Fitting and Assessing Statistical Models for Well Water Samples and Health-Related Quality of Life Variables in Coal Mining Regions of West Virginia

Grant Carney '15, Dr. Brad Hartlaub

Introduction

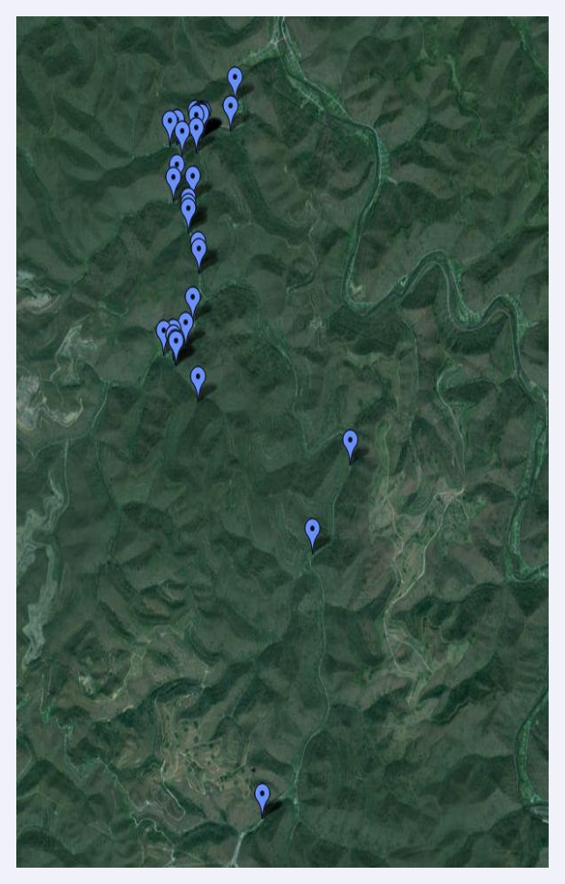
In recent years, certain areas in West Virginia have been under investigation due to claims of high contaminant levels in the local well water. Residents of one town claim that this pollution can be traced back to coal slurry, a by-product of coal mining.

This project uses two major databases in order to create a detailed statistical analysis of water quality levels in West Virginia. The first dataset, provided by Jackson Kelly PLLC, includes on-site samples from nine laboratories. Almost 10,000 samples were obtained by experts and independent consultants, with approximately 125 variables. These variables include hazardous elements like arsenic, water quality measures like pH levels, and spatial identifiers such as GPS coordinates of the wells. The second dataset is from the West Virginia Department of Environmental Protection (WVDEP). The WVDEP data contains river and stream contaminant samples from bodies of water that are in close proximity to the wells. This dataset was mainly used to compare well pollutant levels to contaminants in other nearby bodies of water. Due to the nature of on-site sampling, many of the pollutants are classified as below the detection limit and are given a special code in the file.

The goal of this study is to examine pollutant levels over time using existing EPA guidelines. The EPA suggests nonparametric procedures instead of parametric procedures when the distributions of the pollutants are not normal. The importance of using nonparametric procedures is examined throughout the study. Linear, bootstrapped, and best subsets models are created in order to identify the best possible predictors for a given pollutant. These models are useful for predicting the level of a particular pollutant. Another goal of this project is to perform a spatial analysis using grouping techniques. Nonparametric multiple comparisons outlined in Hollander and Wolfe's "Nonparametric Statistical Methods" are used throughout the analysis.

Map of the Region

Using Google Maps and GPS coordinates that were given in the dataset, a map was created of the region in question. In total, 72 wells were plotted in order to make spatial comparisons simpler. The wells are all within five miles of one another, so pollutant concentrations should be fairly similar. However, there may be differences in pollutants due to any of the following reasons: Coal slurry runoff, secondary source pollution, improper maintenance of the well, or variation in environmental conditions.



National data on health issues is available from The Behavioral Risk Factor Surveillance System that was created by The Center for Disease Control, but county level identifiers are not available for the 2012 database due to confidentiality restrictions. Therefore, the main focus of the project shifted to modeling.

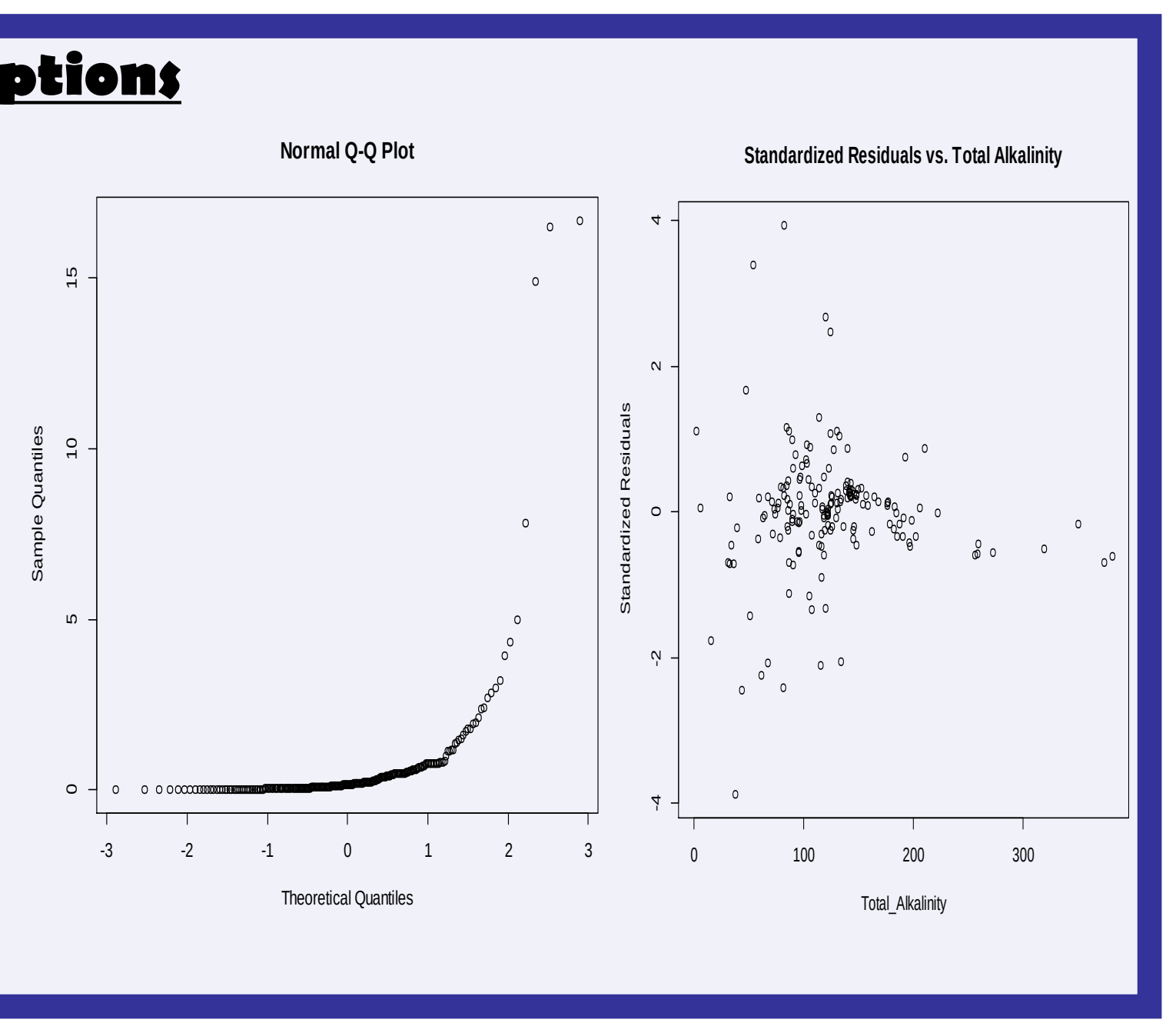
Formal Model

The general linear model that is used throughout this project is shown below. Each βk is a coefficient and each Xk is a predictor such as a pollutant, latitude, or elevation.

$$Pollutant = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

Examining Classical Assumptions

Many parametric procedures inherently contain assumptions that the data have a constant variance and are normal. The graphs to the right show the well dataset rarely satisfies these assumptions. If the data were normal, we would expect to see a linear normal probability plot and an unstructured band of points in the residual plot. The Q-Q shows clear curvature, and the residual plot shows heteroskedasticity. We will rely on distribution free models when these assumptions are not satisfied.

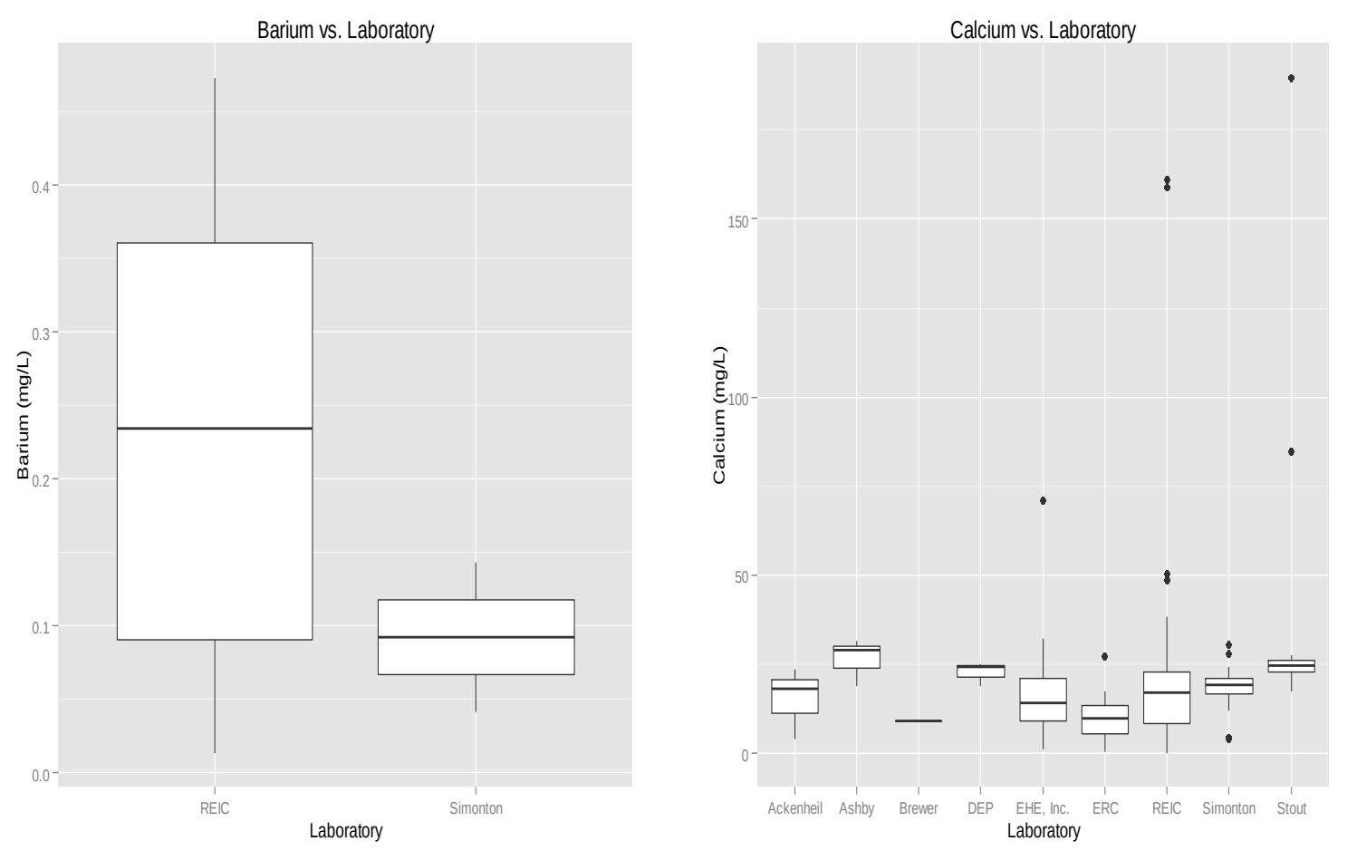


Analysis of Laboratories

Calcium								
mean	sd	median	min	max	range	Q1	Q3	IQR
17.54	18.25	16.9	.07	189.81	189.74	8.86	22.8	13.94

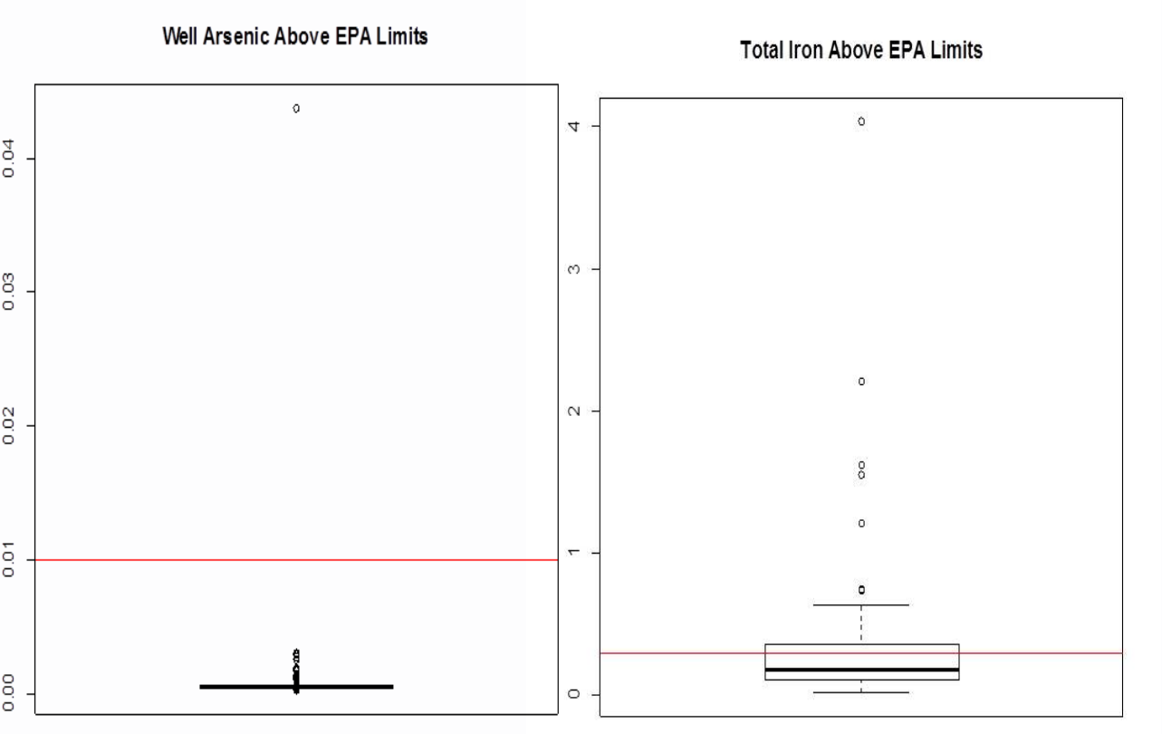
Barium								
mean	sd	median	min	max	range	Q1	Q3	IQR
.21	.14	.23	.01	.47	.46	.0708	.3595	.2887

Multiple laboratories sampled water sources across West Virginia. In order to examine the variation across labs, descriptive statistics, boxplots, and Tukey multiple comparisons were used. The "Barium vs. Laboratory" boxplot to the right is unusual because in general there were not great differences between the measurements of the labs. Therefore, we do not need to control for the laboratory that took the sample when performing the analysis.



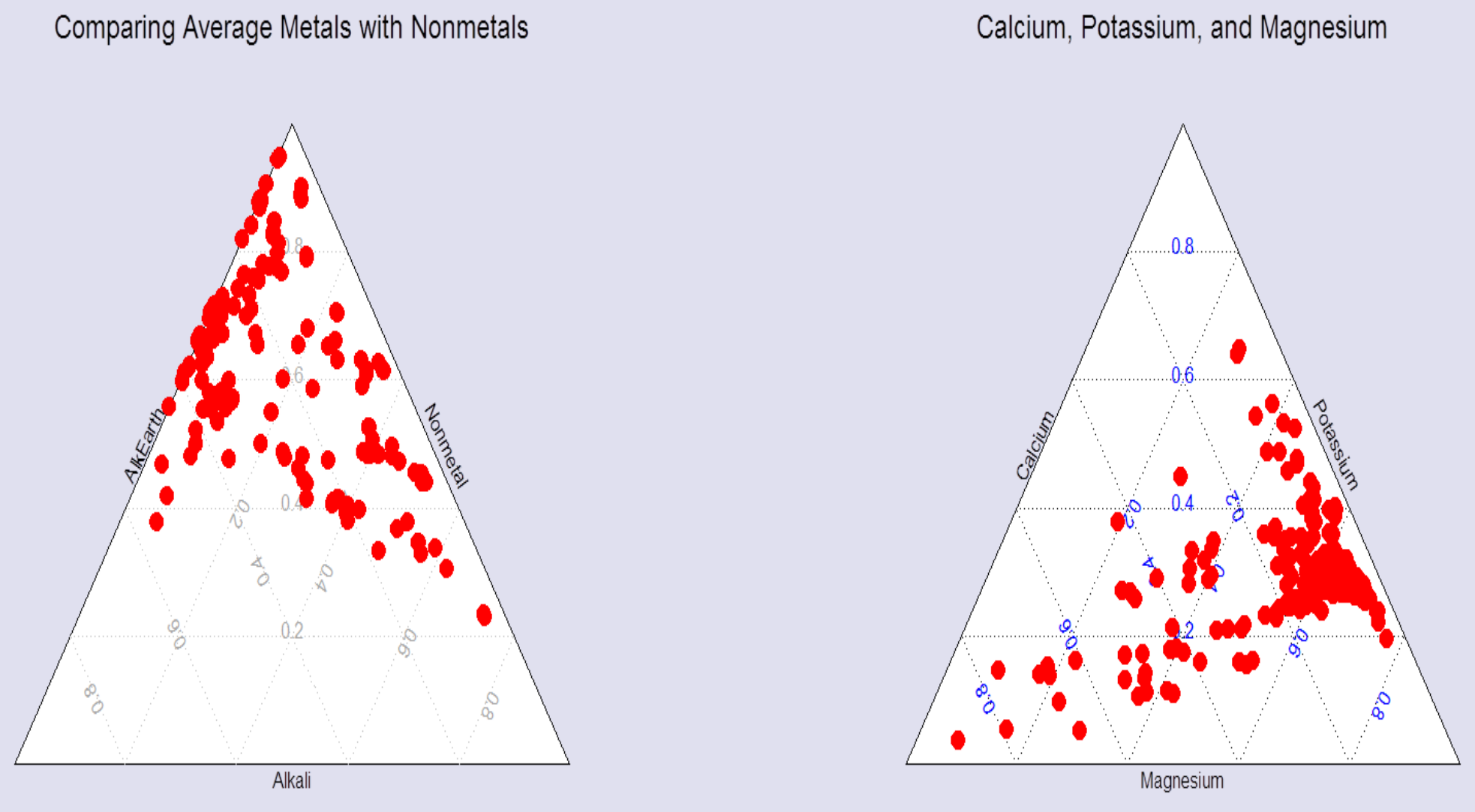
EPA Secondary Concentration Limits

A 2002 EPA report, titled "Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites", details the enforceable standards of "maximum contaminant levels" in drinking water. Boxplots were created in R in order to determine how often the well data contained pollutant concentrations above the recommended EPA guidelines. The red line is the EPA's MCL for the given element. The plots below show that some elements such as iron have serious contamination issues, while other elements such as arsenic have almost no indication of high concentrations.



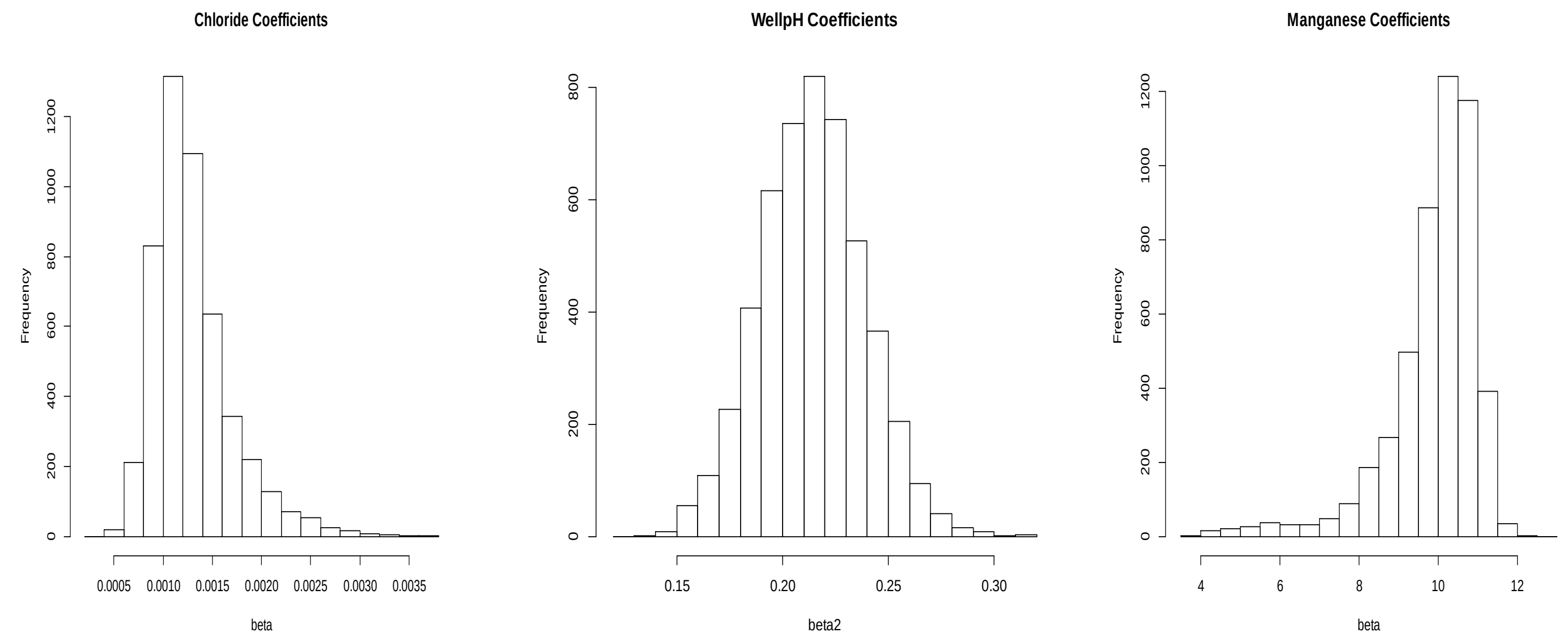
Ternary Diagrams

Ternary diagrams, also known as triangle plots, are often used in the physical sciences in order to examine the ratios of three variables to one another. Knowing more about the composition of the pollutants can help us to determine the complexity of the relationship. For example, the plot "Comparing Average Metals with Nonmetals" shows that almost every one of the wells contains a greater proportion (mg/L) of alkali metals to alkali earth and nonmetals. All except for a few wells have more than 40% alkali metals and around 0% alkali earth metals. The second plot shows the higher concentration of calcium compared to potassium.



Bootstrapping

Because the necessary conditions for a linear model do not apply for the well data, an alternate procedure called bootstrapping is used to make inferences for the regression models. 5000 bootstrapped simulations are created in order to model every element of interest. The bootstrap simulation is created in R. Interaction was considered in the bootstrap models when applicable. The right skew of the bootstrapped distribution for the Chloride coefficients histogram shows that outliers are prevalent in the original data set. The distribution for Well pH is approximately normal, and the distribution for Manganese is skewed left. These histograms show that some pollutants will require nonparametric procedures while some may not.



In order to check the utility of these bootstrapped models against the classical linear model, the mean absolute deviation (MAD) was used as a metric to compare the error variability. The bootstrapped models often outperformed the classical models.

$$MAD = \sum_{i=1}^{n} \left| \frac{(\widehat{Y}_i - Y)}{n} \right|$$

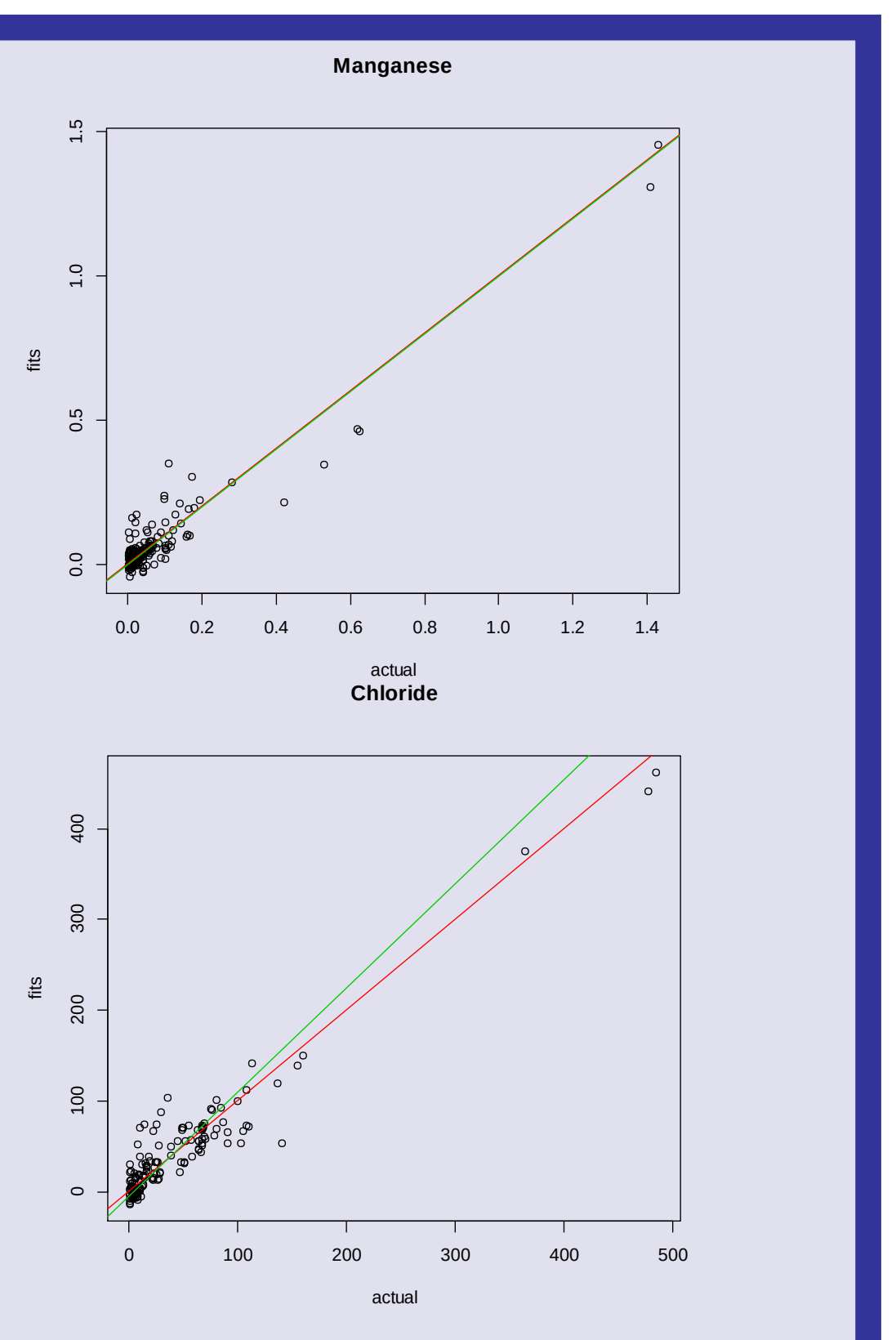
MAD		
Alkalinity	Bootstrapped	Classical
	24.96287	25.01254
Calcium	Bootstrapped	Classical
	0.09660572	0.0968556

Best Subsets

R can create another set of models to be compared to the bootstrap models. Both stepwise and backward elimination techniques are considered. To determine the best model, Mallow's CP is used.

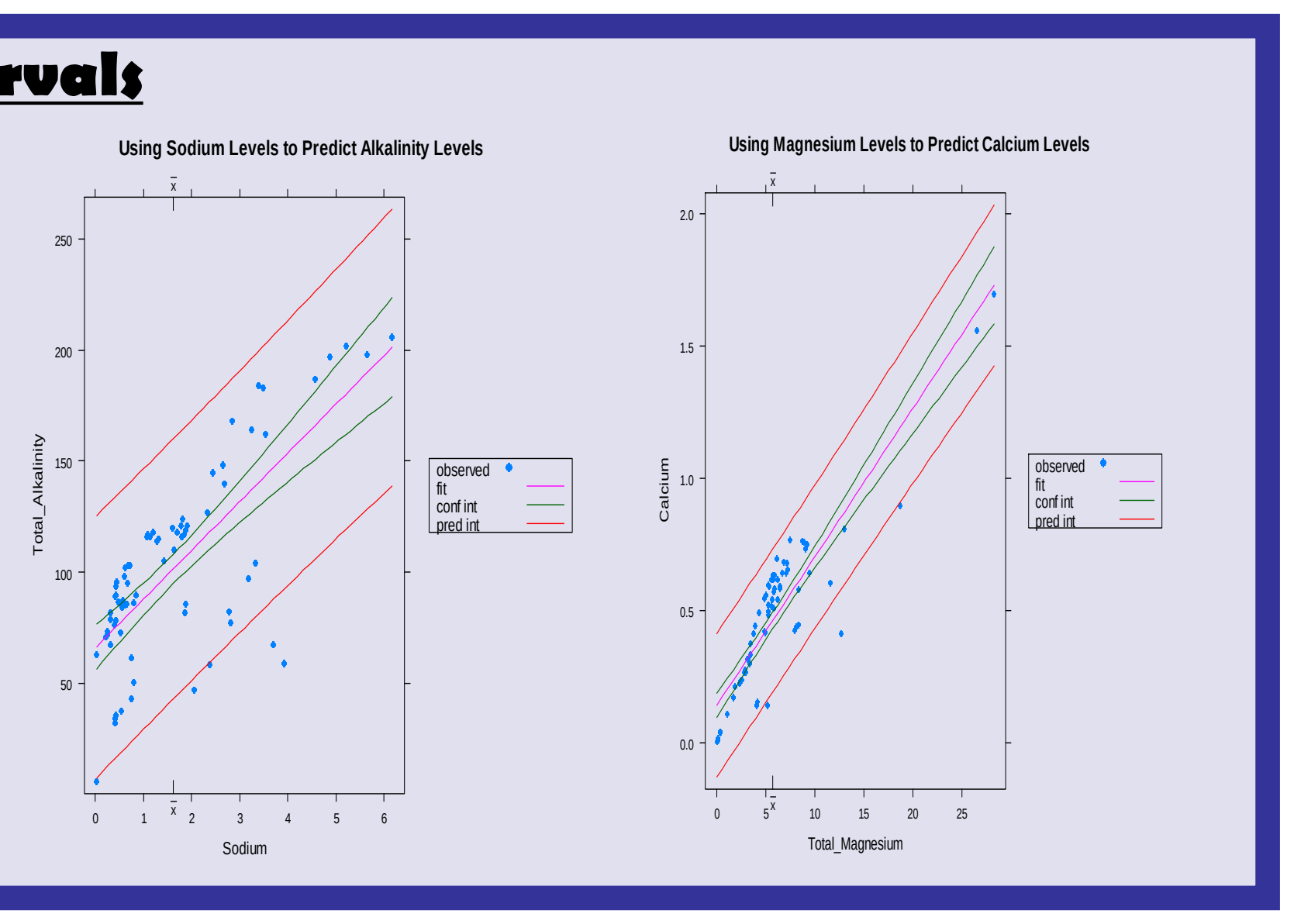
$$C_p = \left(\frac{SSE_m}{MSE_k}\right) + 2(m+1) - n$$

Plots of the fitted values against the actual values can help to determine the quality of the best subsets models. Both graphs contain two lines, one with the line of best fit with the outliers, and one without the outliers. The graph for Manganese shows that the outliers have relatively little influence on the regression while the graph for Chloride says the opposite. The green line is the fitted line when the outliers are not included. It is obvious that the three points in the top right of the plot have a large effect on the line of best fit. Due to the influence of the outliers, an alternative least squares regression approach would be preferred.



Confidence and Prediction Intervals

For Alkalinity and Calcium, 95% confidence and prediction intervals are created for simple linear regression models. The first regression successfully uses Sodium to predict Alkalinity levels. The influence of outliers on the slope is obvious in the graph using Magnesium to predict Calcium levels. With the outliers, the confidence interval for the slope is (.0499 mg/L, .0625 mg/L). Without the outliers, the interval for the slope increases to (.0744 mg/L, .0842 mg/L).



Sample Script

The code to the right is an example of a typical script that is created in R. This particular script runs a for loop for the bootstrapped simulation for the coefficients predicting Calcium. The final command "bootbetas" saves the coefficients for every predictor in the model.

```
for (i in 1:nboot) {
        bootmag=sample(magnesium,replace=TRUE)
        bootmodel=lm(Calcium~Total_Magnesium+Sodium,data=mydata3[bootmag,])
        bootbetas[i,]=coef(bootmodel)
}
```

Transformations, Interaction Terms, and Final Models

Interaction terms and transformations are sometimes needed to explain the relationship between a pollutant and its set of predictors. Residual plots were examined, and logarithmic, exponential, or quadratic transformations were performed. The graph in the "Examining Classical Assumptions" section is a great example of a residual plot that would require a transformation. Here are some of the final models:

Response	Predictors	t-statistic	p-value	Adj. R-Sq
Total Manganese	Magnesium $\sqrt{Total\ Alkalinity}$ Total Iron Potassium	8.229 3.996 31.987 -1.964	p<.0001 p<.0001 p<.0001 p<.1	.8484
Potassium	Sodium (Dissolved Calcium)^2 Interaction (Na*(DC)^2)	4.739 2.79 1.781	p<.0001 p<.01 p<.1	.758
Sodium	Total Alkalinity Chloride	21.12 26.96	p<.0001 p<.0001	.9399
Sulfate	Total Selenium Latitude Interaction (Lat*Se)	7.346 4.626 -7.343	p<.0001 p<.0001 p<.0001	.8481

$$Total\ \widehat{Manganese} = -0.087392 + 0.011821 * Total\ Magnesium + .008207 * \sqrt{Alkalinity} + .081467 * Total\ Iron - .620993 * Potassium$$

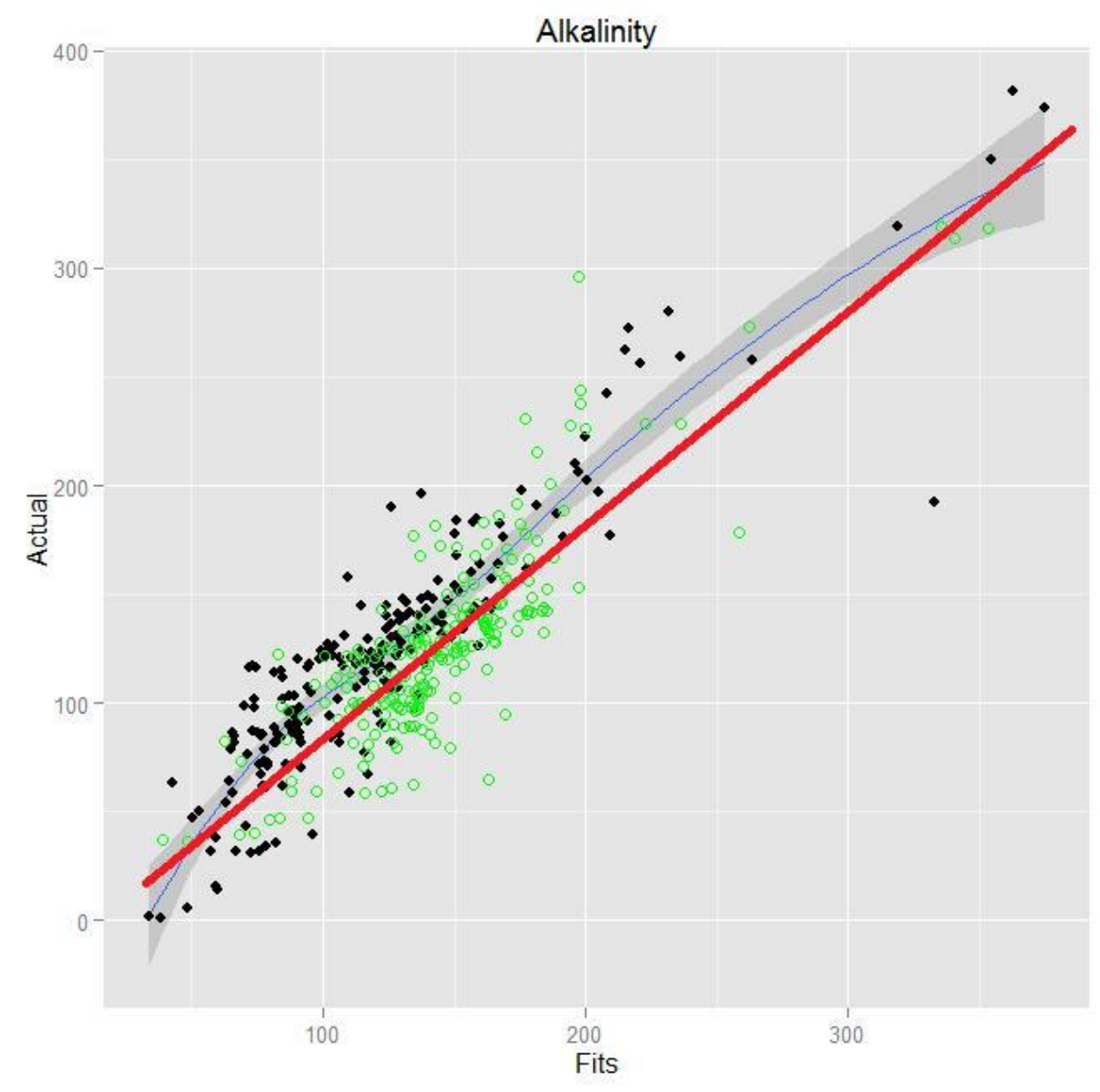
$$\widehat{Potassium} = .03549 + .001803 * Sodium + .0007239 * (Dissolved\ Calcium)^2 + I(Sodium: (Dissolved\ Calcium)^2)$$

$$\widehat{Sodium} = -1.280126 + .023656 * Total\ Alkalinity + .028736 * Chloride$$

$$\widehat{Sulfate} = -94480 + 218400000 * Selenium + 2481 * Latitude - 7736000 * I(Selenium: Latitude)$$

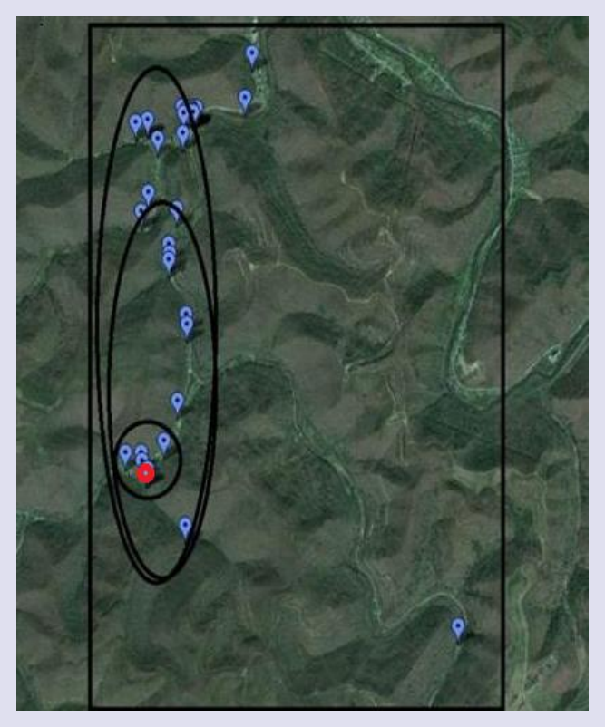
Comparing the Models

In order to compare the best subsets, bootstrapped and simple OLS models, the fitted values for each model were plotted against the actual values on the same graph. A good model would contain points that closely follow the 45° line. The black points, plotted according to the best subsets model for alkalinity, often are above the line of best fit, meaning that the model often underestimates the actual values of alkalinity. The blue lowess trend line confirms that the actual values are almost always greater than the fitted values. The green bootstrapped points provide a much better model as almost half of the points are above the line of best fit, while the other half are below. For the variable alkalinity, the nonparametric model outperformed the classical regression procedures.



Hotspots

Using a distance metric, a hotpot was identified and subsequent groupings of wells were created. The hotpot, which was known to contain pollutants, was compared to the other groups through Kruskal-Wallis and ANOVA procedures. The first equation is the Kruskal-Wallis test and the second is for multiple comparisons.



$$K = (N-1) * \frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n}(r_{ij} - \bar{r})^2}$$

$$W*_{ij} = \frac{W_{ij} - \frac{nj(ni+nj+1)}{2}}{\left(\frac{ninj(ni+nj+1)}{24}\right)\left(\frac{1}{2}\right)}; Decide\ \tau_u \neq \tau_v if \ \geq q_\alpha$$

The Kruskal-Wallis test found general differences, so Dwass multiple comparisons were run to find out which hotspot groupings differed. Unlike the parametric multiple comparisons, the nonparametric test statistic of W_{uv}=2.906 was allowed us to conclude that there are differences between the Hotspot and Well Group 3.

Year-to-Year Analysis

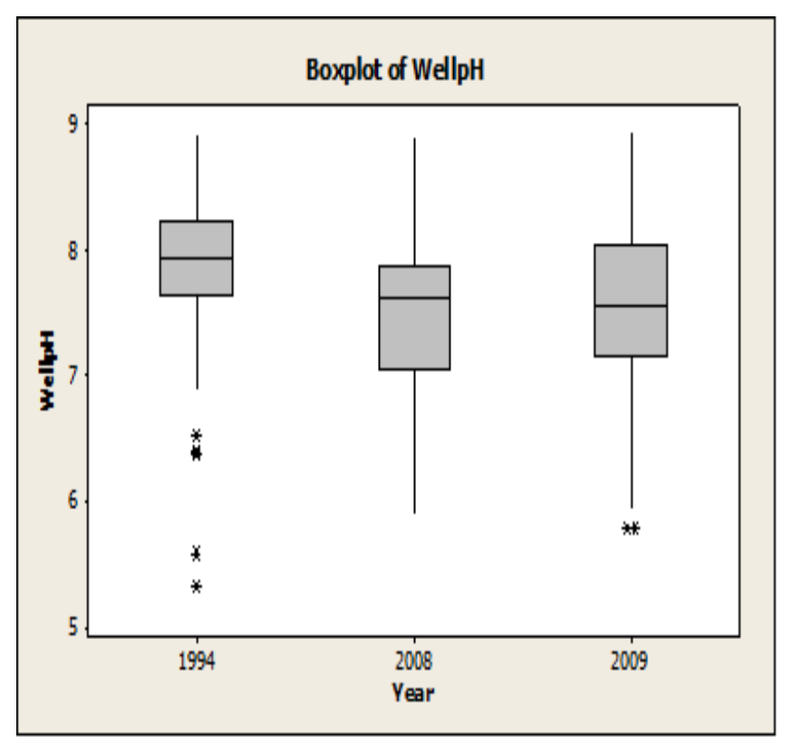
In order to check for changes in pollutants over time, a Jonckheere-Terpstra ordered alternatives test is used, with τ equal to the yearly treatment effect. The ordered alternatives test for pH provided a test statistic of J*= 2.89 and a p-value less than .01. There is evidence that pH is decreasing over time. However, this is not a negative result because the EPA's secondary concentration guidelines for pH range from 6.5 to 8.5.

$$H_0: \tau_{1994} = \tau_{2008} = \tau_{2009}$$

$$H_a: \tau_{1994} \geq \tau_{2008} \geq \tau_{2009}$$

$$J = \sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}$$

$$J^* = \frac{J - E_0(J)}{\sqrt{var_o(J)}}$$



Results

This Summer Science Research project supports evidence that nonparametric statistics often outperform the competing parametric procedures when outliers are present. Analysts working with hazardous materials must follow recent EPA guidelines that emphasize the use of distribution free tests. The analysis in this project supports the claim that certain areas in West Virginia must continue to sample local well water in order to find the root source of pollution. To eliminate health concerns, city water is now available for all residents in the area!

Acknowledgements

I would like to thank Brad Hartlaub for his assistance and guidance during this project. Recognitions are also due to Kenyon College and The Kenyon Summer Scholars Program for providing funding. I would also like to thank Rob Passmore for allowing us to use the Jackson & Kelly dataset. I express my gratitude to John Wirts of the WVDEP Division of Water and Waste Management for his willingness to provide water quality data on area streams and rivers. Lastly, I would like to thank Zubair Saleem of the EPA for his discussions on water pollutant simulation.

References

-Cannon, Cobb, Hartlaub, Legler, Lock, Moore, Rossman, and Witmer. STAT2- Building Models for a World of Data. W.H. Freeman, 2012.
-Hollander, Myles, and Douglas Wolfe. Nonparametric Statistical Methods. 2nd ed. New York: John Wiley & Sons, Inc. , 1999.
-Jones, Owen, Robert Maillardet, and Andrew Robinson. Scientific Programming and Simulation Using R. 1st ed. . Boca Raton: Taylor & Francis Group, LLC, 2009.
-R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://cran.us.r-project.org/.
-United States EPA. Office of Emergency and Remedial Response. "Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites." Washington, D.C. : , 2002.
- Zullig, Keith, and Michael Hendryx. "Health-Related Quality of Life Among Central Appalachian Residents in Mountaintop Mining Counties." American Journal of Public Health. 101.5 (2011). 9 Oct. 2013.