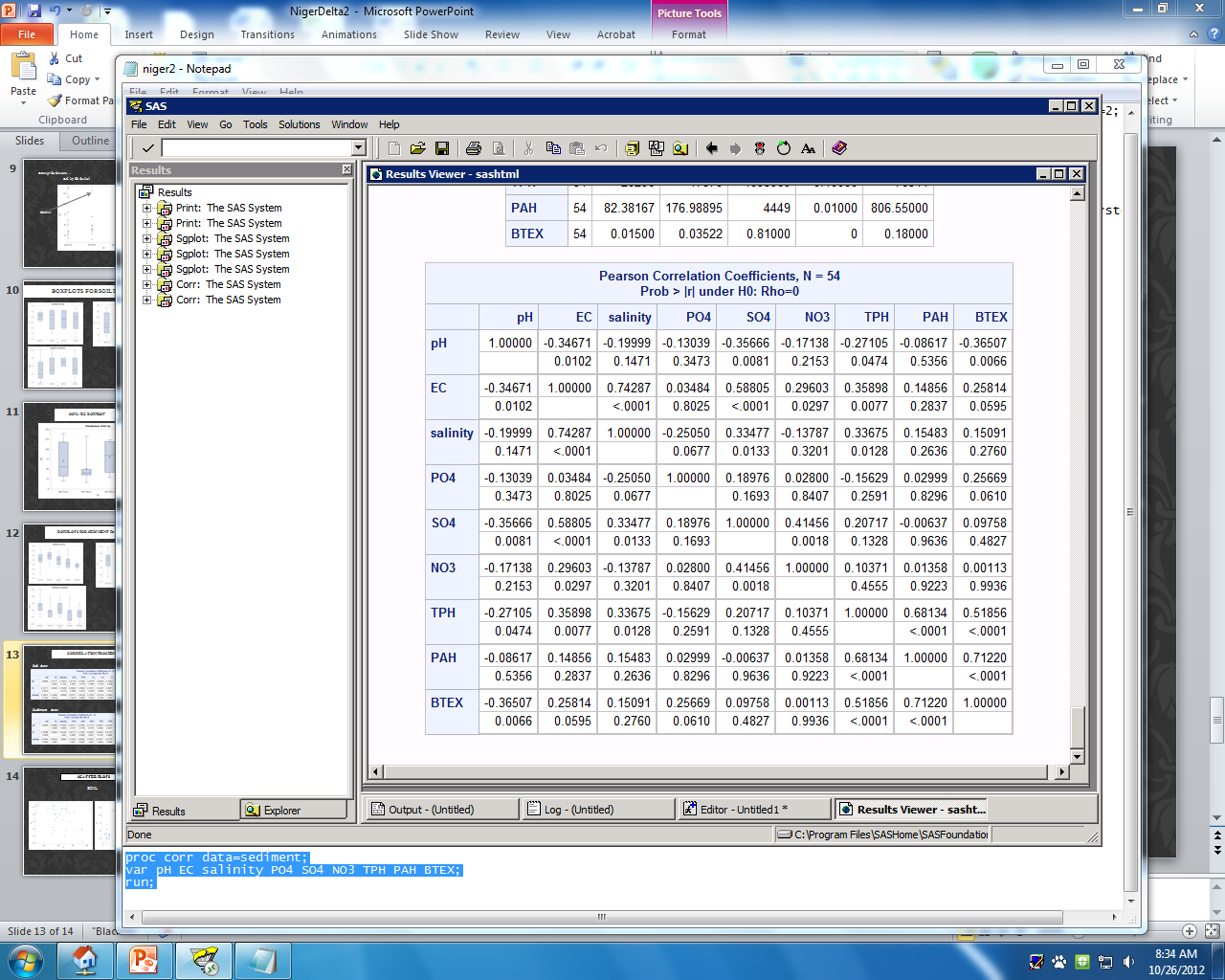
var pH EC salinity PO4 SO4 NO3 TPH PAH BTEX;

by ID;

**RUN**;

This code shows the correlation for the sediment data. For the soil data, the “data=” statement would need to be changed. Also, the “var” statement would need to be changed as well as to reflect the chemicals that are present in the soil data as opposed to those present in the sediment data.

The correlation output for the *sediment* data is:



This row shows the p-value

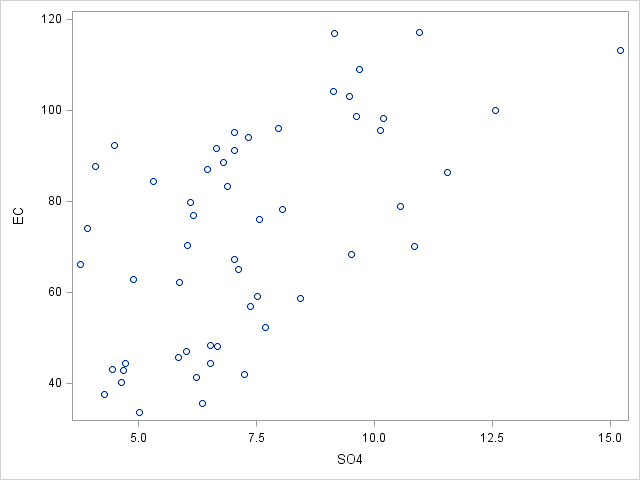
The correlation coefficient values fall between +1

This row shows the correlation coefficient

and -1. The closer to zero, the less correlated the

variables are. A negative correlation coefficient shows a negative relationship between the two variables and, likewise, a positive correlation coefficient shows a positive relationship.

The p-values mean nothing without a clear hypothesis. Here out hypothesis would be that there is no significant correlation. Our accepted significance level is 0.05, meaning that when the value is less than 0.05 we reject the null and claim that there is a significant correlation. We have several significant correlations above. One example would be EC and SO4. When we run a scatterplot of the EC vs. SO4, as follows, we see that we have a strong positive relationship, which agrees with our correlation coefficient of 0.58805 and p-value of <0.001.



***Inferential Statistics and Conclusions***

*Question 1*

In order to answer the first statistical question of whether there are differences in pollutant levels between the different locations, we can perform three separate analysis of variance procedures (ANOVA). One ANOVA is performed for each of the three response variables, pH, EC and salinity. Prior to running the ANOVA’s, we combined the sediment and soil datasets into one datasheet called both. We are interested in results by location so we combine the datasets so we have all location information in one place. An example of one of these ANOVA’s is seen below. The response variable for this ANOVA is EC. The code to obtain this ANOVA is:

**PROC** **MIXED** data= both;

class Location type;

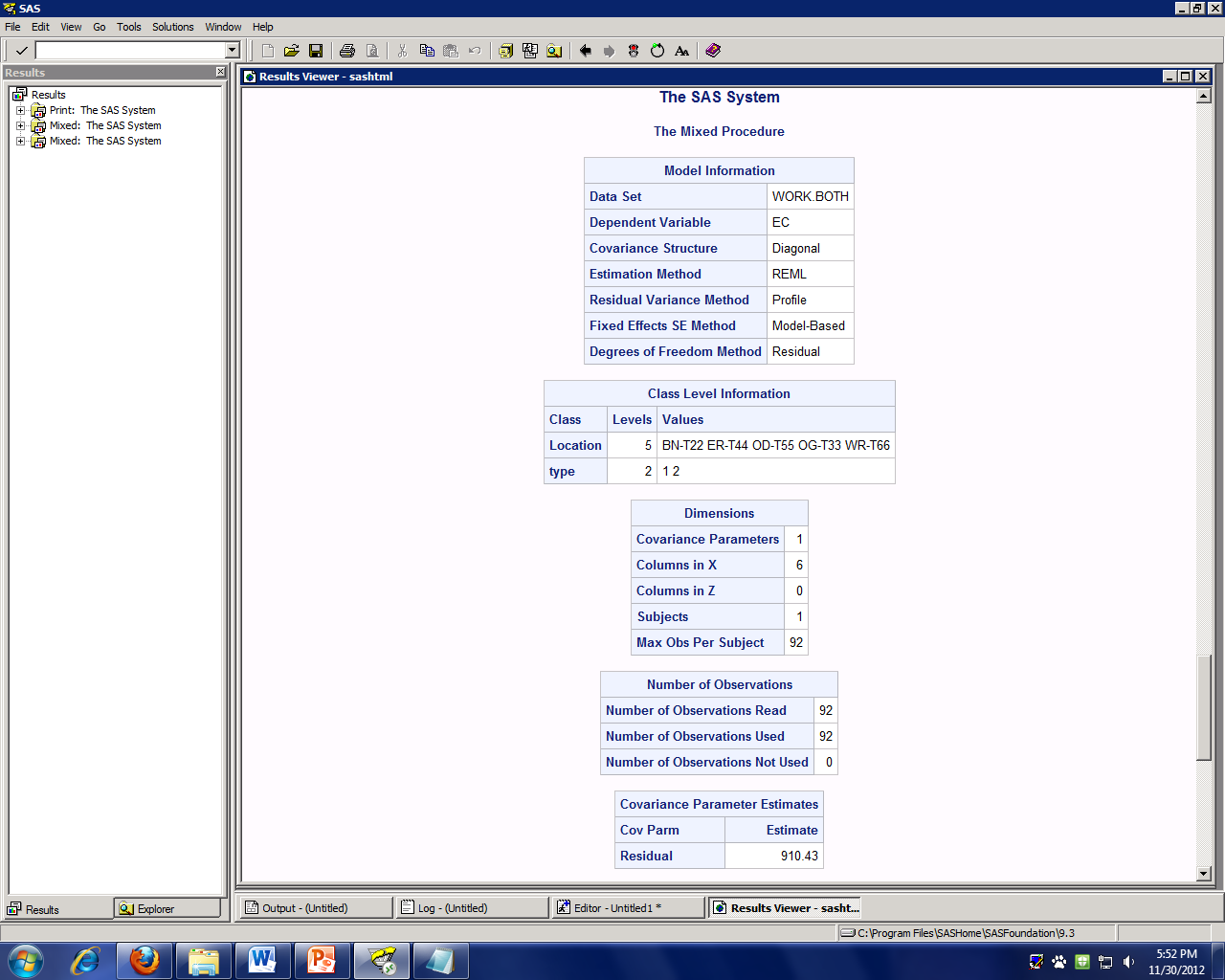
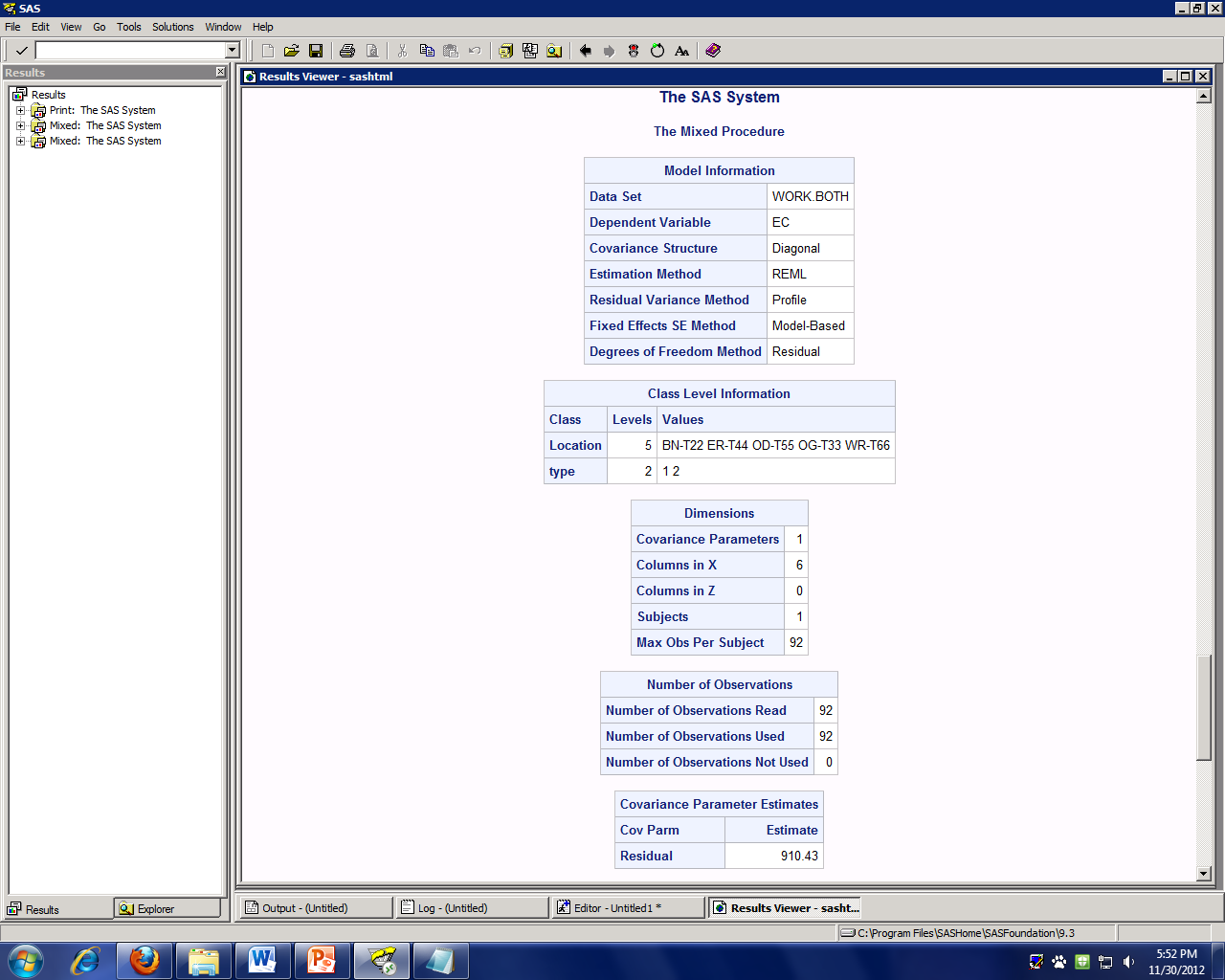
model ec = location;

lsmeans location / adjust=tukey;

**RUN**;

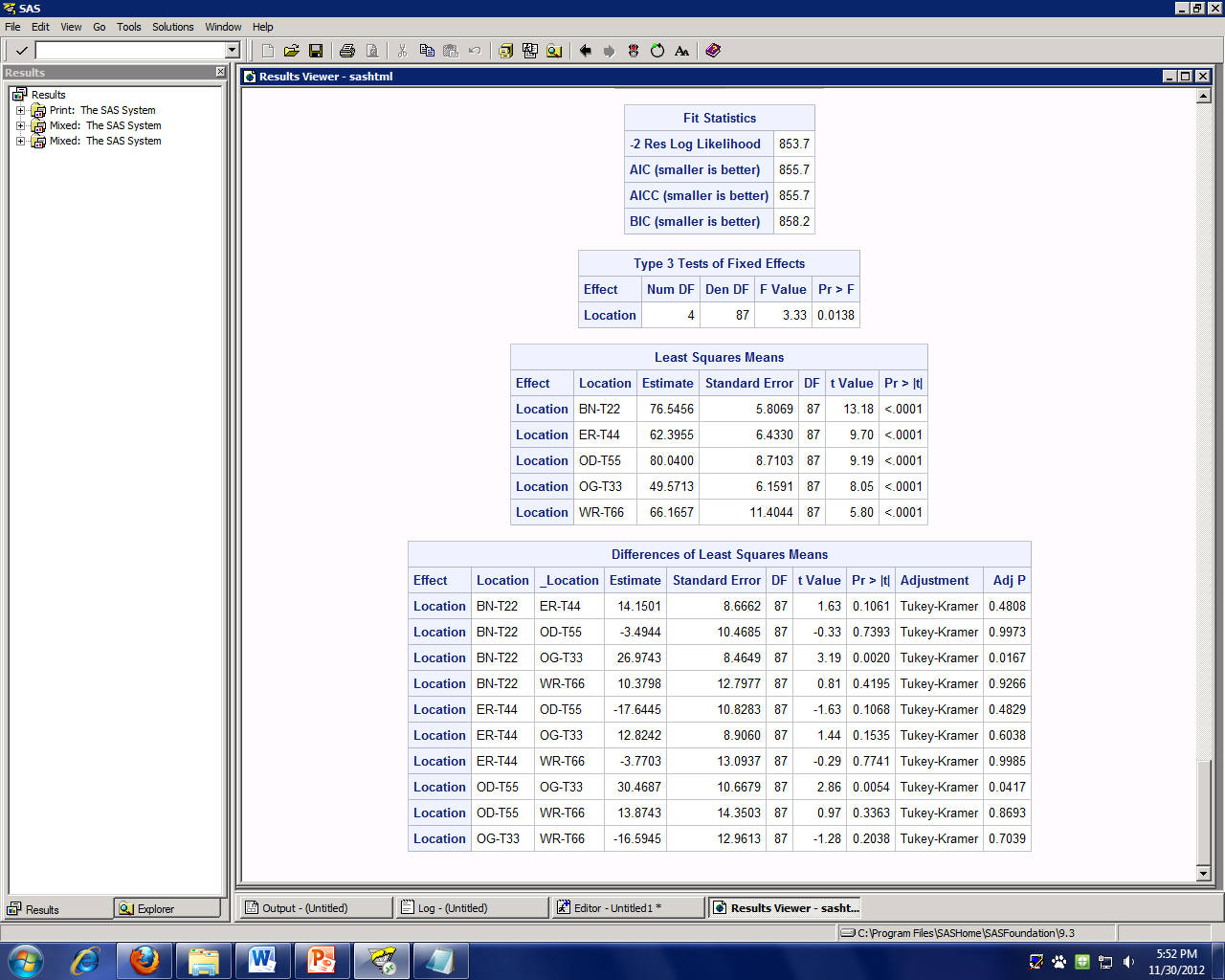
The class statement tells what variables we are interested in. The model statement shows the response variable we are interested in and sets that equal to the fixed effects, in this case the fixed effect is the location. The lsmeans statement applies a tukey adjustment to the location variable. By using this adjustment, the ANOVA is separated by location and is not clumped together. We use the same code for the other two response variables; the only difference is we change the model statement to reflect the new response variables.

The PROC MIXED output for the EC response variable is:



The above output obtained from the PROC MIXED procedure tells us what our data look like. It breaks the information into sections based on the class and model statements we provided in the code.

The output below, also obtained from the PROC MIXED procedure, gives us the numbers that are able to be analyzed.

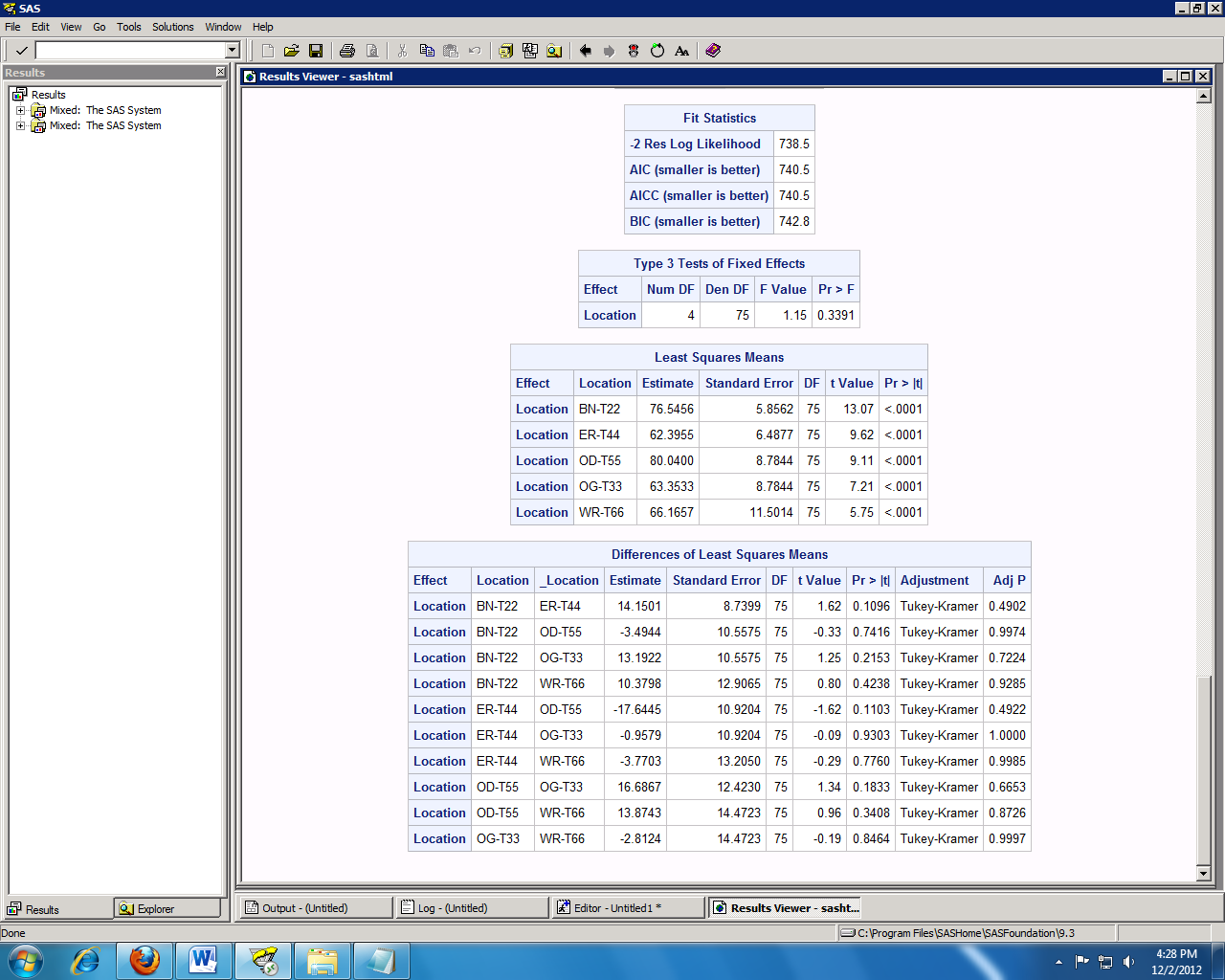


Since these values are less than the accepted value of 0.05, we claim that that there is a difference between the locations BN-T22 and OG-T33 and also between OD-T55 and OG-T33.

Since this value is less than the accepted value of 0.05, we claim that the EC variable is significant. This result agrees with the work done in the EDA section.

The common location here is OG-T33. By removing this location from the dataset and rerunning the PROC MIXED procedure with the new dataset, we see that the EC variable becomes insignificant and we also see that there are no significant differences between the locations. From this output we can claim that the OG-T33 location is statistically significant and is different than the other locations.

The output from this new dataset without the OG-T33 location is shown below. We see in this output that the overall p-value is 0.3391 so the EC variable is not significant. We also see that all the adjusted p-values are greater than 0.05 so there are no significant differences between the locations.



Overall p-value

*Question 2*

In order to answer the second question of whether concentrations of different chemicals can be used to predict overall pollutant levels, we need to perform a principal component analysis (PCA) on the sediment data and the soil data. The code used for this procedure is:

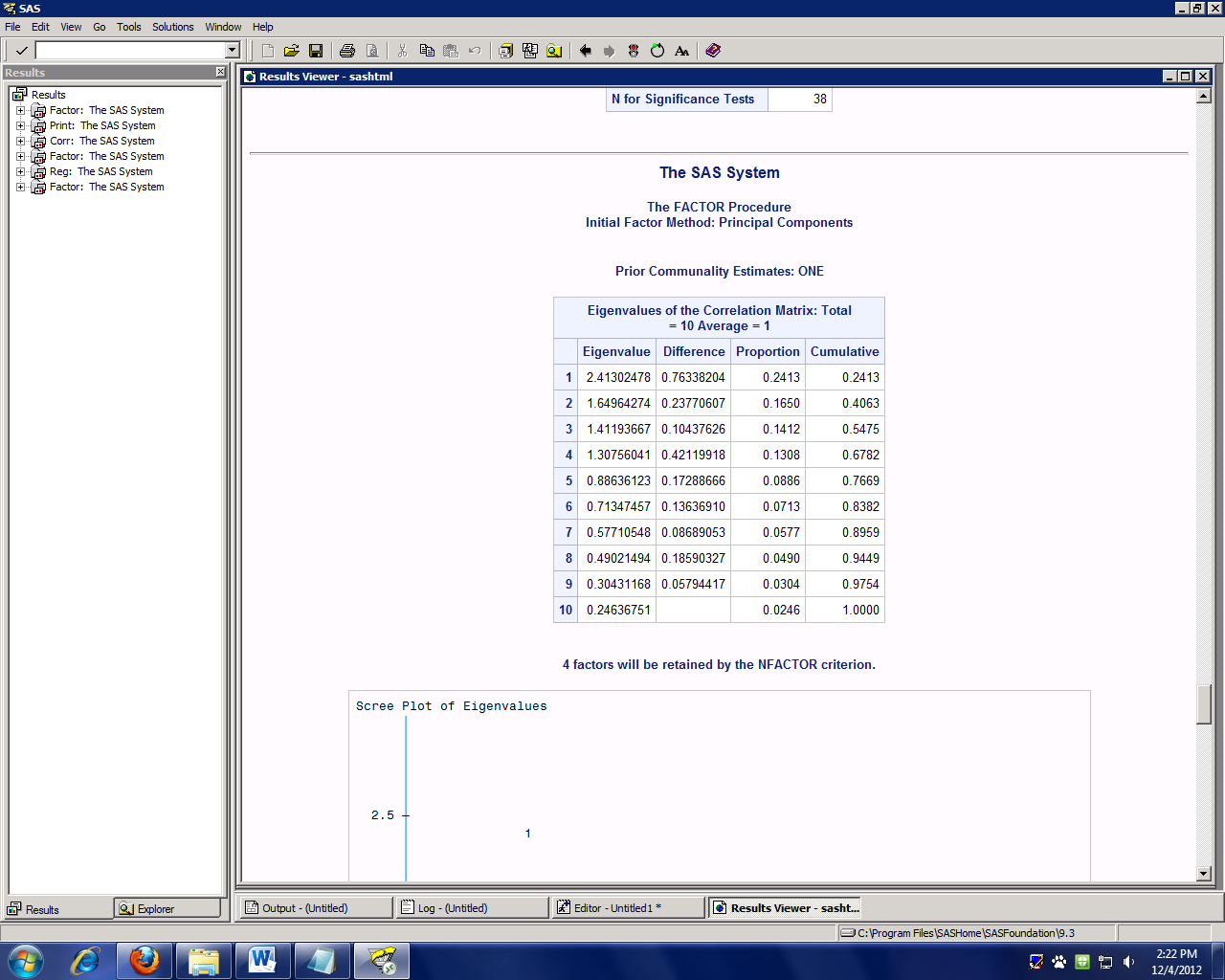
**PROC FACTOR** data = soil rotate=varimax scree nfactors=**4** out=soilfactors fuzz=**.4**;

var SO4 PO4 Zn Cd Cr Cu Pb Ni Mn Fe;

**RUN**;

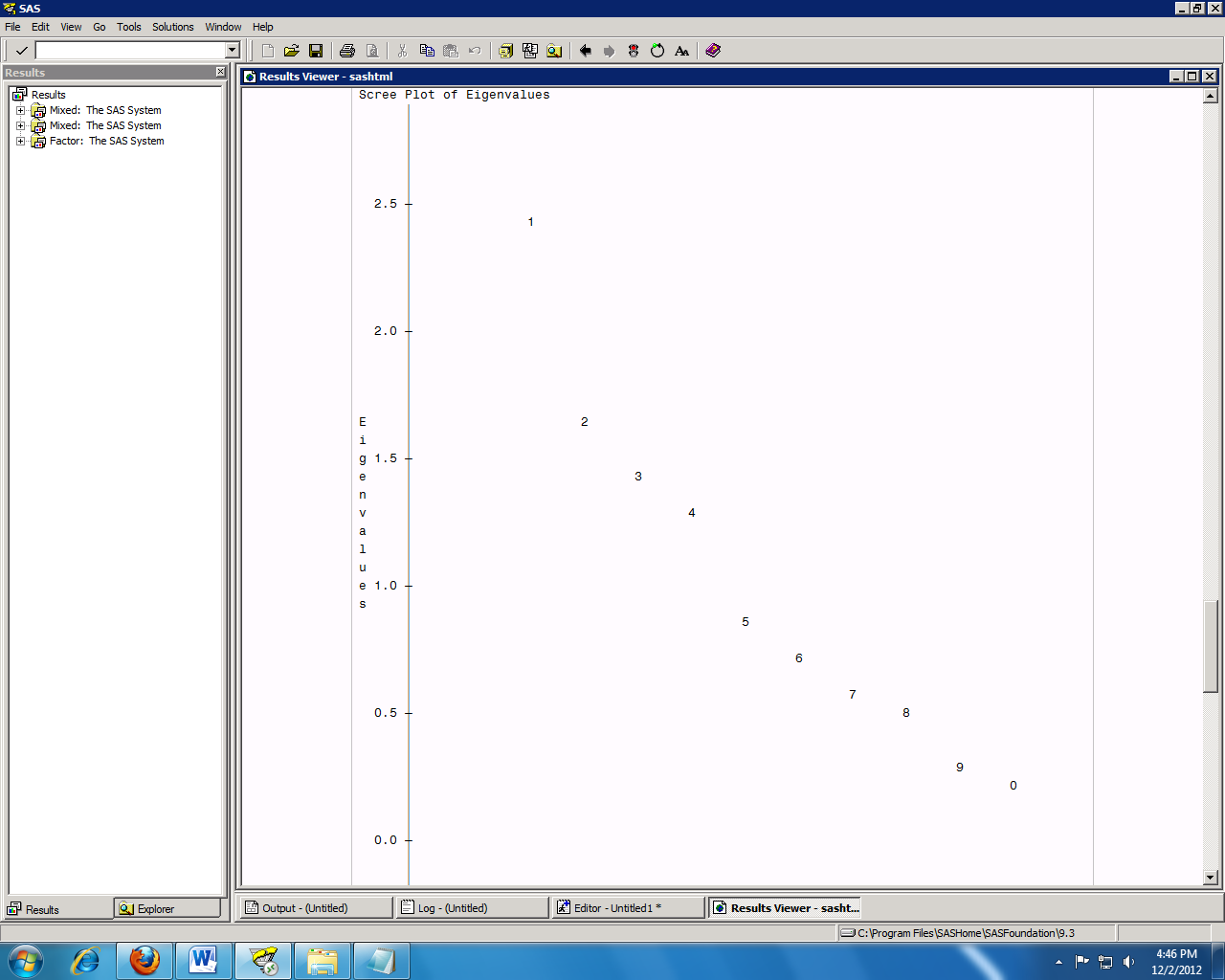
The purpose of PCA is to reduce the number of factors in the model. We want to remove variables that are highly correlated with each other because they are redundant. We will then be left with the most important factors, or the principal components.

Upon running the code above, we get the following output:

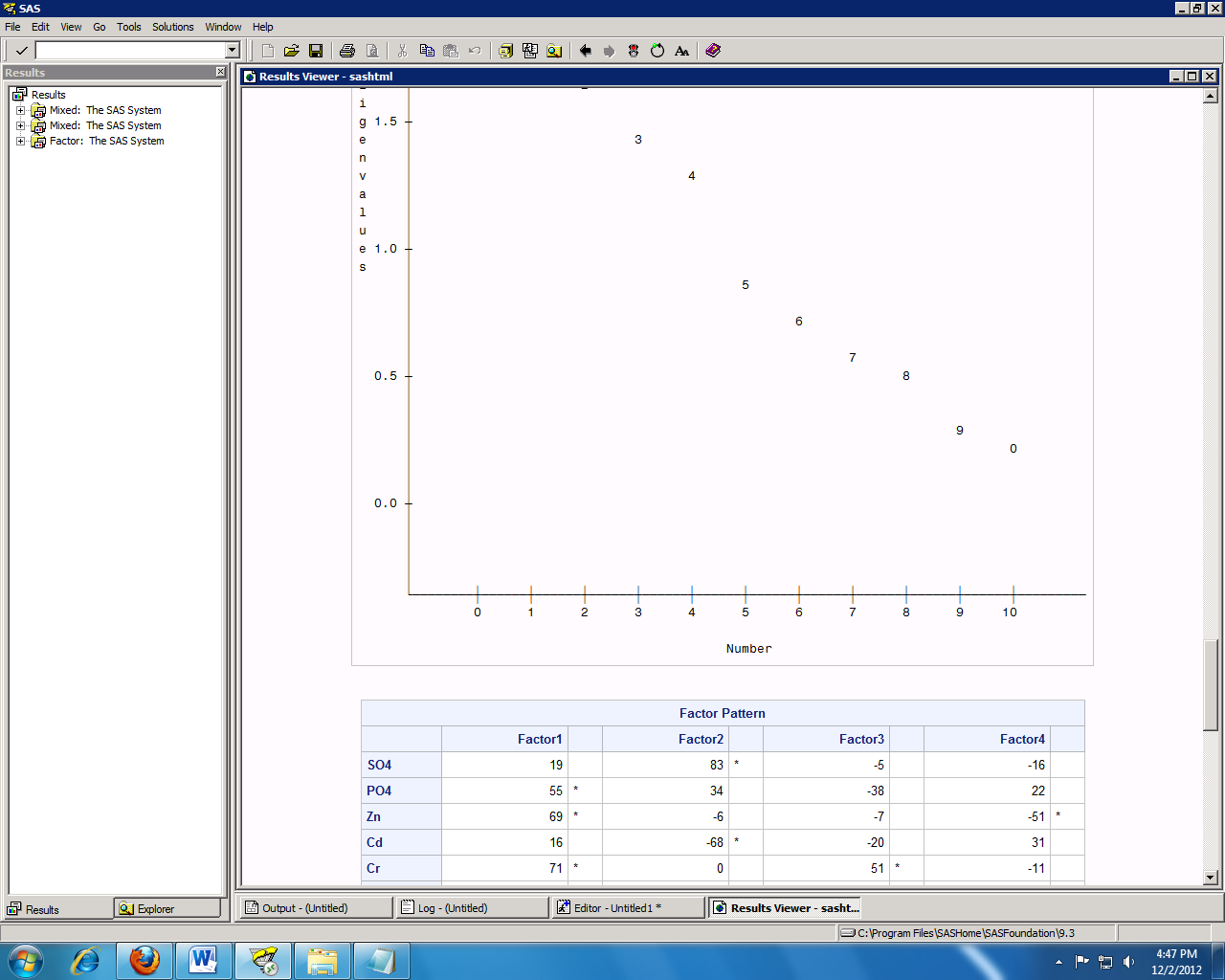


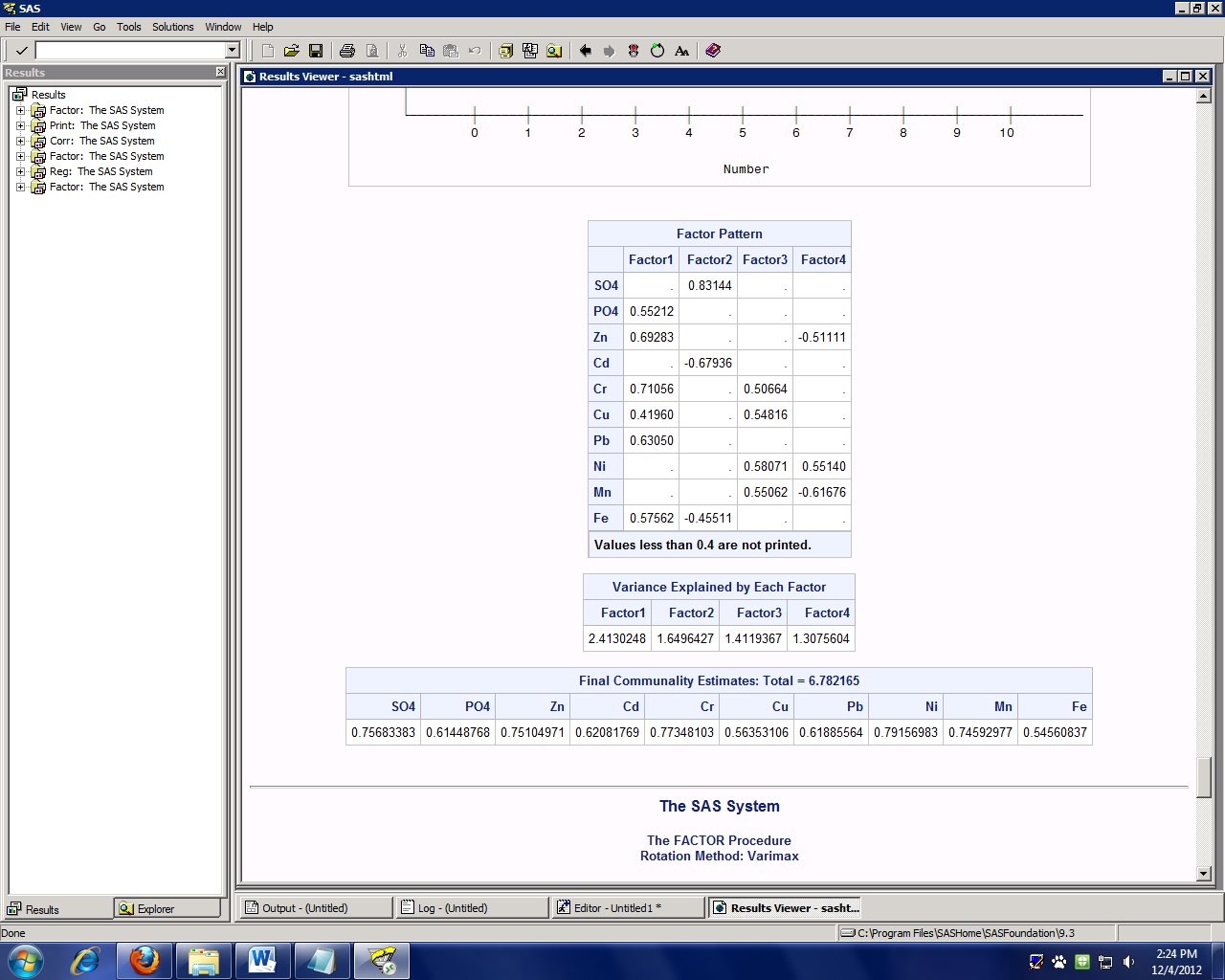
We have an accepted eigenvalue cutoff value of 1, so with that cutoff, we would say that we retain 4 factors. The cumulative amount of variability accounted for by these 4 factors is 67.8% as shown in the right most column.

The statement at the bottom of this piece of output agrees with what is stated above and says that 4 factors will be retained.

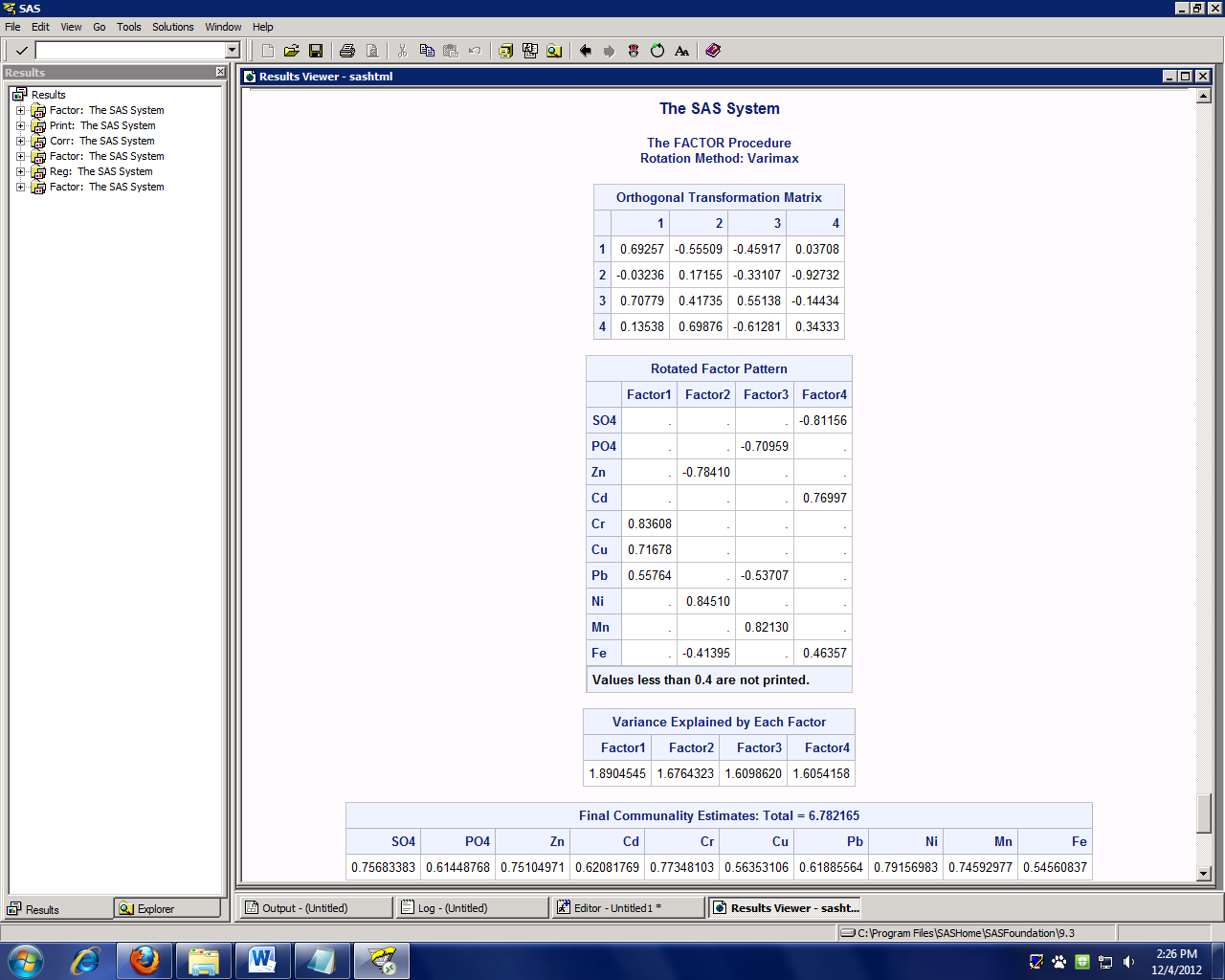


The SCREE plot to the left results in the same analysis as above. We see that the eigenvalues fall below 1 after the 4th factor. We also see that the plot starts to level off starting at factor 5.





Where the arrows are pointing we see that the variables are loading on more than one factor. We will apply the varimax rotation to help fix this issue.



After the varimax rotation, we see that there are only two variables loading on more than one factor, Pb and Fe. Neither one of these is loading very highly.

The rotated factor pattern shows how highly the variables are loading on each variable. For example, SO4 is loading highly negatively on factor 4, while Cd is loading highly positively on factor 4.

Now that we have determined that the proper number of factors that should be included in the model is four, we can perform regression including those four factors. The code to run the regression is:

**PROC REG** data = soilfactors;

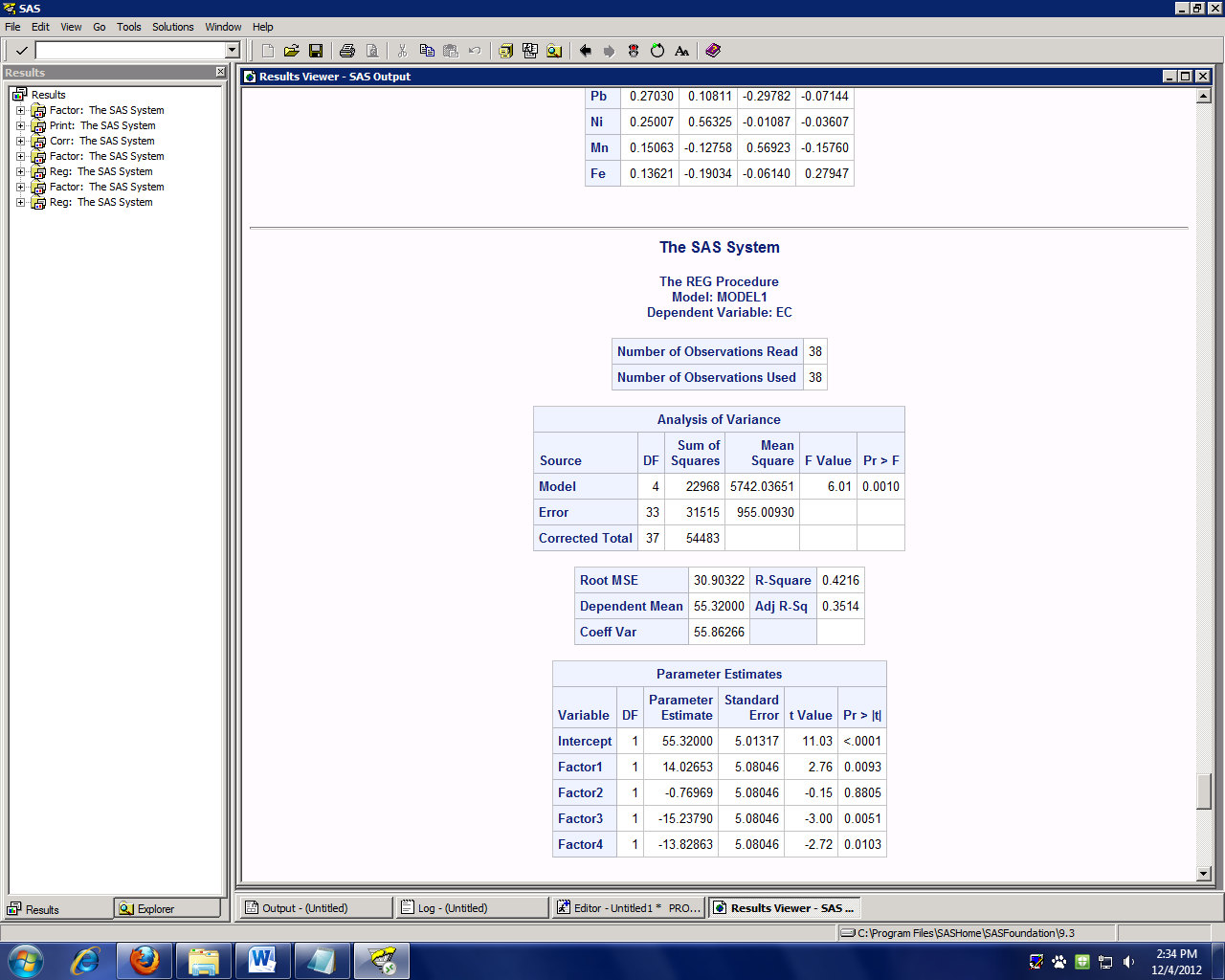
model ec=factor1 factor2 factor3 factor4;

**RUN**;

There should be three regression models, one for EC, one for pH, and one for salinity. For this example, we will just run the one for the response variable EC.

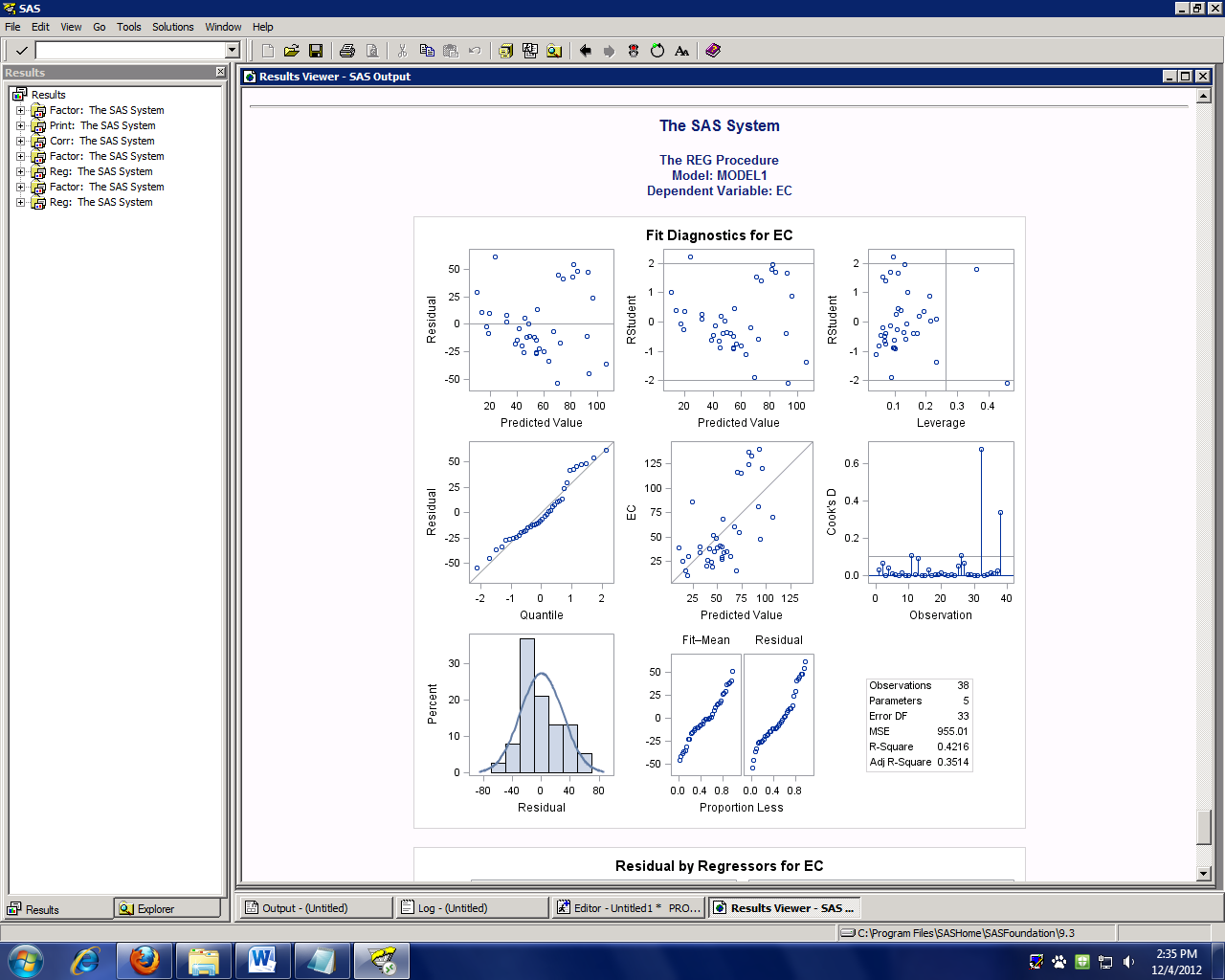
The data file used is called ‘soilfactors’. This file was created during the PROC FACTOR step. The model here is the response is equal to the four factors we obtained in the PROC FACTOR step.

After running the code we get the following output:



The overall p-value is 0.0010 which is less than the accepted of 0.05. This value tells us that the we have a significant model. This again, agrees with what we expected due to the EDA from the EC variable.

Three of the four factors are significant since their p-values fall below 0.05.



**8**

**7**

**6**

**5**

**4**

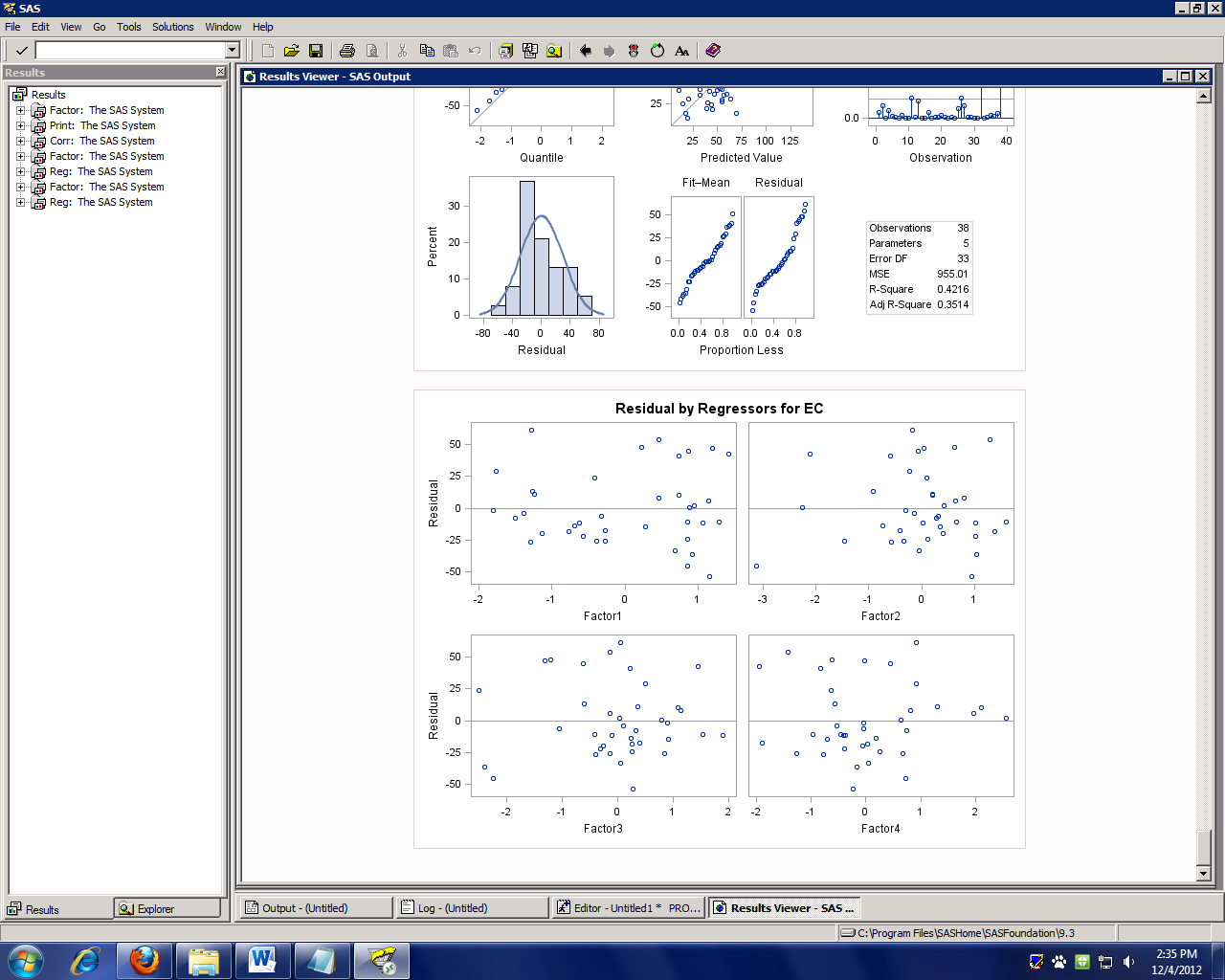
**1**

**3**

**2**

Using the above plots, we can check the assumptions for linear regression. All the assumptions must be met in order to use linear regression. The assumptions are linearity, independence, normality, and equal variances.

Firstly, using plots 4 and 7 we can check the linearity and normality assumptions. Since the points on plot number 4 follow the diagonal line pretty well, we will claim that the assumption for linearity is met. In plot 7 we see that the data follow a normal curve pretty well with the possibility of an outlier or two in the third bar (since this bar is much taller than the others). Even though the third bar is taller than expected, the data still follow a normal distribution so we will claim that the normality assumption is met. Plot 1 meets the criteria for both the independence and equal variances assumptions. There is no pattern to this plot and it looks randomly distributed. All the assumptions are met, so linear regression is appropriate in this case.



The residual plots above do not really show any specific patterns. Factor 2 might show a slight fanning pattern, but it does not show any obvious pattern.

***Summaries***

*Question 1*

The OG-T33 is a significant location in the ANOVA analysis of the EC response variable. We can claim this to be true because this location shows significant relationships with two different locations. Also, when the OG-T33 site is removed from the dataset, we see that the EC variable is no longer significant. So, the answer to the question of whether there are differences in pollutant levels between the different locations is yes. There are differences between the different locations.

*Question 2*

Because the overall p-value obtained from the linear regression following the PCA is significant, we conclude that we can predict overall pollutant levels based on the concentrations of different chemicals. So, the answer to the second question of whether concentrations of different chemicals can be used to predict overall pollutant levels is yes.