An Exploration of Non-Detects in Environmental Data

A Senior Project

presented to

the Faculty of the Statistics Department

California Polytechnic State University, San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science

by

Juliana Fajardo

June, 2011

Table of Contents

Introduction

The goal of my project is to explore how non-detects may affect the results of standard analyses and to discover methods to address this possible affect. To look at this problem of non-detects, I will be analyzing a specific case study.

To explore the affect of non-detects, I will first consider a "standard analysis" of the data for which the non-detects are treated as actual data values at half the detection limit. Second, I will consider the same analysis for which the non-detects are treated as values of zero. Finally, I again will consider the same analysis for which the non-detects are treated as the detection limit values. Ultimately, I want to know if the updated treatment plant has made a difference on the stream water quality, and through comparisons, we want to see if the bias of non-detects affect these results.

To address the bias of the non-detects, I will be considering a simulation method. I will simulate values for the non-detects based on a uniform distribution and then analyze the simulated data sets repeatedly and average the results. With the simulated results, comparisons will be made to the results of previous substitution treatments.

The Problem of Non-Detects

When analyzing environmental data, non-detects can pose a problem for data analyses. In many environmental problems, non-detects occur when chemical concentrations are below the detection limit of the measuring devices. Also called "less thans" or "censored data", these low values, in our study described below for nitrates and phosphates, are known inexactly (Helsel, 2005). The non-detect values can affect the look of the data graphically as well as affect models, summaries, and test results.

There are different ways to deal with non-detects: substitution, probability, and simulation. For example, a few forms of the substitution method could be to use the values of the threshold limit for the non-detects, replace all non-detects with zero, or replace all non-detects with half of the threshold value. Ultimately, using substitution may be the simplest procedure, but it is the most inaccurate route to take when dealing with non-detects (Helsel, 2005). Better alternatives to deal with the issue of non-detects include making a probability model to find non-detect values or using simulation to generate values for non-detect substitution.

Chorro Creek/Men's Colony Case-Study

Background Information:

One of the most important resources to any society can be summed up into one word: water. Water is important not only for individual, household, and agricultural use but for environmental stability, for example through local streams. In San Luis Obispo Country, these streams and creeks are important for many reasons. With environmental issues present in our community, it is important to know what things, such as our streams, are providing for our community. In general, creeks maintain a "cleansed" environment by carrying impurities away instead of keeping water stagnant in a single area. Creeks maintain a stable surrounding environment, such as providing a living space for organisms. Stream water considered to be unpolluted have less than 1 mg/L of nitrates where the Department of Environmental Protection (DEP) water quality standard is 10 mg/L (Lehigh Earth Observatory, 2006). Phosphate levels below 0.03 mg/L are considered to be unpolluted and the critical level for avoiding accelerated eutrophication is 0.1 mg/L (Lehigh Earth Observatory, 2006).

Nitrates can be found in sewage water and phosphates are usually found in water containing laundry detergent. Phosphorous is considered a "limiting nutrient", which is a nutrient that limits plant growth, which can affect the ecosystems of the stream system. Also, both phosphorus and nitrogen causes eutrophication, especially phosphates found in laundry detergents. Eutrophication is a rapid increase in algae or plant growth in an aquatic system due to the influx of a limiting nutrient that was in short supply previously (McKinney, 2007). Essentially, "too much of a good thing can be a pollutant" (McKinney, 2007).

Case Study Information:

I will be analyzing nitrate and phosphate data collected from the San Luis Obispo Country Creeks over a period of 10 years, from January 2000 to May 2010. Within this time span, a new treatment facility for the California Men's Colony waste water went on-line Sept. 2007. In addition, the data was collected both upstream of the treatment plant and downstream of the treatment plant, with 3 collection sites upstream, R-01, CHO, and R-02, and 6 collection sites downstream, R-03, UCR, CER, CAN, UCF, and TWB (See Appendix Site Details). In our data set, there are some missing data either based on different dates that samples were collected or data that wasn't collected until later years at some of the sites.

To explore the effect of non-detects on environmental data, we look at the problem posed by this Chorro creek data. Because the county's creeks run through the community, our first observation was the effect of the Men's Colony on the creeks. The creek runs through the Men's Colony in San Luis Obispo and nitrates and phosphates are introduced from sewage and laundry water. Because both nitrates and phosphates have regularly leaked into the creeks, the colony updated their private waste water treatment plant. We want to find out if the update of the treatment plant had any effect on the amount of trace nitrates and phosphates in the creek. For this specific case study, traces of nitrates and phosphates found in the creek downstream of the Men's Colony are generally higher than upstream of the colony.

Data Collection:

When the data was collected at each site, it was taken by a collection of different labs. Each lab had their own testing device to read levels of nitrates and phosphates. In fact, these devices for each lab had different threshold limits of chemical detection. Because there are different measurement thresholds, different non-detect values are present in the data, such as .1, .4, .25 etc., which further complicates analysis (Helsel, 2005). The method in which these labs recorded censored values was to record one half of the detection limit. Like in this case, substitution is one standard approach when working with this sort of censored data.

Case Study Objective:

The objective of the study is to compare the changes, if any, in the amount of nitrates and phosphates in the water. By comparing the samples from upstream and from downstream of the plant and comparing the samples over time from before and after the plant instillation, we can determine whether or not significant changes in pollutant levels have occurred, while taking into consideration the non-detects.

Brief Introduction to the Analysis:

To execute the study, I will first provide traditional analysis to show initial observations of the data. This will be done when non-detects are ignored and the data is just used as is and what the results show about the raw data. To analyze the original data, I will use a 2 sample t test, ANOVA, and regression to analyze data before plant and after plant up date and to compare the difference between nitrate and phosphate traces both upstream and downstream from the plant.

After the initial testing, I will change the non-detect values. To view substitutions at both ends of the spectrum, the non-detect substitution values will be changed to zero and the values of the threshold limit. After the non-detects are changed, we will fit the same models and compare the change between each substituted non-detect value. Because we will be dealing with non-detects, a questions we can ask is: when the non-detect values are changed, how is the data set affected?

The final goal will be to find a better solution by developing a simulation for the non-detects. I will use simulation to find the distribution of the results for the non-detect values and change the values that were originally assigned substitutions. Because the non-detect values will not be simple substitutions, we would like to see what this does to the non-detects data that substitution does not.

Analysis

The original data for nitrates and phosphates were not normal, possibly from the impact of non-detect substitution values. Histograms of the raw nitrate and phosphate data demonstrate a strong right skew, in Figure 1 and 2, thus the analyses considers log transformed data. When using the log transformation when the non-detect substitutions were changed to zero, log(0) does not have a value. To adjust for this problem, the log values are calculated by log(value+1) for all of the data.

Figure 1: Histograms of the Raw Nitrate Data (Normal)

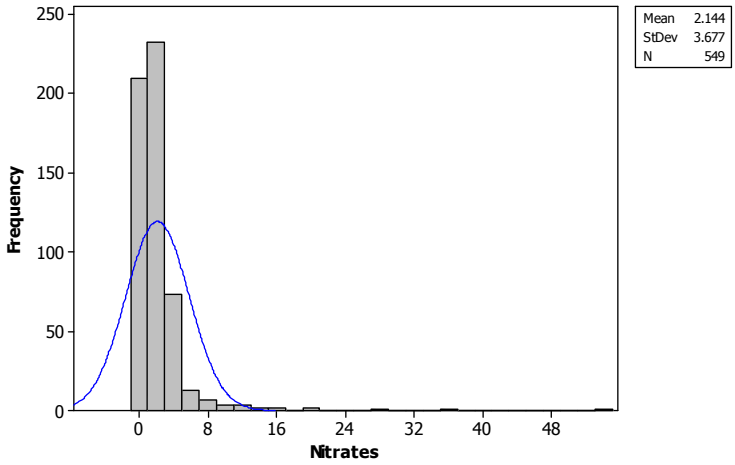
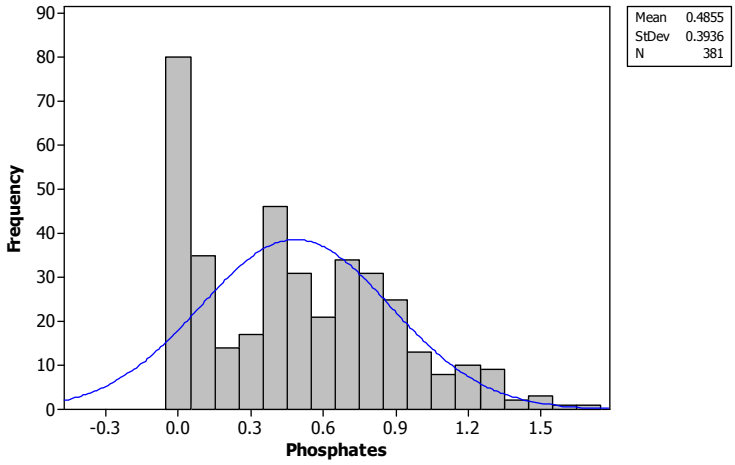


Figure 2: Histograms of the Raw Phosphate Data (Normal)



For reference, Table 1 and Table 2 below both display basic summary statistics of the logged data as well as medians of the raw data for each location and the percent of data values below the threshold.

Table 1: Summary Statistics for Log Nitrate Data

		Raw Data Median	Mean	SD	SE	N	Percent Below Threshold
R-01	**Before**	0.050	0.050	0.038	0.005	53	86.79
	After	0.055	0.032	0.031	0.006	30	86.67
CHO	**Before**	0.050	0.047	0.045	0.016	8	62.50
	After	0.125	0.066	0.055	0.010	30	46.67
R-02	**Before**	0.500	0.219	0.133	0.018	53	1.88
	After	0.480	0.213	0.102	0.018	31	3.22
R-03	**Before**	4.050	0.760	0.323	0.045	52	0
	After	2.600	0.515	0.152	0.027	31	3.22
UCR	**Before**	3.200	0.659	0.236	0.032	53	0
	After	1.700	0.429	0.101	0.012	74	1.35
CER	**Before**	2.550	0.578	0.137	0.056	6	0
	After	1.700	0.406	0.098	0.035	8	0
CAN	**Before**	2.900	0.591	0.142	0.039	13	0
	After	1.200	0.344	0.093	0.015	38	2.63
UCF	**Before**	NA	NA	NA	NA	0	0
	After	1.700	0.425	0.065	0.023	8	0
TWB	**Before**	3.350	0.612	0.142	0.029	24	0
	After	2.200	0.508	0.073	0.011	45	0

Table 2: Summary Statistics for Log Phosphate Data

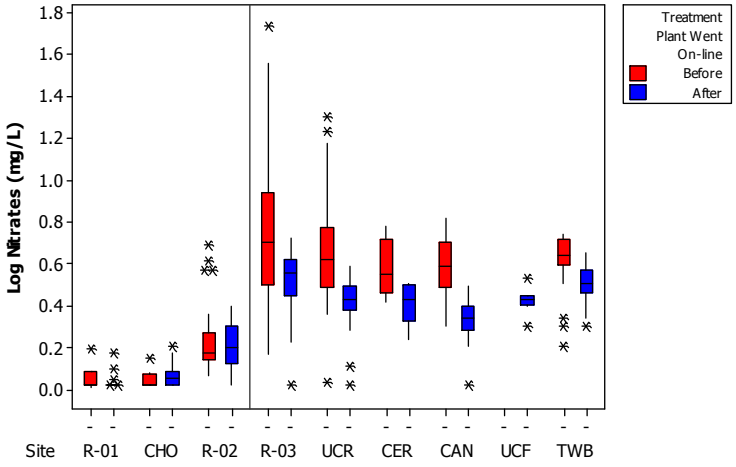
		Raw Data Median	Mean	SD	SE	N	Percent Below Threshold
R-01	**Before**	0.060	0.055	0.086	0.024	13	23.07
	After	0.023	0.019	0.024	0.004	30	26.67
CHO	**Before**	0.025	0.016	0.018	0.008	6	16.67
	After	0.020	0.009	0.005	0.001	29	6.89
R-02	**Before**	0.060	0.036	0.033	0.009	13	7.69
	After	0.046	0.028	0.032	0.006	31	12.90
R-03	**Before**	0.740	0.229	0.118	0.033	13	0
	After	0.900	0.277	0.086	0.015	31	0
UCR	**Before**	0.720	0.248	0.075	0.021	13	0
	After	0.770	0.246	0.071	0.008	74	0
CER	**Before**	0.855	0.234	0.068	0.028	6	0
	After	0.925	0.236	0.092	0.032	8	0
CAN	**Before**	0.770	0.233	0.048	0.014	12	0
	After	0.700	0.215	0.056	0.009	37	0
UCF	**Before**	NA	NA	NA	NA	0	0
	After	0.425	0.146	0.050	0.018	8	0
TWB	**Before**	0.390	0.151	0.032	0.007	21	0
	After	0.415	0.150	0.030	0.005	44	0

Note: The data below the threshold appears to only occur in the sites above the treatment plant.

Non-Detects with Original Substitution

Generally, it is expected that the nitrate and phosphate concentration is higher down stream from the plant, but we want to first look at boxplots of both the log nitrates and log phosphates to visualize any of the changes in the data over time. Figure 3 and 4 display boxplots comparing nitrates and phosphates (log-scale) over time and across locations.

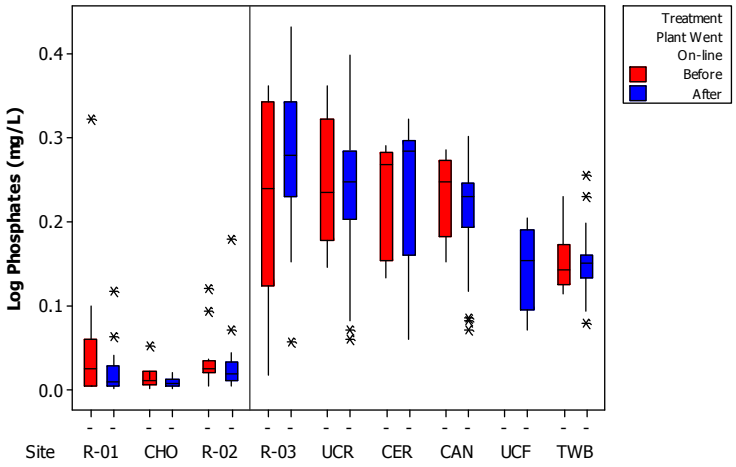
Figure 3: Boxplot of Log Nitrates



To visualize the data, before the treatment plant went on-line is labeled in red and after the plant went on-line is labeled in blue. The line between R-02 and R-03 indicates where the Men's Colony treatment plant is located, which distinguishes between the upstream and downstream locations.

After the treatment plant went on-line, the concentration of nitrates down stream appears to decrease. In comparison, the upstream values appear to be nearly the same both before and after the treatment, which is expected. The downstream boxplots show the differences between the before and after treatment factor, where the blue boxplots are generally lower with smaller medians and ranges that the red boxplots.

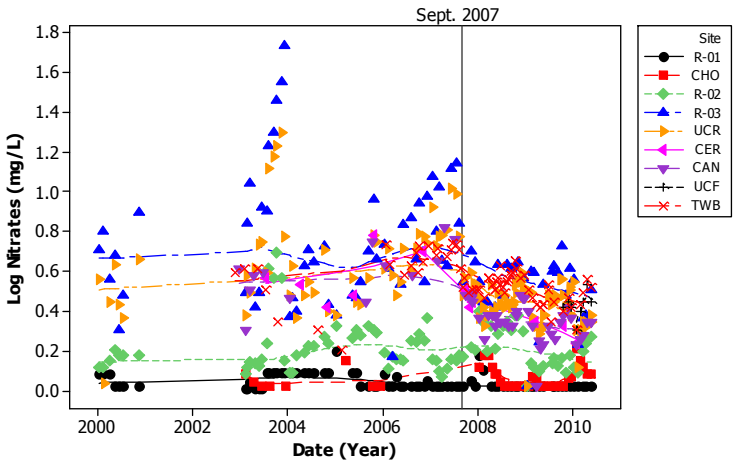
Figure 4: Boxplots of Log Phosphates



For the log phosphate data in Figure 4, we see a slight difference between the before and after data downstream from the plant, but a few appear to have larger means after the plant went on-line. Visually, the box plots downstream don't seem to have as much of a change as nitrates, but there is a large variability between upstream and downstream data.

To visualize the data for each site over the span of the time frame, Figure 5 and 6 display plots of the log nitrates and log phosphates.

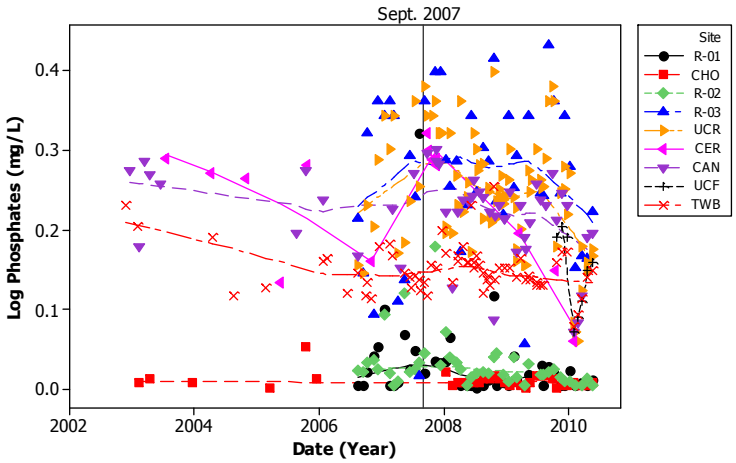
Figure 5: Plot of Log Nitrates from 2000 to 2010



Note: The plot shows the installation of the water treatment plant in September 2007.

As the plot shows, many recording from the R-01 and CHO locations consist non-detect values, as these values appear to form a straight line. There is a slight downward shift in the data after the treatment plant installation at sites R-03 through TWB.

Figure 6: Plot of Log Phosphates from 2000 to 2010



The non-detect values fall within the upstream sites as well, mostly including R-01, CHO, and R-02. Again there is visually no definite indication for a decrease in log phosphates after the installation of the water treatment plant.

2 Sample T-Tests:

To conduct a very simple assessment of the significance of the effect of the water treatment plant, we used a two-sample t-test to compare the mean differences before and after treatment plant went on-line and the mean differences of the upstream data to the downstream data. I want to see if the log values before and after the plant went on line differ for both nitrates and phosphates. This test will allow us to determine if there is a significant difference between the two means.

The data requirements for the 2 sample T-tests include: randomly sampled data, normality and independent values. I assume the data is random because there was no set pattern in the data collection. From the log transformation indicated on Page 7, the log data follows a normal distribution. Independence can't be verified because of the relationships that will occur within the before and after groups, but to execute the test, I assume independence.

To compare the means between before and after the plant went on-line:

H_O: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

Where μ_1 is the mean before the plant went on-line and μ_2 is mean after the plant went on-line.

There is strong evidence that there is a significant difference in log nitrates before and after the treatment plant went on-line ($t = 4.16$, $P < 0.001$). There is no evidence that

there is a significant difference in log phosphates before and after the treatment plant went on-line (t = -0.38, P = 0.701). See Appendix Tables 1A and 2A for detailed results.

Now to compare the means between the upstream and downstream data:

H_O: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

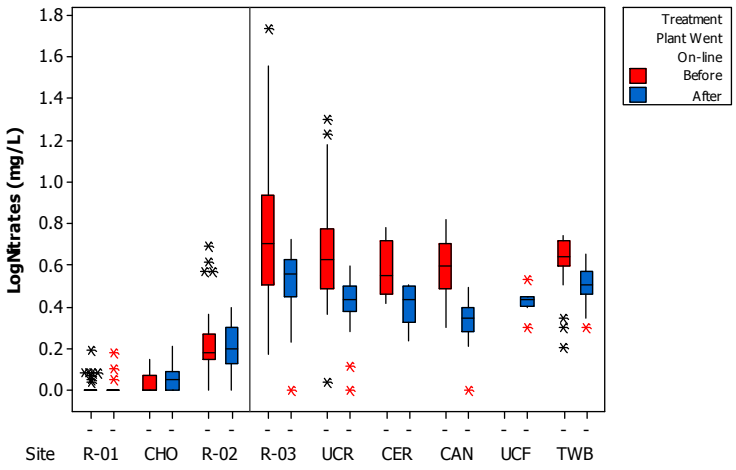
Where μ_1 is mean upstream from the water treatment plant and μ_2 is mean downstream from the water treatment plant.

There is strong evidence that there is a significant difference in log nitrates upstream and downstream from the water treatment plant (t = -29.38, P < 0.001). There is strong evidence that there is a significant difference in log phosphates upstream and downstream from the water treatment plant (t = -32.63, P < 0.001). See Appendix Tables 3A and 4A for detailed results.

Non-Detect Substitution of Zero

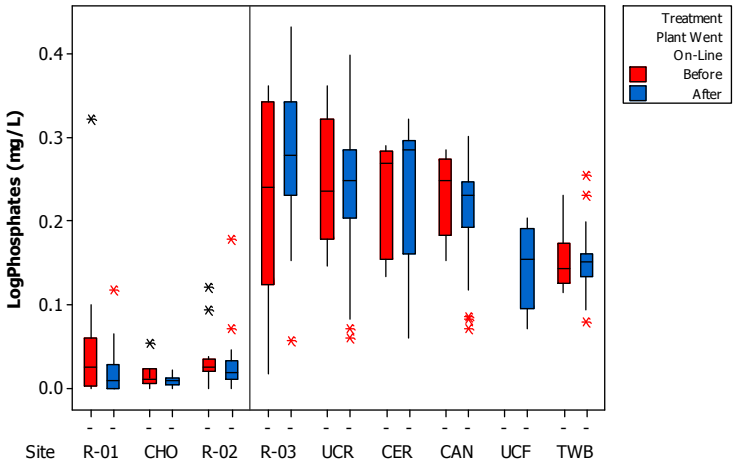
To address the affect of using the substitution method for non-detect data, a new substitution is made. A substitution of zero is made to see how it may affect or change the data visually and test results.

Figure 7: Boxplot of Log Nitrates with Non-Detect Substitution of Zero



As seen in Figure 7, after the treatment plant went on-line, the concentration of nitrates down stream appears to decrease. According to the log nitrate scale, the differences between the values are smaller than the original data. Further exploration of the differences will be made to find significant differences.

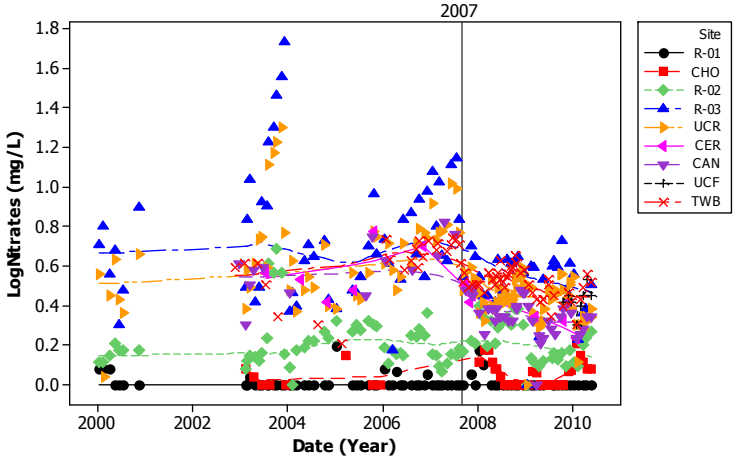
Figure 8: Boxplot of Log Phosphates with Non-Detect Substitution of Zero



Similarly to the original substituted data, Figure 8 shows the variability between the upstream and downstream data, where many of the values upstream were non-detect values. The values after the treatment plant went on-line appear to be larger than the values before the plant went on-line, which is generally an unexpected trend.

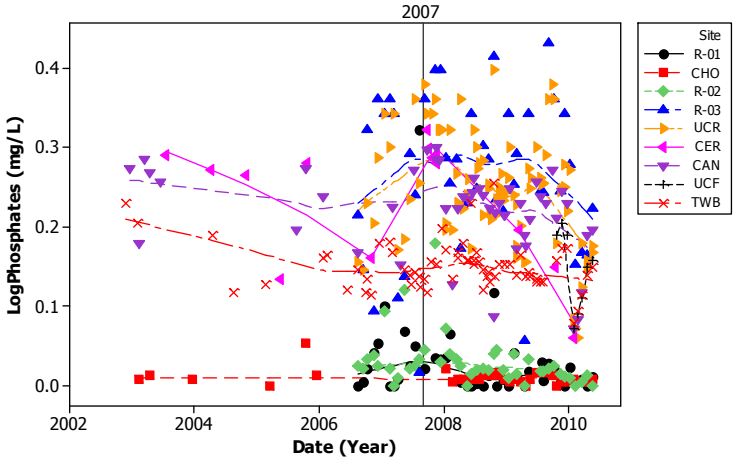
Figure 9 and 10 show the plots of this substituted data across the sites over the collection period.

Figure 9: Plot of Log Nitrates with Non-Detect Substitution of Zero from 2000 to 2010



Similarly to the original log nitrate substitution, Figure 9 shows the effect of the non-detects on the data. The non-detect values from R-01 and CHO sites are clearly shown as they form a straight line at the bottom of the log scale at 0 mg/L.

Figure 10: Plot of Log Phosphates with Non-Detect Substitution of Zero from 2000 to 2010



Compared to the values in Figure 6, Figure 10 shows that the zero substitution has the same effect visually. The non-detect values fall low and close to 0 mg/L.

2 Sample T-Tests:

To compare the means between before and after the plant went on-line:

H_O: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

Where μ_1 is the mean before the plant went on line and μ_2 is mean after the plant went on line.

There is strong evidence that there is a significant difference in log nitrates before and after the treatment plant went on-line ($t = 3.85$, $P < 0.001$).There is no evidence that there is a significant difference in log phosphates before and after the treatment plant went on-line ($t = -0.38$, $P = 0.702$). See Appendix Tables 5A and 6A for detailed results.

Now to compare the means between the upstream and downstream data:

H_O: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

Where μ_1 is mean upstream from the water treatment plant and μ_2 is mean downstream from the water treatment plant.

There is strong evidence that there is a significant difference in log nitrates upstream and downstream from the water treatment plant ($t = -29.51$, $P < 0.001$). There is no evidence that there is a significant difference in log phosphates upstream and downstream from the water treatment plant ($t = -32.63$, $P < 0.001$). See Appendix Tables 7A and 8A for detailed results.

Non-Detect Substitution of the Threshold

The final substitution made was the threshold levels from the original measuring devices. Note, these values are doubled the original substitution values. Thus intuitively, I anticipate the values to fall on a larger scale. To view the results, the boxplots and plots, Figures 11-14, are shown below.

Figure 11: Boxplot of Log Nitrates with Non-Detect Substitution of the Threshold

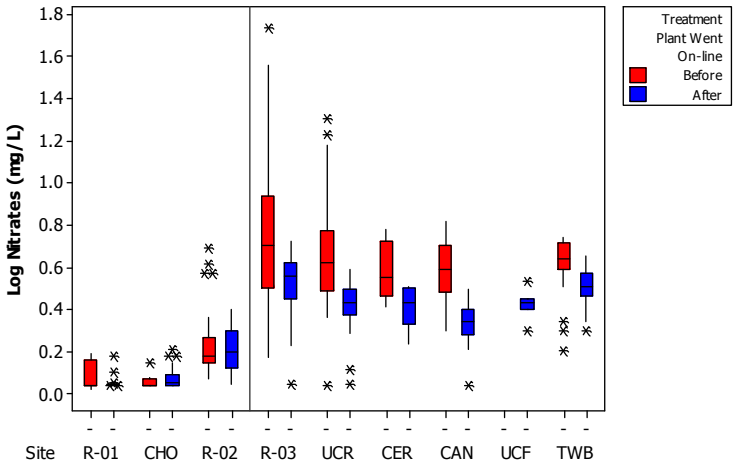
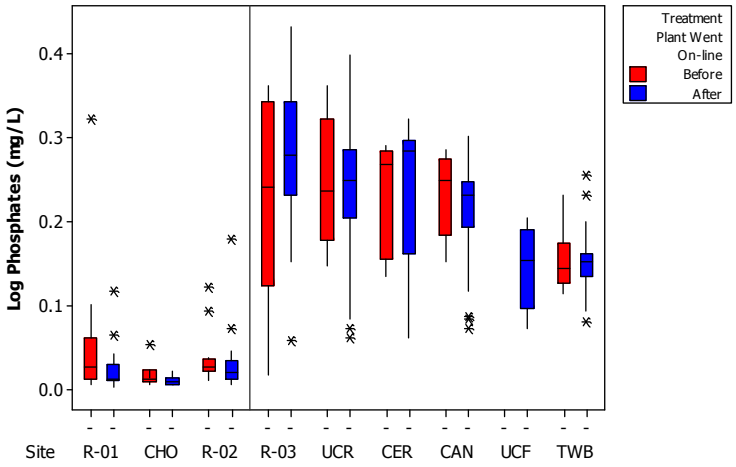


Figure 11 clearly shows a decrease in log nitrate concentration after the treatment plan went on-line over each site, but ultimately showing no visual difference from the original substitution.

Figure 12: Boxplot of Log Phosphates with Non-Detect Substitution of the Threshold



Similarly for the log phosphates, Figure 12 resembles the original substitution, in Figure 4. For further investigation, Figure 13 and 14 are the plots of the threshold substituted data.

Figure 13: Plot of Log Nitrates with Non-Detect Substitution of the Threshold from 2000 to 2010

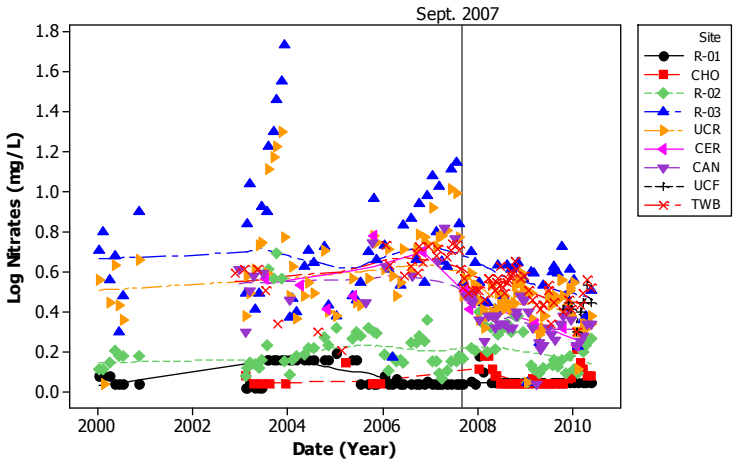
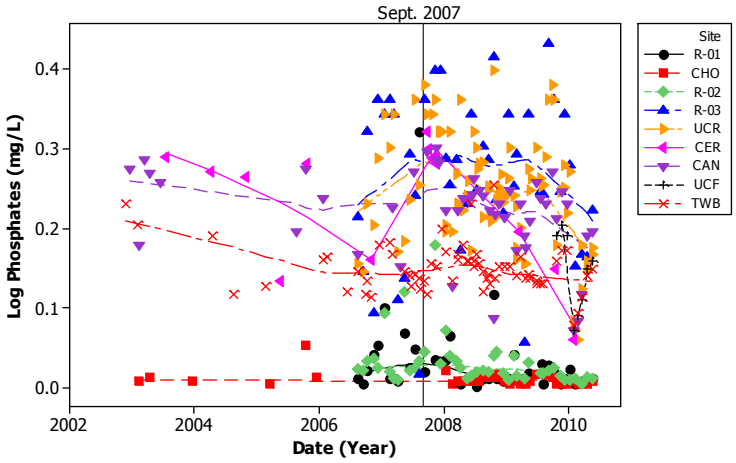


Figure 14: Plot of Log Phosphates with Non-Detect Substitution of the Threshold from 2000 to 2010



Similarly to the original substituted data, the non-detect values are clearly shown towards the bottom of the scale.

2 Sample T-Tests:

Again, to compare the means between before and after the plant went on-line:

H_O: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

Where μ_1 is the mean before the plant went on line and μ_2 is mean after the plant went on line.

There is strong evidence that there is a significant difference in log nitrates before and after the treatment plant went on-line ($t = 4.43$, $P < 0.001$). There is no evidence that

there is a significant difference in log phosphates before and after the treatment plant went on-line ($t = -0.38$, $P = 0.701$). See Appendix Tables 9A and 10A for detailed results.

Now to compare the means between the upstream and downstream data:

$H_O: \mu_1 = \mu_2$
$H_A: \mu_1 \neq \mu_2$

Where μ_1 is mean upstream from the water treatment plant and μ_2 is mean downstream from the water treatment plant.

There is strong evidence that there is a significant difference in log nitrates upstream and downstream from the water treatment plant ($t = -28.97$, $P < 0.001$). There is no evidence that there is a significant difference in log phosphates upstream and downstream from the water treatment plant ($t = -32.6$, $P < 0.001$). See Appendix Tables 11A and 12A for detailed results.

Comparison of all Substitutions

Table 3: Summary Statistics for Log Nitrate Data

Log Nitrates		Median	Mean			SD			SE			N	Percent below threshold
		Raw	Original	Zero	Threshold	Original	Zero	Threshold	Original	Zero	Threshold		
R-01	**Before**	0.050	0.050	0.013	0.082	0.038	0.035	0.058	0.005	0.005	0.008	53	86.79
	After	0.055	0.032	0.011	0.051	0.031	0.037	0.026	0.006	0.007	0.005	30	86.67
CHO	**Before**	0.050	0.047	0.033	0.059	0.045	0.054	0.038	0.016	0.019	0.013	8	62.50
	After	0.125	0.066	0.056	0.075	0.055	0.064	0.048	0.010	0.012	0.009	30	46.67
R-02	**Before**	0.500	0.219	0.218	0.221	0.133	0.136	0.132	0.018	0.019	0.018	53	1.88
	After	0.480	0.213	0.212	0.214	0.102	0.103	0.100	0.018	0.019	0.018	31	3.22
R-03	**Before**	4.050	0.760	0.760	0.760	0.323	0.323	0.323	0.045	0.045	0.045	52	0
	After	2.600	0.515	0.514	0.516	0.152	0.155	0.150	0.027	0.028	0.027	31	3.22
UCR	**Before**	3.200	0.659	0.659	0.659	0.236	0.236	0.236	0.032	0.032	0.032	53	0
	After	1.700	0.429	0.429	0.429	0.101	0.102	0.100	0.012	0.012	0.012	74	1.35
CER	**Before**	2.550	0.578	0.578	0.578	0.137	0.137	0.137	0.056	0.056	0.056	6	0
	After	1.700	0.406	0.406	0.406	0.098	0.098	0.098	0.035	0.035	0.035	8	0
CAN	**Before**	2.900	0.591	0.591	0.591	0.142	0.142	0.142	0.039	0.039	0.039	13	0
	After	1.200	0.344	0.343	0.344	0.093	0.095	0.091	0.015	0.015	0.015	38	2.63
UCF	**Before**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0
	After	1.700	0.425	0.425	0.425	0.065	0.065	0.065	0.023	0.023	0.023	8	0
TWB	**Before**	3.350	0.612	0.612	0.612	0.142	0.142	0.142	0.029	0.029	0.029	24	0
	After	2.200	0.508	0.508	0.508	0.073	0.073	0.073	0.011	0.011	0.011	45	0

Table 4: Summary Statistics for Log Phosphate Data

Log Phosphates		Median	Mean			SD			SE			N	Percent below threshold
		Raw	Original	Zero	Threshold	Original	Zero	Threshold	Original	Zero	Threshold		
R-01	**Before**	0.060	0.055	0.053	0.056	0.086	0.087	0.085	0.024	0.024	0.024	13	23.07
	After	0.023	0.019	0.018	0.021	0.024	0.025	0.023	0.004	0.004	0.004	30	26.67
CHO	**Before**	0.025	0.016	0.016	0.017	0.018	0.019	0.018	0.008	0.008	0.007	6	16.67
	After	0.020	0.009	0.009	0.009	0.005	0.005	0.005	0.001	0.001	0.001	29	6.89
R-02	**Before**	0.060	0.036	0.036	0.037	0.033	0.034	0.033	0.009	0.009	0.009	13	7.69
	After	0.046	0.028	0.027	0.028	0.032	0.032	0.031	0.006	0.006	0.006	31	12.90
R-03	**Before**	0.740	0.229	0.229	0.229	0.118	0.118	0.118	0.033	0.033	0.033	13	0
	After	0.900	0.277	0.277	0.277	0.086	0.086	0.086	0.015	0.015	0.015	31	0
UCR	**Before**	0.720	0.248	0.248	0.248	0.075	0.075	0.075	0.021	0.021	0.021	13	0
	After	0.770	0.246	0.246	0.246	0.071	0.071	0.071	0.008	0.008	0.008	74	0
CER	**Before**	0.855	0.234	0.234	0.234	0.068	0.068	0.068	0.028	0.028	0.028	6	0
	After	0.925	0.236	0.236	0.236	0.092	0.092	0.092	0.032	0.032	0.032	8	0
CAN	**Before**	0.770	0.233	0.233	0.233	0.048	0.048	0.048	0.014	0.014	0.014	12	0
	After	0.700	0.215	0.215	0.215	0.056	0.056	0.056	0.009	0.009	0.009	37	0
UCF	**Before**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0
	After	0.425	0.146	0.146	0.146	0.050	0.050	0.050	0.018	0.018	0.018	8	0
TWB	**Before**	0.390	0.151	0.151	0.151	0.032	0.032	0.032	0.007	0.007	0.007	21	0
	After	0.415	0.150	0.150	0.150	0.030	0.030	0.030	0.005	0.005	0.005	44	0

To compare the results and values from the three substitutions, Tables 3 and 4 include the descriptive statistics for these substitutions for both log nitrates and log phosphates. To further compare the results between each substitution, the appropriate analysis is an Analysis of Variance (ANOVA) test.

ANOVA:

Because the analysis deals with the nine specific stream locations, there is data for nine sites with two distinguishing groups of before and after the treatment plant went on-line. The ANOVA test that will be performed will be a Blocked ANOVA test, with site as the blocking variable, to compare the differences before and after the treatment plant went on-line. Ultimately, the goal is to find out if the effect of the before and after the treatment plant went on-line factor depends on the upstream and downstream effect.
Note, the data assumptions are: independence, normality, equal variances.

H_O: $\mu = 0$
H_A: $\mu \neq 0$

Where μ is the mean for the interaction term between before and after the treatment plant went on-line and the upstream and downstream factor.

Table 5a and b: ANOVA Results for Log Nitrates

Log Nitrates	Original			Zero			Threshold		
	Adj. MS	**F**	**P-value**	**Adj. MS**	**F**	**P-value**	**Adj. MS**	**F**	**P-value**
Treatment	1.337	57.11	<0.001	1.246	52.68	<0.001	1.409	60.19	<0.001
Location	14.597	623.38	<0.001	15.581	658.65	<0.001	13.737	586.93	<0.001
Treatment*Location	1.152	49.20	<0.001	1.249	52.78	<0.001	1.080	46.12	<0.001
Site(Location)	0.362	15.48	<0.001	0.429	18.11	<0.001	0.313	13.35	<0.001
Error	0.023			0.024			0.023		
R-Squared(adjusted)	69.94%			71.14%			68.66%		

	Coefficient	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Constant	0.324	0.009	0.317	0.009	0.330	0.009
Treatment (Before)	0.054	0.007	0.052	0.007	0.055	0.007
Location (up)	-0.216	0.009	-0.224	0.009	-0.210	0.009
Treatment*Location	-0.050	0.007	-0.052	0.007	-0.048	0.007

Note: Treatment is before or after the treatment plant went on-line

To see if there is an effect from before and after the treatment plant went on-line by the groups which were each site, the results from the interaction term of Treatment by Location are addressed from Table 5a for the log nitrates. From the original data, at a 0.05 significance level, there is statistical evidence that there is a difference between the upstream and downstream location according to the before and after effect of when the

treatment plant went on-line (F = 49.2, P < 0.001). This indicates the evidence of the before or after factor depending on location. From the zero substitution, the result is relatively the same but the F statistic is larger, so the strength of the result is greater. Based on the threshold results, again the result is relatively the same, the F statistic is smaller than the original substitution, so there is a slight increase in variability in the threshold data and the evidence is not as strong.

Table 6a and b: ANOVA Results for Log Phosphates

Log Phosphates	Original			Zero			Threshold		
	Adj. MS	F	P-value	Adj. MS	F	P-value	Adj. MS	F	P-value
Treatment	0.003	0.78	0.376	0.003	0.800	0.372	0.003	0.77	0.381
Location	1.653	481.35	<0.001	1.666	483.320	<0.001	1.640	478.96	<0.001
Treatment*Location	0.009	2.62	0.106	0.009	2.640	0.105	0.009	2.60	0.108
Site(Location)	0.072	20.99	<0.001	0.072	20.890	<0.001	0.072	21.09	<0.001
Error	0.003			0.003			0.003		
R-Squared(adjusted)	72.95%			73.01%			72.88%		

	Coefficient	SE	Coefficient	SE	Coefficient	SE
Constant	0.118	0.004	0.118	0.004	0.119	0.004
Treatment (0)	0.003	0.004	0.003	0.004	0.003	0.004
Location (up)	-0.090	0.004	-0.091	0.004	-0.090	0.004
Treatment*Location	0.006	0.004	0.006	0.004	0.006	0.004

Based on Table 6a for the log phosphates, at a 0.05 significance level for the original data results, there is no evidence that there is a difference between the upstream and downstream location according to the before or after effect of when the treatment plant went on-line. With a similar F statistic, the zero substitution obtains the same result as well as the threshold substitution. Thus, the effect of the before and after factor does not depend on location, for any of the log phosphate substitutions.

Simulation

Based on the previous ANOVA results for both the log nitrate and log phosphate data, two questions lead to the next analysis: Do these three substitutions give any further explanation about the effect of non-detects? And would other substituted values provide the same results as these? To answer these questions, simulation of the non-detects is done to distinguish if sound conclusions can be made from these results or possibly if different results would be obtained.

Because the true distribution of the non-detect data is unknown, a uniform distribution from zero to the respective thresholds is used for the simulation generation. Using the program R, the simulation run replaces the non-detects and obtains the ANOVA results from the 1000 simulated datasets.

To see examine the distribution of ANOVA p-values to test for a treatment plant effect downstream, I made a histogram to view the frequency and range of the simulated p-values for the location and upgrade interaction. Figure 15 and 16 show where these results fall in comparison to the previous substitutions. These also lead to answering the question: Do I obtain the same results?

Figure 15: Histogram of Simulated P-Values for Log Nitrates

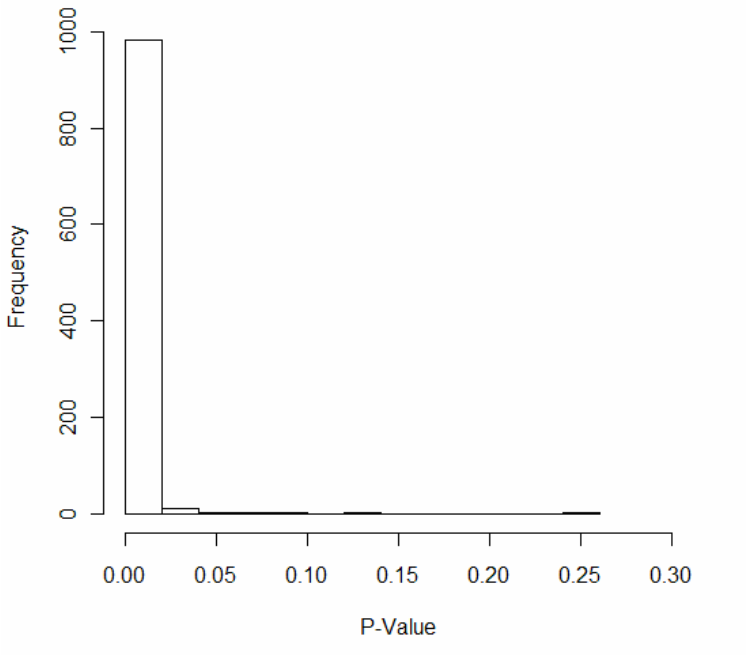


Figure 16: Histogram of Simulated P-Values for Log Phosphates

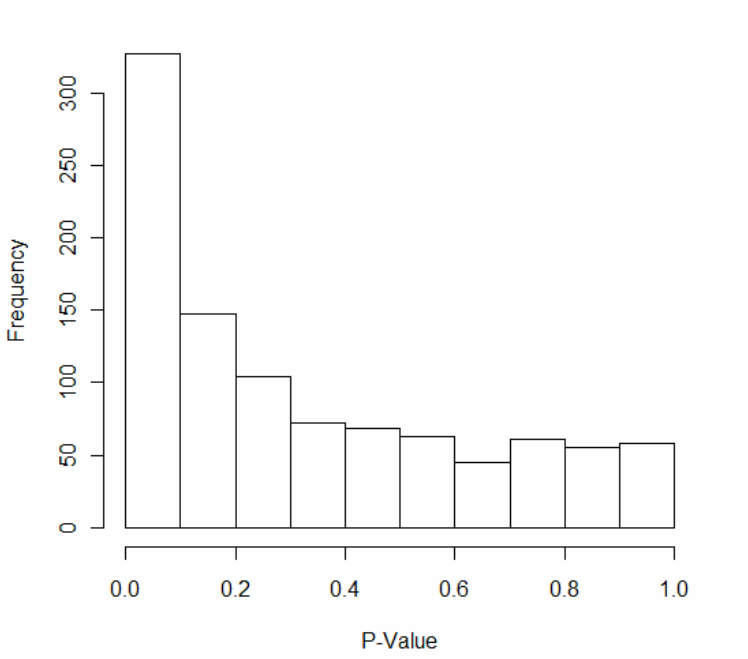


Figure 15 shows that when simulating values for non-detects using a uniform distribution, the results are usually significant ($P < 0.001$), and provide evidence for a significant difference between upstream and downstream interaction with the before and after effect, similar to the results of the previous substitutions ($P < 0.001$ for each).

Figure 16 shows that the simulated results fall across the spectrum ranging from less than 0.001 and 1. Many of the p-values fall at a significant level, around 0.001 and 0.05, with a count of approximately 350 out of 1000, but most results fall in the "insignificant" range. The results from the three previous substitutions fall where the second bar is located ($P = 0.106, 0.105, 0.108$). About 150 results out of 1000 from the simulation were between 0.1 and 0.2, which is only 15% of the total simulated results. Because of the wide range of results, overall, these are not as consistent with the previous results, indicating that the substitution for the log phosphates had an affect on the data and outcomes.

Conclusion

Because of the ambiguity posed by non-detects, they can have an affect on environmental data. Using the Chorro Creek case study, this affect was addressed using tests, comparisons, and simulations. Thus, the main questions were: What was the affect made by the non-detects on the data? Did the original substitutions prove to be valid for the study and what conclusions can be made? For the log nitrates, the simulation reinforced some validity in the previous substitutions made, considering the results obtained were the same. In this case, the original substitution was a decent data organization approach to use. Because the simulation provided different results for the log phosphates, in this case, non-detects affect the conclusions that might made in this case study and conclusions can not be made.

Ultimately, non-detects clearly have an affect on this case study and data set. Due to the differences shown, from the simulation, attention and care need to be made when handling data with non-detect values and making substitutions. Further analysis of the true distribution of the non-detects could provide more accurate comparisons of the affect of these values on environmental data.

Appendix

Locations:
Upstream: R-01, CHO, R-02
Downstream: R-03, UCR, CER, CAN, UCF, and TWB

Site Details:

MBNEP sites:

CHO is upstream of the CMC WWTP.
UCR, CER, CAN and TWB are downstream of the CMC WWTP.

CMC sites:

R-01 is below Chorro Dam.

R-02 is approximately 100' upstream from the CMC outfall.

R-03 is approximately 100' downstream from the CMC outfall.

Notes about Treatment plant:
Time period before the plant went on-line is between 1/12/200 and 8/16/2007, and the time period after the plant went on-line is between 9/11/2007 to 5/18/2010.

Two-Sample T Test Results:

Original Substitution

Table 1A: Two-Sample T-Test for Log Nitrates Before and After the Treatment Plant Went On-line

Treatment	N	Mean	Standard Deviation	SE Mean
0	262	0.439	0.344	0.021
1	295	0.338	0.194	0.011

T-Value	P-Value	DF
4.16	< 0.001	400

Table 2A: Two-Sample T-Test for Log Nitrates Upstream and Downstream from the Water Treatment Plant

Location	N	Mean	Standard Deviation	SE Mean
Up	205	0.118	0.118	0.008
Down	352	0.541	0.222	0.012

T-Value	P-Value	DF
-29.38	< 0.001	551

Table 3A: Two-Sample T-Test for Log Phosphates Before and After the Treatment Plant Went On-line

Treatment	N	Mean	Standard Deviation	SE Mean
0	97	0.153	0.108	0.011
1	292	0.158	0.114	0.007

T-Value	P-Value	DF
-0.38	0.701	172

Table 4A: Two-Sample T-Test for Log Phosphates Upstream and Downstream from the Water Treatment Plant

Location	N	Mean	Standard Deviation	SE Mean
Up	122	0.025	0.038	0.003
Down	267	0.217	0.079	0.005

T-Value	P-Value	DF
-32.63	< 0.001	386

Zero Substitution

Table 5A: Two-Sample T-Test Log Nitrates with Substitution Zero Before and After the Treatment Plant Went On-line

Treatment	N	Mean	Standard Deviation	SE Mean
0	262	0.430	0.354	0.022
1	295	0.335	0.199	0.012

T-Value	P-Value	DF
3.85	< 0.001	400

Table 6A: Two-Sample T-Test Log Nitrates with Substitution Zero Upstream and Downstream from the Water Treatment Plant

Location	N	Mean	Standard Deviation	SE Mean
Up	205	0.103	0.129	0.009
Down	352	0.541	0.222	0.012

T-Value	P-Value	DF
-29.51	< 0.001	554

Table 7A: Two-Sample T-Test Log Phosphates with Substitution Zero Before and After the Treatment Plant Went On-line

Treatment	N	Mean	Standard Deviation	SE Mean
0	97	0.153	0.109	0.011
1	292	0.158	0.115	0.007

T-Value	P-Value	DF
-0.38	0.702	172

Table 8A: Two-Sample T-Test Log Phosphates with Substitution Zero Upstream and Downstream from the Water Treatment Plant

Location	N	Mean	Standard Deviation	SE Mean
Up	122	0.024	0.038	0.004
Down	267	0.217	0.079	0.005

T-Value	P-Value	DF
-32.63	< 0.001	385

Threshold Substitution

Table 9A: Two-Sample T-Test Log Nitrates with Substitution of the Threshold Before and After the Treatment Plant Went On-line

Treatment	N	Mean	Standard Deviation	SE Mean
0	262	0.446	0.337	0.021
1	295	0.342	0.188	0.011

T-Value	P-Value	DF
4.43	<0.001	398

Table 10A: Two-Sample T-Test Log Nitrates with Substitution of the Threshold Upstream and Downstream from the Water Treatment Plant

Location	N	Mean	Standard Deviation	SE Mean
Up	205	0.131	0.112	0.008
Down	352	0.542	0.221	0.012

T-Value	P-Value	DF
-28.97	0<0.001	546

Table 11A: Two-Sample T-Test Log Phosphates with Substitution of the Threshold Before and After the Treatment Plant Went On-line

Treatment	N	Mean	Standard Deviation	SE Mean
0	97	0.153	0.108	0.011
1	292	0.158	0.114	0.007

T-Value	P-Value	DF
-0.38	0.701	172

Table 12A: Two-Sample T-Test Log Phosphates with Substitution of the Threshold Upstream and Downstream from the Water Treatment Plant

Location	N	Mean	Standard Deviation	SE Mean
Up	122	0.025	0.037	0.003
Down	267	0.217	0.079	0.005

T-Value	P-Value	DF
-32.60	<0.001	386

ANOVA Assumptions:

Original Substitution

Figure 1A: ANOVA Assumptions for Log Nitrates with Original Substitution

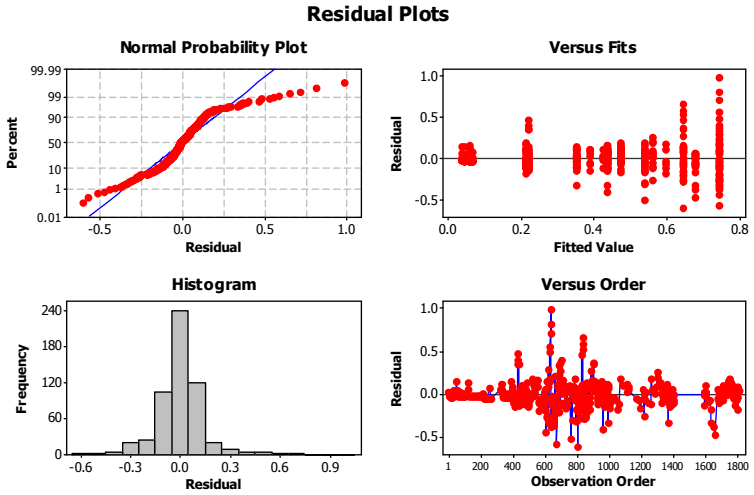
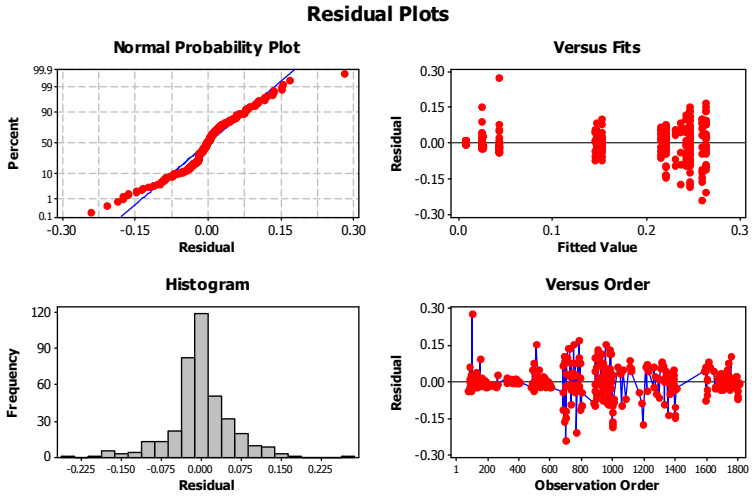


Figure 2A: ANOVA Assumptions for Log Phosphates with Original Substitution



Zero Substitution

Figure 3A: ANOVA Assumptions for Log Nitrates with Zero Substitution

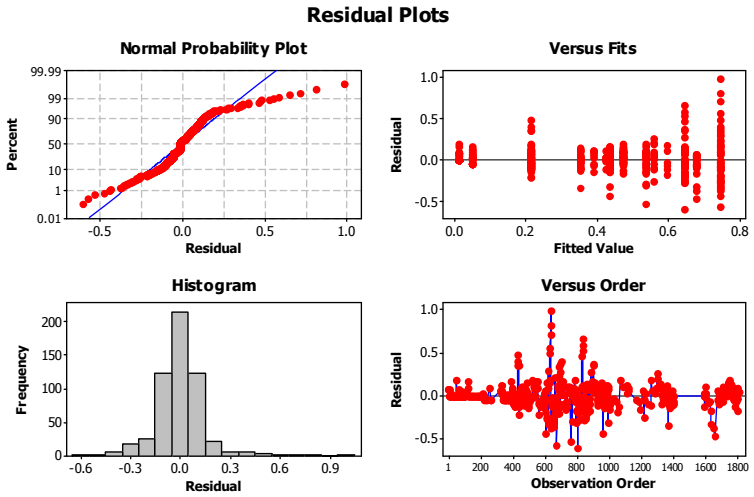
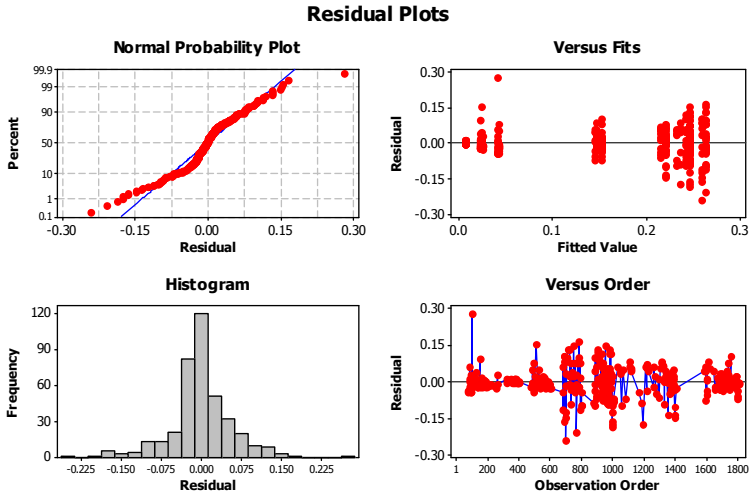


Figure 4A: ANOVA Assumptions for Log Phosphates with Zero Substitution



Threshold Substitution

Figure 5A: ANOVA Assumptions for Log Nitrates with Threshold Substitution

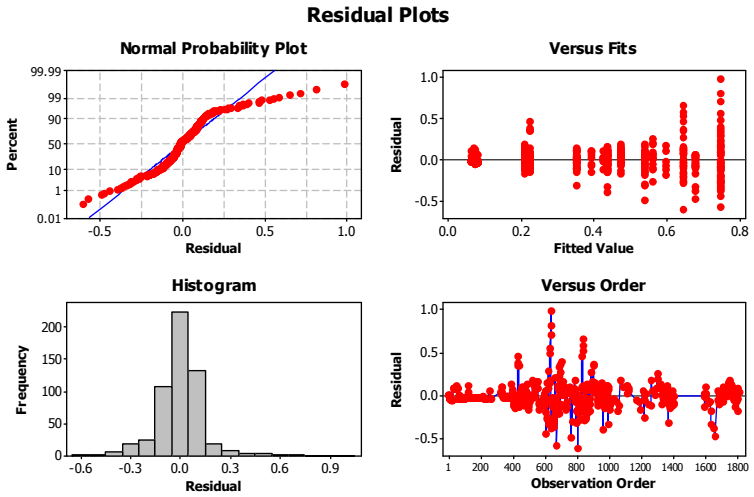
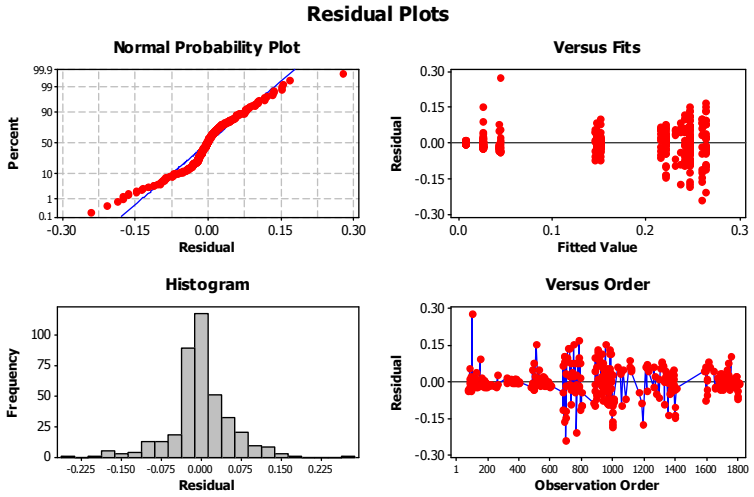


Figure 6A: ANOVA Assumptions for Log Phosphates with Threshold Substitution



ANOVA Main Affects and Interaction Plots:

Figure 7A: ANOVA Main Affects Plots for Log Phosphates with Original Substitution

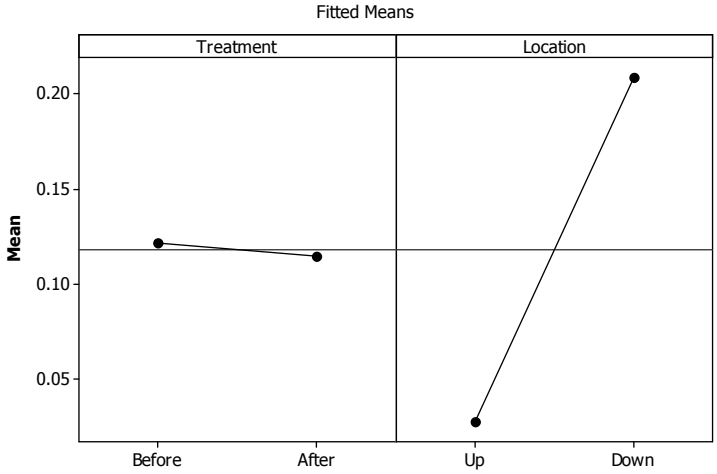


Figure 8A: ANOVA Interaction Plot for Log Phosphates with Original Substitution

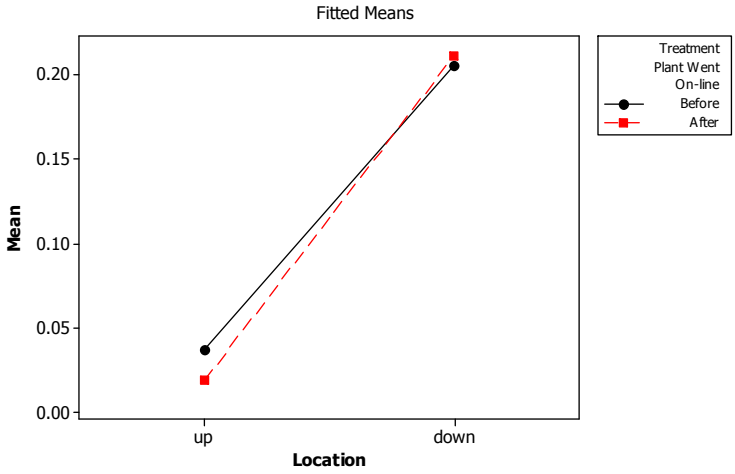


Figure 9A: ANOVA Main Affects Plots for Log Nitrates with Original Substitution

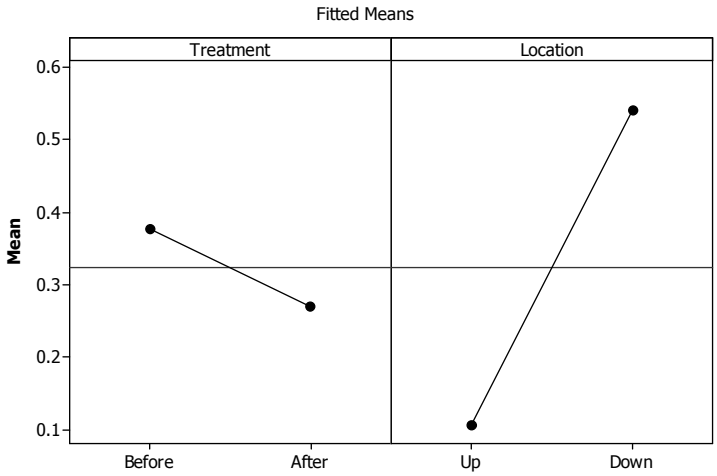


Figure 10A: ANOVA Interaction Plot for Log Nitrates with Original Substitution

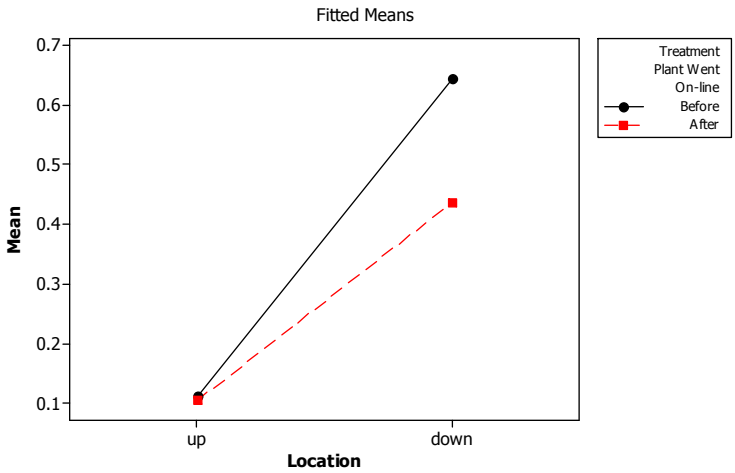


Figure 11A: ANOVA Main Affects Plots for Log Phosphates with Zero Substitution

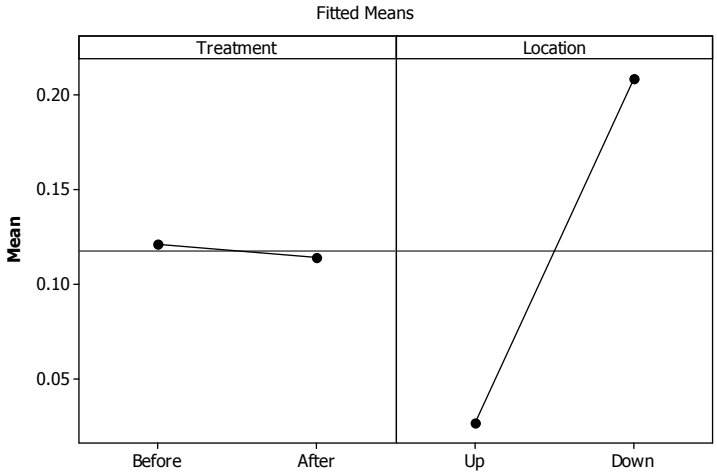


Figure 12A: ANOVA Interaction Plot for Log Phosphates with Zero Substitution

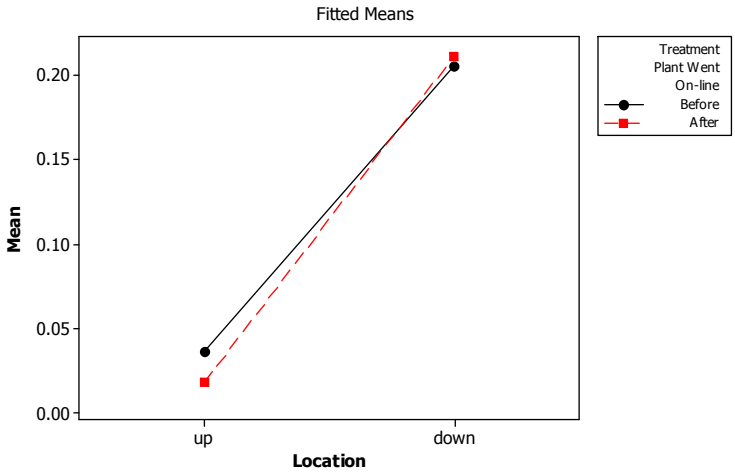


Figure 13A: ANOVA Main Affects Plots for Log Nitrates with Zero Substitution

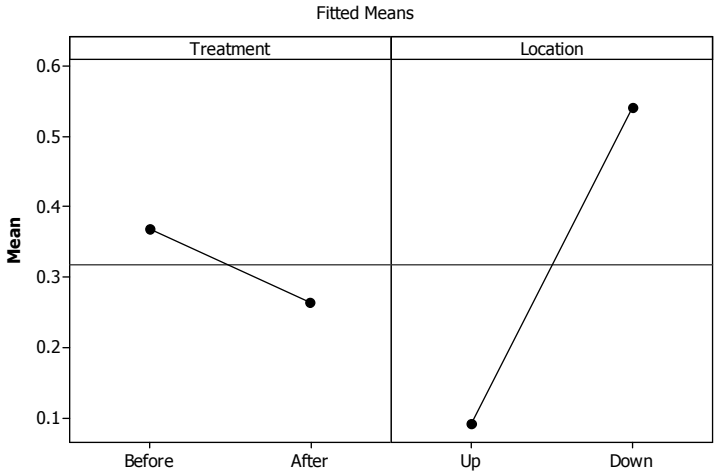


Figure 14A: ANOVA Interaction Plot for Log Nitrates with Zero Substitution

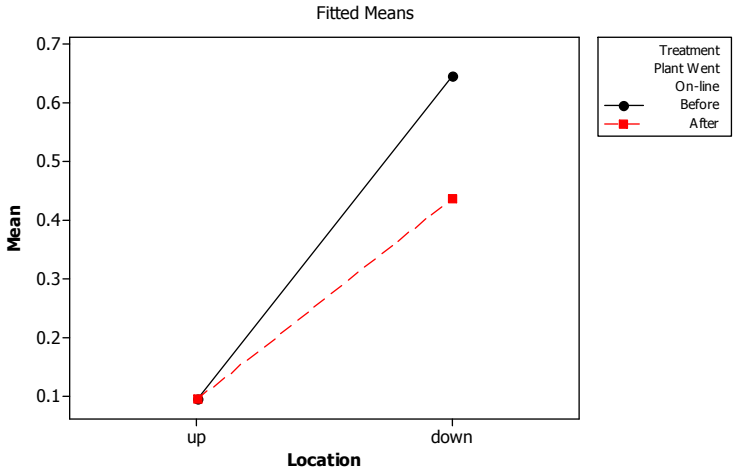


Figure 15A: ANOVA Main Affects Plots for Log Phosphates with Threshold Substitution

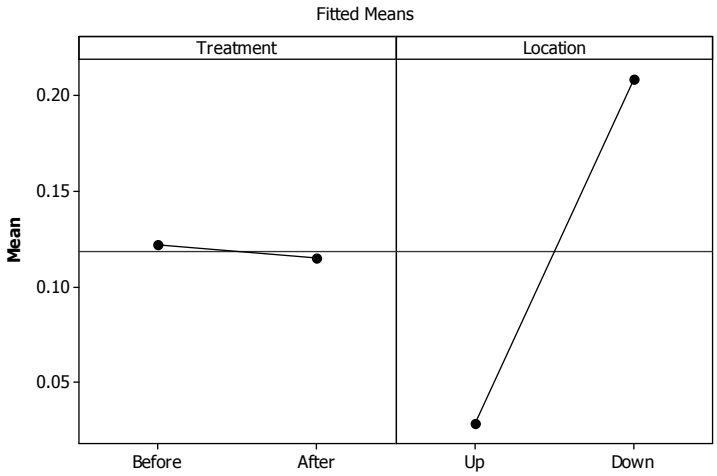


Figure 16A: ANOVA Interaction Plot for Log Phosphates with Threshold Substitution

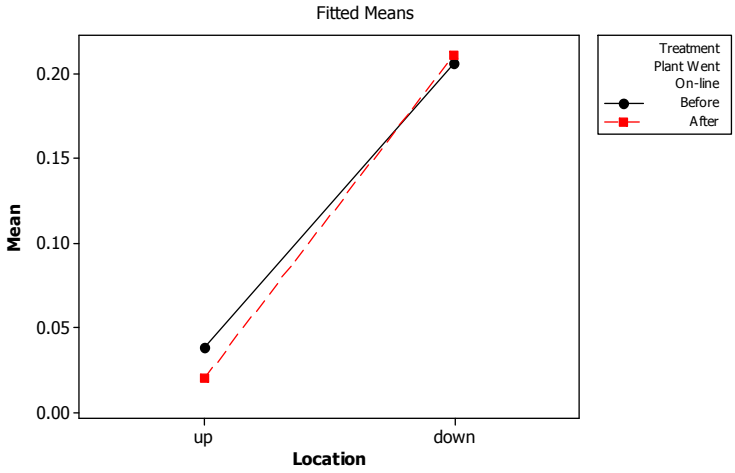


Figure 17A: ANOVA Main Affects Plots for Log Nitrates with Threshold Substitution

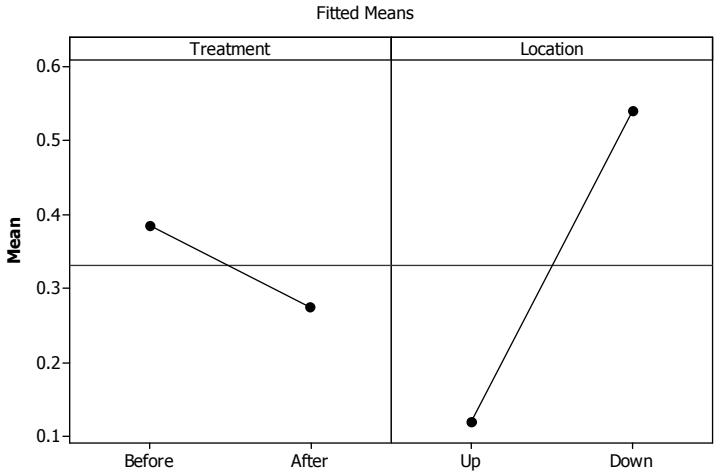
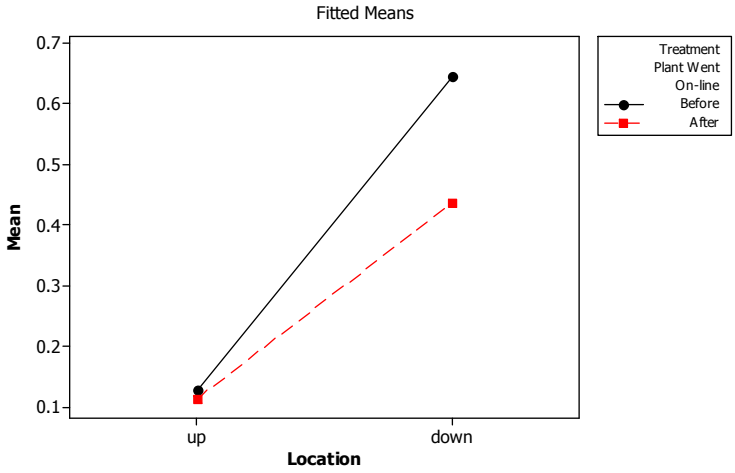


Figure 18A: ANOVA Interaction Plot for Log Nitrates with Threshold Substitution



R Code for Simulation:

```
rm(list=ls())
Creek=read.csv("C:/Users/juliana/Desktop/FinalData.CSV", header=TRUE)
Creek0=Creek
Creek.T=Creek
str(Creek)

PhosRows = which(Creek$Indicators.for.Phosphates == "<")
NitRows = which(Creek$Indicators.for.Nitrates == "<")

### Simulation for Phosphates
p.valueP = matrix(0, nrow = 1000, ncol =4)
for(SimItor in 1:1000){

        CreekS = Creek
        for(i in 1:length(PhosRows)){
                CreekS$logPhosphates[PhosRows[i]] = log10(runif(1, 0,
2*10^Creek$logPhosphates[PhosRows[i]]))
        }
        m1 = lm(logPhosphates~Treatment + Location + Site%in%Location +
Treatment*Location, data=CreekS)
        anova(m1)
        p.valueP[SimItor,] =anova(m1)[[5]][1:4]
        #cat(SimItor)
}
hist(p.valueP[,4],xlim=c(0,1),main="Histogram of Simulated P-Values for Log
Phosphates",xlab="P-Value")

#### Simulation for Nitrates
p.valueN = matrix(0, nrow = 1000, ncol =4)
for(SimItor in 1:1000){

        CreekS = Creek
        for(i in 1:length(NitRows)){
                CreekS$logNitrates[NitRows[i]] = log10(runif(1, 0,
2*10^Creek$logNitrates[NitRows[i]]))
        }
        m1 = lm(logNitrates~Treatment + Location + Site%in%Location +
Treatment*Location, data=CreekS)
        anova(m1)
        p.valueN[SimItor,] =anova(m1)[[5]][1:4]
}
hist(p.valueN[,4],xlim=c(0,.3),main="Histogram of Simulated P-Values for Log
Nitrates",xlab="P-Value")
```

Works Cited

Helsel, Dennis R. *Non-detects and Data Analysis: Statistics for Censored Environmental Data*. Hoboken, NJ: Wiley-Interscience, 2005.

McKinney, Michael L., Robert M. Schoch, and Logan Yonavjak. *Environmental Science: Systems and Solutions.* Sudbury, MA: Jones and Bartlett, 2007.

"DEFINITION." *LEO - Lehigh Earth Observatory*. Lehigh Earth Observatory, 2006. <http://www.leo.lehigh.edu/projects/hydroprobe/wqdef.html>.

"Water Resource Characterization DSS - Phosphorus." *North Carolina Extension Water Quality Information System*. NCSU Water Quality Group, 2007. <http://www.water.ncsu.edu/watershedss/info/phos.html>.