Analytical Support for Identifying Water Quality Thresholds in New Mexico Surface Waters

FINAL REPORT

J. Thad Scott[1] and Brian E. Haggard[2]

[1]Department of Crop, Soil, and Environmental Sciences, Division of Agriculture, University of Arkansas; jts004@uark.edu

[2]Arkansas Water Resources Center, Division of Agriculture, University of Arkansas; haggard@uark.edu

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY 6
2. INTRODUCTION 7
3. DATA REDUCTION AND OUTLIER ANALYSIS 7
3.1 SUMMARY 7
3.2 DESCRIPTIVE STATISTICS 8
3.3 OUTLIER ANALYSIS 8
4. ANALYSIS OF LAKE AND RESERVOIR DATA 10
4.1 SUMMARY 10
4.2 OBJECTIVES AND METHODS 10
4.3. RESULTS 12
Changepoints in total phosphorus concentration 12
Changepoints in total nitrogen concentration 13
Regression tree models 14
Cyanobacteria and Water Quality 16
Analysis details and conclusions 17
5. ANALYSIS OF RIVER DATA 17
5.1 SUMMARY 17
5.2 OBJECTIVES AND METHODS 18
5.3 RESULTS 19
Changepoint in total phosphorus concentration 19
Changepoints in total nitrogen concentration 20
Categorical and Regression Tree Model 20
Analysis details and conclusions 21
6. CONCLUSIONS 22
7. REFERENCES 23
APPENDIX I – R CODE & OUTPUT FOR LAKE AND RESERVOIR DATA 24
APPENDIX II – R CODE & OUTPUT FOR RIVER DATA 57
APPENDIX III – R CODE & OUTPUT FOR REVISIONS 66

LIST OF TABLES

Table 1.1. Total phosphorus and total nitrogen thresholds that at which biological shifts occurred in New Mexico Surface Waters.

Table 5.1. Results of Categorical and Regression Tree (CART) analysis on benthic chlorophyll-a. Primary split predictors are listed in order of prediction strength. Only one split was statistically valid in the tree.

Table 6.1. Total phosphorus and total nitrogen thresholds that at which biological shifts occurred in New Mexico Surface Waters.

LIST OF FIGURES

Figure 4.1. Results of Changepoint Analysis on median total P and median secchi depth values for all New Mexico Lakes and Reservoirs.

Figure 4.2. Results of Changepoint Analysis on median total P and median Zeu values for all New Mexico Lakes and Reservoirs.

Figure 4.3. Results of Changepoint Analysis on median total P and median chlorophyll-a concentrations for all New Mexico Lakes and Reservoirs.

Figure 4.4. Results of Changepoint Analysis on median total N and median secchi depth values for all New Mexico Lakes and Reservoirs.

Figure 4.5. Results of Changepoint Analysis on median total N and median Zeu values for all New Mexico Lakes and Reservoirs.

Figure 4.6. Results of Changepoint Analysis on median total P and median chlorophyll-a concentrations for all New Mexico Lakes and Reservoirs.

Figure 4.7. Regression Tree Analysis of secchi depth across all New Mexico lakes and reservoirs.

Figure 4.8. Regression Tree Analysis of Zeu across all New Mexico lakes and reservoirs.

Figure 4.9. Regression Tree Analysis of chlorophyll-a concentration across all New Mexico lakes and reservoirs.

Figure 4.10. Changepoint analysis of secchi depth versus percent cyanobacteria in New Mexico lakes and reservoirs.

Figure 4.11. Changepoint analysis of chlorophyll-a versus percent cyanobacteria in New Mexico lakes and reservoirs.

Figure 4.12. Changepoint analysis of total phosphorus versus percent cyanobacteria in New Mexico lakes and reservoirs.

Figure 4.13. Changepoint analysis of total nitrogen versus percent cyanobacteria in New Mexico lakes and reservoirs.

Figure 4.14. Changepoint analysis of summer total phosphorus versus percent cyanobacteria in summer in New Mexico lakes and reservoirs.

Figure 4.15. Changepoint analysis of summer total nitrogen versus percent cyanobacteria in summer in New Mexico lakes and reservoirs.

Figure 5.1. Results of Changepoint Analysis on total P and benthic chlorophyll-a from New Mexico rivers.

Figure 5.2. Results of Changepoint Analysis on total P and Trophic Diatom Index from New Mexico rivers.

Figure 5.3. Results of Changepoint Analysis on total N and benthic chlorophyll-a from New Mexico rivers.

Figure 5.4. Results of Changepoint Analysis on total N and Trophic Diatom Index from New Mexico rivers.

1. EXECUTIVE SUMMARY

Nutrient thresholds at which rapid changes in biological and ecological variation occur may represent a useful approach to setting numeric water quality criteria for nitrogen and phosphorus. However, quantifying these thresholds is difficult because the statistical techniques for these type of data are just gaining favor in the environmental and ecological sciences. We conducted threshold analyses on water quality data provided by the New Mexico Environment Department to quantify any potential threshold relationships between nutrient concentrations and biological responses in New Mexico surface waters (lakes, reservoirs, and rivers).

We conducted analyses on a variety of dataset that included: raw data, median data, and data restricted to summer months and found a relatively narrow range of total P thresholds that resulted in biological change (Table 1.1). A broader range of total N thresholds also explained the same variation in these variables. Results of this study can be used to guide the setting and adoption of numeric water quality standards for nutrients in the State of New Mexico.

Table 1.1. Total phosphorus and total nitrogen thresholds that at which biological shifts occurred in New Mexico Surface Waters.

Setting	Response Variable	Data Type	Total P threshold (mg/L)	Total N threshold (mg/L)
Lake	Secchi depth (m)	Median	0.035	0.27 - 0.55
Lake	Euphotic thickness (m)	Median	0.045	0.55
Lake	Chlorophyll-a (µg/L)	Median	0.045	0.93
Lake	Euphotic thickness (m)	All Data	0.044	--
Lake	Chlorophyll-a (µg/L)	All Data	--	1.54
Lake	% Cyanobacteria	All Data	0.045	0.84
Lake	% Cyanobacteria	Summer	0.045	0.73
River	Benthic Chlorophyll-a[1]	All Data	0.028	0.42
River	Trophic Diatom Index	All Data	0.020	0.34

[1]Units = µg cm^{-2}

2. INTRODUCTION

Developing numeric water quality criteria for nutrients is challenging because the environmental impact of accelerated eutrophication of surface waters is subject to value judgements (Hart et al. 1998, Dodds and Welch 2000, Chambers et al. 2008). Therefore, the US Environmental Protection Agency has recommended that states seek to quantify the relationships between nutrient concentrations and the associated biological responses in their surface waters that may result in failures to meet designated uses. However, the relationships between nutrients and biological responses are complicated and often involve non-linear and hierarchical structure (Clements et al. 2010, Dodds et al. 2010). Monitoring data generated by states is often insufficient for identifying these relationships, particularly when using traditional statistical techniques such as correlation analysis.

Changepoint and Regression Tree Analyses (De'Ath and Fabricius 2000) are proving to be useful methods for identifying environmental thresholds that result in ecological change (Scott et al. 2008). These techniques are now being applied to aid in the development of numeric water quality criteria for nutrients (Richardson et al. 2008, Chambers et al. 2011). Many states have sufficient data on both cause (nutrient concentrations) and effect (algal biomass, productivity, biodiversity, etc.) to conduct these types of analyses, but most have not utilized the techniques.

In this study, we analyzed water quality data generated by the New Mexico Environment Department in order to identify environmental thresholds at which biological or chemical data exhibit a clear change in either magnitude or variability. We used Changepoint Analysis, Regression Tree Analysis, and Categorical and Regression Tree Analysis to meet this objective. These techniques allowed us not only to identify environmental thresholds that resulted in ecological change, but also to identify water quality and ecological variation that was embedded within a hierarchical structure.

3. DATA REDUCTION AND OUTLIER ANALYSIS

3.1 SUMMARY

The first task in analytical support was to understand the format and composition of the data provided by the New Mexico Environment Department. Our strategy was to create descriptive statistics for these data and conduct an outlier analysis to determine how specific data could overly influence threshold analysis. Following our analyses, we flagged approximately 6% of data that are outliers. Future analyses will be conducted on data sets that include and exclude these data, unless the outliers are identified as erroneous data. Calculated medians that will be used in analyses on median data included the outlier values unless they were identified as erroneous data.

3.2 DESCRIPTIVE STATISTICS

We calculated mean and median values for the following parameters for each lake: secchi depth, specific conductance, euphotic zone thickness (z_{eu}), alkalinity, total suspended solids (TSS), total Kjeldahl nitrogen (TKN), nitrate-N, total phosphorus, total nitrogen, hardness, chloride, station depth, percent of DO profile below criterion, average DO near surface, average DO near bottom, and orthophosphate. Mean and median values sometimes represented multiple sites within a lake and/or multiple dates that a lake was sampled. In instances where only one observation existed for a lake/parameter combination, we allowed the single value to represent both the mean and median observation for that site. We also quantified the number of samples collected from each lake to inform our interpretation of future analyses. However, we did not quantify the number of observation for each particular parameter within each lake because the statistical analyses we plan on conducting are robust against missing data.

3.3 OUTLIER ANALYSIS

The following instances of outliers were found in the lakes and reservoirs database. Outlier analysis was not conducted on the rivers database because there were fewer data to work with and no data reduction (analysis of median data) was conducted for the river data. Outliers in lake and reservoir data were identified by calculating means and medians for each parameter for each lake/reservoir. Median values were plotted against mean values for visual inspection, and the mean values were divided by the median values to determine which observations were outliers according to the following: Any parameter for each lake where the mean/median was <0.5 or >2 were flagged as containing potential outliers. Of the 429 lake/parameter combinations (33 lakes with multiple observations x 13 parameters), 27 were shown to have possible outliers. The raw data for these lakes were checked. The following list describes the outliers that were identified and any action that was taken (in italics):

- Line 80 of excel file sorted by water name. Caballo Reservoir should be Cabresto Lake. This was indicated by outliers in Secchi Depth, TSS, Nitrate-N, and chlorophyll-a. *Spreadsheet was correct with red text showing change. Means and medians corrected.*
- Specific conductance values for Lake Avalon are highly variable. Appears as though observations from 7/18/1990 are incorrect. *Outlier flagged*.
- The TSS value for Sumner Reservoir on 4/22/1997 appears high. The value was 95 mg/L when no other values exceeded 30. Also, the high TSS did not correspond with particularly low transparency or Zeu. *Outlier flagged*.

- TSS in Ute Reservoir showed up as possibly having outliers. However, most of the TSS values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- TSS in Heron Reservoir showed up as possibly having outliers. However, most of the TSS values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- The TSS values for Lake Farmington on 7/17/2002 appears high. The value was 114 mg/L when no other values exceeded 5. *Outliers flagged*.
- TSS in Ramah Reservoir showed up as possibly having outliers. However, most of the TSS values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- TSS in Wall Lake showed up as possibly having outliers. However, most of the TSS values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- Nitrate-N in Elephant Butte Reservoir showed up as possibly having outliers. However, most of the nitrate-N values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- Nitrate-N in Brantley Reservoir showed up as possibly having outliers. However, most of the nitrate-N values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- Nitrate-N at the deep site in Lost Lake from 09/07/1988 appears high. *Outlier flagged*.
- Total P in Abiquiu Reservoir showed up as possibly having outliers. However, most of the total P values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- Total P at the Dam site in El Vado reservoir on 11/06/2007 appears very high. *Outlier flagged*.
- Total P at the Dam site in Heron reservoir on 09/27/2007 appears very high. *Outlier flagged*.
- Total P in Lake Maloya showed up as possibly having outliers. However, most of the total P values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- Total P in Avalon Reservoir showed up as possibly having outliers. However, all of the total P values for this lake were below the detection level and the few uncensored measurements do not appear unreasonably high. *No action taken.*
- Total P in Santa Cruz Lake showed up as possibly having outliers. However, all of the total P values for this lake were below the detection level. *No action taken.*
- Total P at Snow Lake on 10/23/1996 appears very high. *Outlier flagged*.

- Chloride at inflow site on Santa Rosa Reservoir from 10/30/2001 appears high. *Outlier flagged*.
- Chloride in Lost Lake showed up as possibly having outliers. However, most of the chloride values for this lake were below the detection level and the one uncensored measurements do not appear unreasonably high. *No action taken.*
- Chlorophyll-a in Maxwell Lake showed up as possibly having outliers. However, most of the chlorophyll values for this lake were near the detection level and the other measurement did not appear unreasonably high. *No action taken.* Are these possibly separate lakes?
- Chlorophyll-a in Canjilon Lake showed up as possibly having outliers. However, most of the chlorophyll values for this lake were near the detection level and the other measurement did not appear unreasonably high. *No action taken.* Are these possibly separate lakes?
- Chlorophyll-a in Charette Lake appeared high on 07/18/2006. *Outlier flagged*.
- Chlorophyll-a in Lake Roberts appeared high on 07/16/1996. *Outlier flagged*.

4. ANALYSIS OF LAKE AND RESERVOIR DATA

4.1 SUMMARY

We used Changepoint and Regression Tree Analyses on environmental and biological data from New Mexico lakes and reservoirs to identify quantitative thresholds in nutrient concentrations that were correlated with common biological response variables. Thresholds total P that explained variability in secchi depth, Z_{eu}, and chlorophyll-a ranged from 0.035 to 0.051 mg/L P. Thresholds in total N concentration were much more variable, ranging from 0.27 to 1.54 mg/L N. Total P or total N concentrations were always the best predictors of these biological response variables all but one analysis. The thresholds reported from the Changepoint and Regression Tree Analyses in this study provide quantitative evidence for the link between nutrient concentrations and commonly measured biological response data in New Mexico lakes and reservoirs.

4.2 OBJECTIVES AND METHODS

In order to develop numeric nutrient criteria for surface water quality standards, there is a need to develop quantitative links between nutrient concentrations and biological responses. However, most states have limited data on both nutrient concentrations and associated biological responses. Inadequate data inhibits the effectiveness of most traditional statistical techniques. But newer methods, such as Changepoint Analysis and Regression Tree Analysis, are providing scientists,

regulators, and policy makers the capacity to extract useful information from relatively limited data sets. The objective of this was to quantify thresholds in nutrient concentrations (total nitrogen and total phosphorus) at which biological changes may occur in lakes and reservoirs in the State of New Mexico.

We obtained a lake and reservoir water quality database from the New Mexico Environment Department and screened the data for outliers and potentially erroneous data. Data from individual lakes were combined into a single median value for each parameter to reduce the uncertainty associated with individual sampling events and create the most reasonable source of information to generalize patterns amongst lakes. We used Changepoint Analysis to evaluate thresholds in median total P or median total N across all lakes and reservoirs which were correlated with measurable changes in either secchi depth, euphotic zone thickness (Z_{eu}), and chlorophyll-a concentration. We also analyzed thresholds in median and raw data by combining potential cause (total N and P), effect (secchi depth, Z_{eu}, chlorophyll-a), and covariate (conducitivity, alkalinity, hardness, total suspended solids, chloride, and water depth) data into Regression Tree Analysis. These analyses were intended to identify thresholds in nutrient concentrations that resulted in biological responses. However, the regression tree models were also intended to indentify thresholds in covariates that could be masking relationships between nutrients and biological responses.

We also utilized the data on cyanobacteria (blue-green algae) to evaluate nutrient thresholds. Cyanobacteria are a group of phytoplankton that generally represent a higher proportion of biomass under nutrient-rich conditions. There were a substantial number of samples from the NMED lakes and reservoirs database that also had information on the percent abundance of cyanobacteria in the phytoplankton. We used Changepoint Analysis to determine if cyanobacterial abundance was related to algal biomass (measured as secchi transparency and chlorophyll-a). We conducted Changepoint Analysis to evaluate thresholds in total P and total N that were correlated with increases in percent cyanobacteria. We also conducted a Regression Tree Analysis on the percent cyanobacteria data using predictors described above. Cyanobacteria are also known to grow better during summer when water temperature is high. Therefore, we also extracted data on percent cyanobacteria and total N and P concentrations that were collected between June and August for Changepoint Analyses on these data only.

Changepoint and Regression Tree analyses are conceptually similar in that both seek to identify the a threshold in an independent variable that explains some change in variability in the dependent variable. Regression Tree Analysis is simply an extension of changepoint in that multiple independent variables can be used to build a Regression Tree model. In Regression Tree Analysis data are partitioned into increasingly homogeneous subsets based changepoints identified for multiple independent variables. The process is repeated in an iterative fashion which results in a tree-like

distribution of predictors (independent variables) associated with the dependent variable.

Changepoint and Regression Tree Analyses were conducted using the R (2.8.1) software. Both analyses in R use recursive partitioning the divide data into homogeneous subsets based on numeric predictors. A threshold relationship between the independent and dependent variable(s) is identified through deviance reduction. In other words, all possible thresholds in independent variable are evaluated based on some initial parameterization and the threshold that results in the greatest reduction in deviance is selected. Initial parameterization involves identifying the minimum number of observations that should be on either side of the threshold, and the intensity of statistical resampling that will occur to validate the model. We required that a minimum of five observations occur on either side of any threshold and that the model size be evaluated with ten cross-validations.

4.3. RESULTS

Changepoints in total phosphorus concentration – Median total P concentrations in New Mexico lakes and reservoirs were correlated with median secchi depth, median euphotic thickness and median chlorophyll-a concentration. Variability in median secchi depth across all lakes and reservoirs was greatest when median total P concentrations were less than 0.035 mg/L (Figure 4.1). Median secchi depth ranged from 0.5 to 16 m when median total P values were less than 0.035 mg/L and from 0.1 to 4 m when median total P values exceeded 0.035 mg/L. The median total P changepoint explained 14% of the variability in median secchi depth across all lakes and reservoirs. Median secchi depth was less than 2.45 m in 51 New Mexico lakes and reservoirs, which is often cited as the boundary between mesotrophic and eutrophic lakes (OECD 1982).

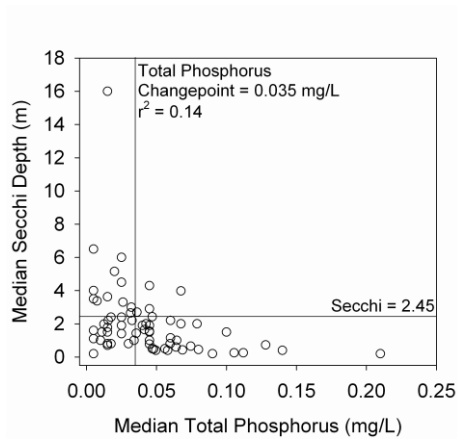


Figure 4.1. Results of Changepoint Analysis on median total P and median secchi depth values for all New Mexico Lakes and Reservoirs.

Variability median euphotic thickness (Z_{eu}) was more evenly distributed among reservoirs than variability in median secchi depth. Median Z_{eu} across all lakes and reservoirs was greatest when median total P concentrations were less than 0.045 mg/L (Figure 4.2). Median Z_{eu} ranged from approximately 0.5 to 27 m when median total P values were less than 0.045 mg/L and from 0.1 to 7 m when median total P values exceeded 0.045 mg/L. The median total P changepoint explained 22% of the variability in median Z_{eu} across all lakes and reservoirs.

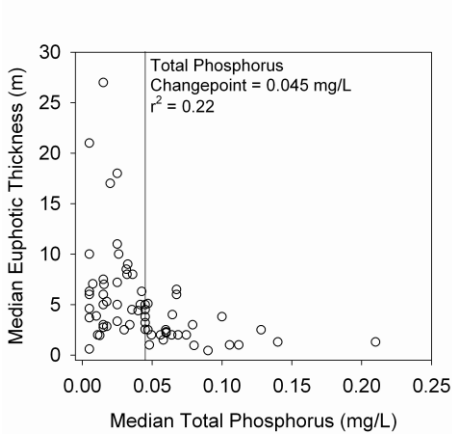


Figure 4.2. Results of Changepoint Analysis on median total P and median Zeu values for all New Mexico Lakes and Reservoirs.

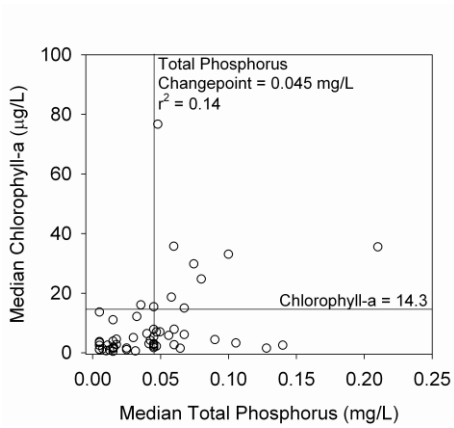


Figure 4.3. Results of Changepoint Analysis on median total P and median chlorophyll-a concentrations for all New Mexico Lakes and Reservoirs.

Median chlorophyll-a concentrations among lakes was highly variable, but was generally least when median total P concentrations were less than 0.045 mg/L and greatest when median total P exceeded this threshold (Figure 4.3). Median chlorophyll-a ranged from approximately 1 to 15 µg/L when median total P values were less than 0.045 mg/L and from 4 to 75 µg/L when median total P values exceeded 0.045 mg/L.

The median total P changepoint explained 14% of the variability in median chlorophyll-a concentrations across all lakes and reservoirs. Median chlorophyll-a concentrations were greater than 14.3 µg/L in 10 New Mexico lakes and reservoirs, which is often cited as the boundary between mesotrophic and eutrophic lakes (OECD 1982). Overall, the observed changepoints in median total P concentrations across all New Mexico lakes and reservoirs were 0.035 – 0.045 mg/L. The relationship between median total P and median Z_{eu} represented the strongest changepoint in total P that resulted in a measureable biological change.

Changepoints in total nitrogen concentration – Median total N concentrations in New Mexico lakes and reservoirs were correlated with median secchi depth, median euphotic thickness and median chlorophyll-a concentration. Median secchi depth was greatest when median total N concentrations were less than 0.27 mg/L (Figure 4.4). A second changepoint was also identified where median secchi depth was least when total N exceeded 0.55 mg/L (Figure 4.4). These two changepoints in total N explained 27% of the variation in median secchi depth across all New Mexico lakes and reservoirs. Median Z_{eu} across all lakes and reservoirs was greatest when median total N concentrations were less than 0.55 mg/L (Figure 4.5). The median total N changepoint explained 14% of the variability in median Z_{eu} across all lakes and reservoirs.

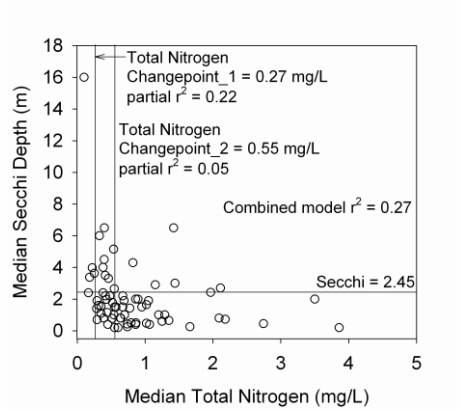


Figure 4.4. Results of Changepoint Analysis on median total N and median secchi depth values for all New Mexico Lakes and Reservoirs.

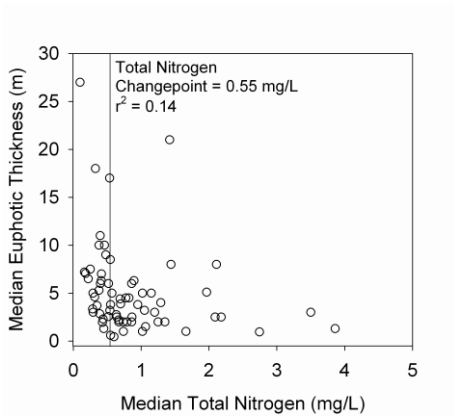


Figure 4.5. Results of Changepoint Analysis on median total N and median Zeu values for all New Mexico Lakes and Reservoirs.

Median chlorophyll-a concentrations were generally least when median total N concentrations were less than 0.93 mg/L and greatest when median total P exceeded this threshold (Figure 4.6). Median chlorophyll-a ranged from approximately 1 to 15 µg/L when median total N values were less than 0.93 mg/L and from 1 to 75 µg/L when median total N values exceeded 0.93 mg/L. The median total P changepoint explained 26% of the variability in median chlorophyll-a concentrations across all lakes and reservoirs. Overall, the observed changepoints in median total N concentrations across all New Mexico lakes and reservoirs were highly variable, ranging from as low as 0.27 mg/L when explaining the greatest amount of variation in sechhi depth to 0.93 mg/L when explaining variation in chlorophyll-a concentration. The relationship between median total N and median Z_{eu} represented the strongest changepoint in total N that resulted in a measureable biological change.

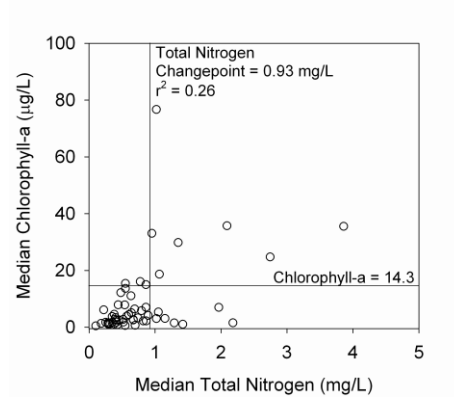


Figure 4.6. Results of Changepoint Analysis on median total P and median chlorophyll-a concentrations for all New Mexico Lakes and Reservoirs.

Regression tree models – Regression tree models on median data only revealed relationships between total nutrients and the measured responses and were therefore identical to the changepoint analysis results. These models are not shown in the results section but output was provided in Appendix I. A two branch model explained the greatest amount of variation in secchi depth raw data across all New Mexico lakes and reservoirs (Figure 4.7). Mean secchi depth was greatest in five lakes and reservoirs exhibiting the least water hardness. When these lakes and reservoirs were split out, mean secchi depth across the remainder of waterbodies was only 0.88 m when total P concentrations exceeded 0.051 mg/L. Mean secchi depth across these lakes was 1.95 m when total P concentrations were less than this threshold. In order to confirm the small number of soft-water lakes did not have too large an effect

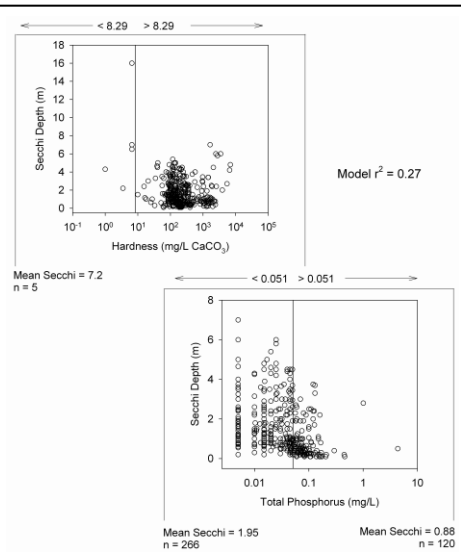


Figure 4.7. Regression Tree Analysis of secchi depth across all New Mexico lakes and reservoirs.

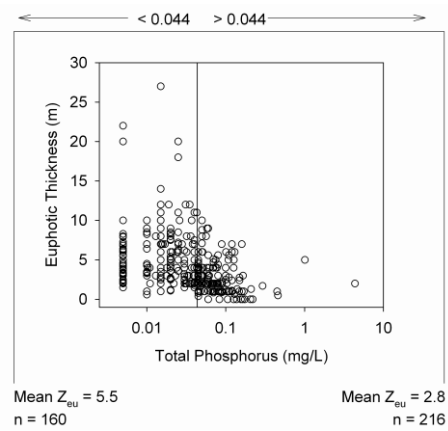


Figure 4.8. Regression Tree Analysis of Zeu across all New Mexico lakes and reservoirs.

on secchi depth, ran the same Regression Tree Analysis but excluded hardness as an independent variable. Indeed, the total P threshold of 0.051 (r^2 = 0.13) was found to be the only branch in the subsequent model. A number of other variables did explain some variation in raw secchi depth. Those variables in order of importance were: depth (r^2 = 0.10), total suspended solids (r^2 = 0.09), total N (r^2 = 0.07), and conductivity (r^2 = 0.05).

Although full regression tree models were attempted for predicting both Z_{eu} and chlorophyll-a concentrations, total nutrient concentrations were the only statistically validpredictors of these variables. Total P concentration was the best predictor Z_{eu} (r^2 = 0.16) and made up the only predictor in the cross-validated regression tree model (Figure 4.8). Mean Z_{eu} was 5.5 m when total P concentrations were less than 0.044 but only 2.8 m when total P concentrations exceeded this threshold (Figure 4.8).

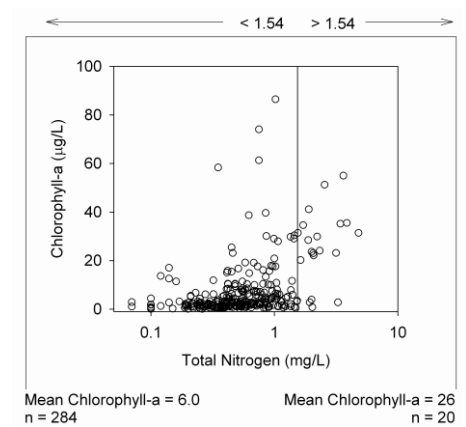


Figure 4.9. Regression Tree Analysis of chlorophyll-a concentration across all New Mexico lakes and reservoirs.

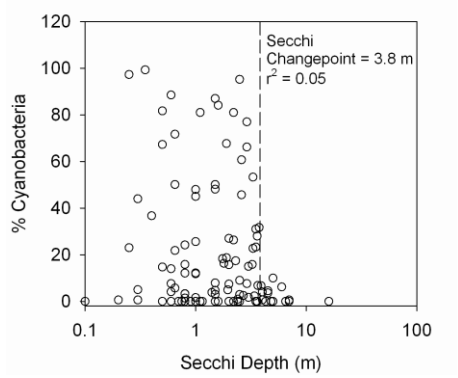


Figure 4.10. Changepoint analysis of secchi depth versus percent cyanobacteria in New Mexico lakes and reservoirs.

Chlorophyll-a concentrations were best predicted by total N concentrations (r^2 = 0.17) in New Mexico lakes and reservoirs (Figure 4.9). Mean chlorophyll-a concentrations were 6.0 µg/L when total N concentration was less than 1.54 mg/L, but mean chlorophyll-a concentrations were 26 µg/L when total N exceeded this threshold.

Although the regression tree models for Z_{eu} and chlorophyll-a concentrations showed only one important predictor, some other variables could have substituted for total P and total N in these models with reduced strength of association. For

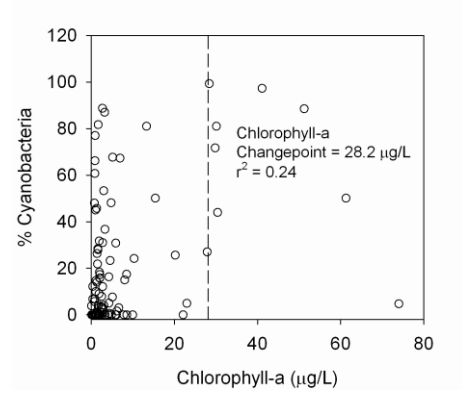


Figure 4.11. Changepoint analysis of chlorophyll-a versus percent cyanobacteria in New Mexico lakes and reservoirs.

example, Z_{eu} could have also been predicted with thresholds in depth (12.5 m; r^2 = 0.12), hardness (2445 mg/L $CaCO_3$; r^2 = 0.09), total suspended solids (10.5 mg/L; r^2 = 0.09), and total nitrogen (0.41 mg/L; r^2 = 0.07). Chlorophyll-a concentrations could have also been predicted with thresholds in total suspended solids (r^2 = 0.17), hardness (r^2 = 0.17), chloride (r^2 = 0.17), and conductivity (r^2 = 0.17).

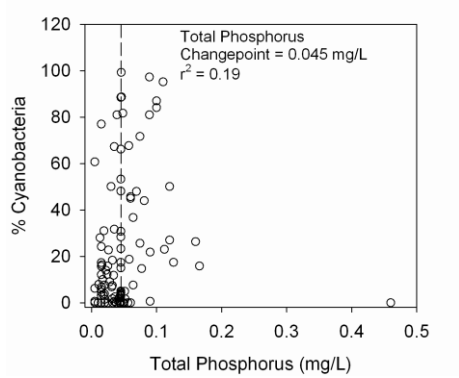


Figure 4.12. Changepoint analysis of total phosphorus versus percent cyanobacteria in New Mexico lakes and reservoirs.

Cyanobacteria and Water Quality – The percentage of phytoplankton biomass comprised of cyanobacteria was greater in New Mexico lakes and reservoirs that were experiencing some level of eutrophication. Cyanobacteria made up only 3% of phytoplankton biomass when Secchi depth was greater than 3.8m, but was more than 22% of phytoplankton biomass when Secchi depth was less than 3.8m (Figure 4.10). Similarly, cyanobacteria made up only 16% (range 0 - 85%) of phytoplankton biomass when chlorophyll-a was less than 28 µg/L, but comprised 67% (range 15 - 100%) when chlorophyll-a was greater than 28 µg/L (Figure 4.11).

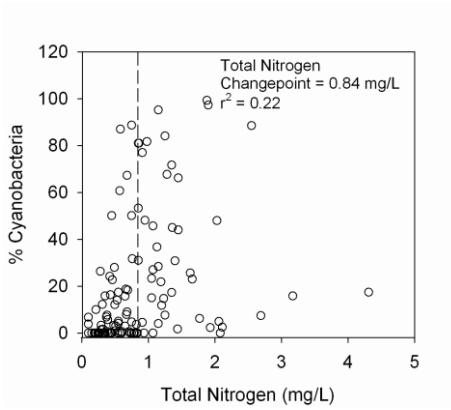


Figure 4.13. Changepoint analysis of total nitrogen versus percent cyanobacteria in New Mexico lakes and reservoirs.

The percent cyanobacteria in New Mexico lakes and reservoirs was also related to nutrient concentrations. Percent cyanobacteria increased from 12% to 39% when total P concentrations went above 0.045 mg/L (Figure 4.12). A total N threshold of 0.84 mg/L was separated cyanobacterial abundances of 10% (low TN) and 37% (high TN), respectively (Figure 4.13). The relationship between total P and the abundance of cyanobacteria in the phytoplankton was even stronger when data were limited to summer months, but resulted in an identical threshold value (0.045 mg/L) (Figure 4.14). The relationship between total N and percent cyanobacteria did not improve when the analysis was limited to data from the summer months, but the total N threshold did decrease slightly (Figure 4.15).

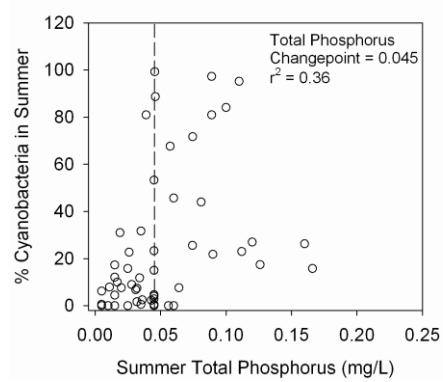


Figure 4.14. Changepoint analysis of summer total phosphorus versus percent cyanobacteria in summer in New Mexico lakes and reservoirs.

Analysis details and conclusions – The R code and model output for each of the modeling scenarios described above are contained in Appendix I of this document. As discussed above, these models showed that no hierarchy existed in the environmental parameters predicting most of the biological responses measured for New Mexico lakes and reservoirs, with one exception. Hierarchical structure was found for secchi depth where water hardness resulted in the exclusion of five lakes with extremely transparent water. However, the resulting total P threshold from this model was very similar in range to total P thresholds identified in other regression tree and changepoint models.

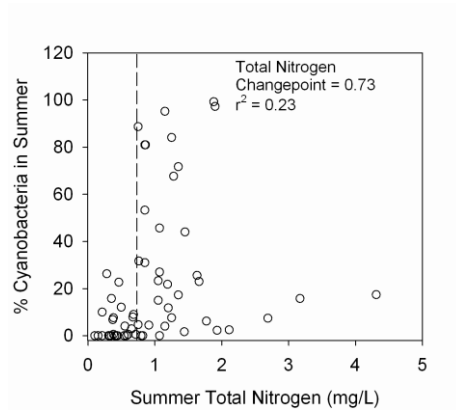


Figure 4.15. Changepoint analysis of summer total nitrogen versus percent cyanobacteria in summer in New Mexico lakes and reservoirs.

Thresholds occurred in the total P and total N data that explained variability in secchi depth, Z_{eu}, chlorophyll-a concentrations, and the percent of cyanobacteria in the phytoplankton across all New Mexico Lakes and reservoirs. Thresholds in total P ranged from 0.035 to 0.051 mg/L P. Thresholds in total N concentration were much more variable, ranging from 0.27 to 1.54 mg/L N. Either total P or total N concentrations were always the best predictors of the biological response variables, except for secchi depth in the 5 most soft water lakes in the state. The thresholds reported from the Changepoint and Regression Tree Analyses in this study provide quantitative evidence for the link between nutrient concentrations and commonly measured biological response data in New Mexico lakes and reservoirs.

5. ANALYSIS OF RIVER DATA

5.1 SUMMARY

We used Changepoint and Categorical and Regression Tree Analysis (CART) on environmental and biological data from New Mexico rivers to identify quantitative thresholds in nutrient concentrations that were correlated with common biological response variables. A threshold in total P was useful for predicting benthic chlorophyll-a concentrations and the Trophic Diatom Index. The total N threshold explained much less of the variation in benthic chlorophyll-a and the Trophic Diatom Index across rivers. A CART model that included total nutrient concentrations as well as temperature, pH, electrical conductivity, dissolved oxygen, total dissolved solids, turbidity, total suspended solids, alkalinity and substrate type (categorical variable) confirmed that total P was the strongest predictor of benthic chlorophyll-a in the database. However, the

relationship between total P and benthic chlorophyll-a that was observed in both the changepoint and CART analyses was counterintuitive; benthic chlorophyll-a decreased as total phosphorus concentrations in river water increased. This pattern could be an artifact of the sampling regime if diatom biofilms were specifically targeted and filamentous green algae were avoided. Or, the pattern may be spurious and underlie some other control on benthic algae for which no data were collected (such as current velocity or riparian cover). If the periphyton data does represent biofilm-type sampling, then the quantitative threshold observed in this study may be very useful for predicting the total P concentrations at which New Mexico streams shift from diatom to green algae-dominated.

5.2 OBJECTIVES AND METHODS

In order to develop numeric nutrient criteria for surface water quality standards, there is a need to develop quantitative links between nutrient concentrations and biological responses. However, most states have limited data on both nutrient concentrations and associated biological responses. Inadequate data inhibits the effectiveness of most traditional statistical techniques. But newer methods, such as Changepoint Analysis and Categorical and Regression Tree Analysis (CART), are providing scientists, regulators, and policy makers the capacity to extract useful information from relatively limited data sets. The objective of this study was to quantify thresholds in nutrient concentrations (total nitrogen and total phosphorus) at which biological changes may occur in rivers in the State of New Mexico.

We obtained a water quality database from the New Mexico Environment Department and screened the data for outliers and potentially erroneous data. We explored relationships between nutrients (total N and total P) and benthic chlorophyll-a concentrations and the Trophic Diatom Index (Kelly 1998) in order to quantify these potential cause-effect relationships. Due to the relatively small amount of data ($n < 150$; benthic chlorophyll-a $n = 67$), we did not reduce data to median or mean values for these analyses and instead used raw data in both changepoint and CART models. We used Changepoint Analysis to evaluate thresholds in total P concentration or total N which were correlated with measurable changes in chlorophyll-a and Trophic Diatom Index concentration. We also analyzed thresholds in raw data by combining potential cause (total N and P), effect (benthic chlorophyll-a), and covariate (temperature, pH, electrical conductivity, dissolved oxygen, total dissolved solids, turbidity, total suspended solids, alkalinity and substrate type) data into a CART analysis. The CART analysis was intended to identify thresholds in nutrient concentrations that resulted in biological responses. However, the CART model was also intended to indentify thresholds in covariates that could be masking relationships between nutrients and

biological responses. Changepoint and CART Analyses were conducted using the R (2.8.1) software as described in section 4 of this document.

5.3 RESULTS

Changepoint in total phosphorus concentration – Benthic chlorophyll-a concentration, which is a surrogate for benthic algal biomass, was greatest (mean = 16 µg cm^{-2}) when total P concentrations were less than 0.028 mg/L and least (mean = 4.4 µg cm^{-2}) when total P concentrations were above this threshold (Figure 5.1). The total P threshold explained 41% of the variation in benthic chlorophyll-a among samples. The Trophic Diatom Index tended to increase with increasing total P concentration as it is intended to function. The Trophic Diatom Index ~50 when total P concentrations were less than 0.02 µg/L and ~70 when total P was above this threshold (Figure 5.2). Total P explained 12% of the variation in Tropic Diatom Index across all samples.

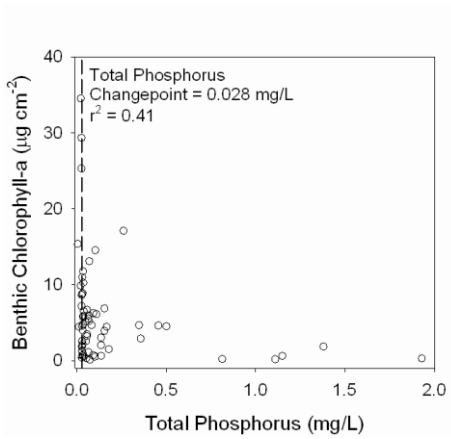


Figure 5.1. Results of Changepoint Analysis on total P and benthic chlorophyll-a from New Mexico rivers.

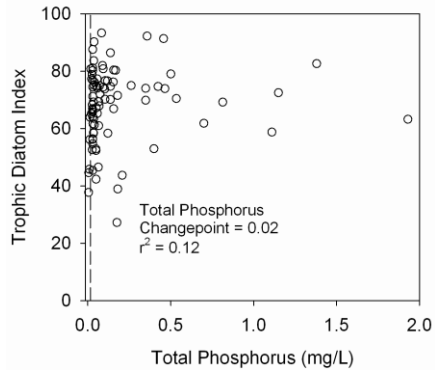


Figure 5.2. Results of Changepoint Analysis on total P and Trophic Diatom Index from New Mexico rivers.

This pattern of decreasing algal biomass with increasing total P is counterintuitive and raises questions regarding how algal biomass was sampled. If diatom biofilms were specifically targeted in the sampling regime, this threshold may represent an important changepoint in New Mexico rivers where diatom biofilms are replace by some other dominant algal growth pattern that were not sampled (such as green algae). However, if sampling was not targeting diatom biofilms specifically (i.e. samples were collected completely randomly or under some systematic randomization) then the relationship between total P and benthic chlorophyll-a is likely spurious.

The response of the Trophic Diatom Index to increasing nutrient concentrations was intuitive. The index is designed to increase in magnitude between 0-100 as the trophic state of a stream or river increases (Kelly 1998). The mechanism for this is the change in diatom taxonomic composition that is reflective of a stream or river which is either organic matter rich (TDI > 50) or poor (TDI < 50). The assumption is that the increase in organic matter is related to nutrient availability in streams and therefore is a biological indicator of changing environmental conditions.

Changepoints in total nitrogen concentration – Benthic chlorophyll-a concentration was greatest (mean = 8.2 µg cm^{-2}) when total N concentrations were less than 0.42 mg/L and least (mean = 4.4 µg cm^{-2}) when total N concentrations were above this threshold (Figure 5.3). The total N threshold explained only 0.08% of the variation in benthic chlorophyll-a among samples. The Trophic Diatom Index also increased with increasing total N, but the relationship was also weak (Figure 5.4).

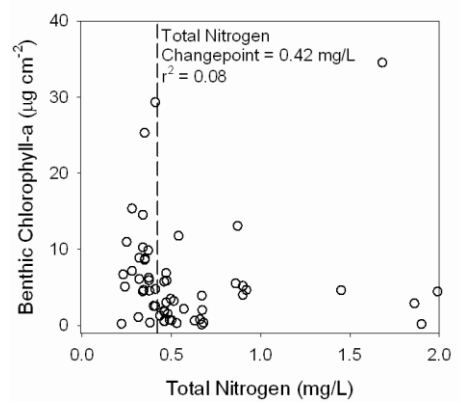


Figure 5.3. Results of Changepoint Analysis on total N and benthic chlorophyll-a from New Mexico rivers.

Similar to the pattern with total P, the pattern of decreasing algal biomass with increasing total N is counterintuitive. The pattern may be defensible is diatom biofilms were specifically targeted in the sampling regime. However, if sampling was not targeting diatom biofilms specifically (i.e. samples were collected completely randomly or under some systematic randomization) then the relationship between total N and benthic chlorophyll-a is also likely spurious. It is suspicious, however, that the counterintuitive relationship was observed for both total N and total P. This pattern warrants further investigation into methods and sampling design.

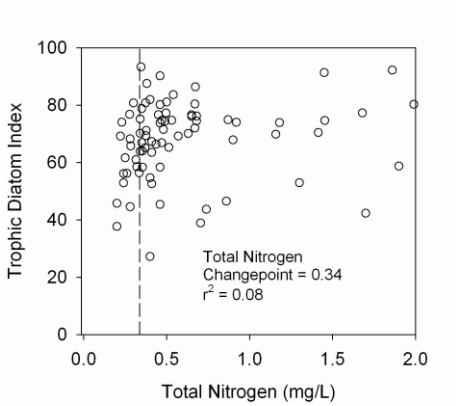


Figure 5.4. Results of Changepoint Analysis on total N and Trophic Diatom Index from New Mexico rivers.

Categorical and Regression Tree Model – Due to the relatively small amount of data in the analysis (n = 67 samples that had nutrients and benthic chlorophyll-a), only one split was identified as statistically viable in the CART model. Total P was the strongest predictor of benthic chlorophyll-a, followed by temperature, dissolved oxygen, turbidity and total N (Table 5.1). Because there was only one viable split in the model, the thresholds for total P and total N at which benthic chlorophyll-a experienced at major shift were identical to the thresholds identified in the changepoint analysis. Similarly, the pattern of correlation was counterintuitive, as previously mentioned.

Table 5.1. Results of Categorical and Regression Tree (CART) analysis on benthic chlorophyll-a. Primary split predictors are listed in order of prediction strength. Only one split was statistically valid in the tree.

Predictor Variable	Threshold	partial r^2
Total phosphorus	0.028 mg/L	0.41
Temperature	8.8 ºC	0.19
Dissolved oxygen	11.0 mg/L	0.19
Turbidity	76 NTU	0.08
Total nitrogen	0.42 mg/L	0.08

Analysis details and conclusions – The R code and model output for each of the modeling scenarios described above are contained in Appendix II to this document. As discussed above, no hierarchy existed in the relationships between environmental variables and the biological responses to nutrient enrichment with the available data.

Thresholds occurred in the total P and total N data that explained variability in benthic chlorophyll-a concentrations and the Trophic Diatom Index across all New Mexico rivers. The thresholds in total P (0.02 - 0.03 mg/L) were by far the strongest predictor of benthic chlorophyll-a (r^2 = 0.41), and Trophic Diatom Index (r^2 = 0.41). Total N thresholds were also useful for predicting variation in benthic chlorophyll-a (0.42 mg/L TN) and the Trophic Diatom Index (0.34 mg/L TN), but these relationships were much weaker and other environmental variables were stronger predictors in the CART model (Table 1). Interestingly, the relationships between total P and chlorophyll-a, and total N and chlorophyll-a, were counterintuitive. Benthic chlorophyll-a concentrations (µg cm^{-2}) decreased as both total P and total N increased. In other words, algal biomass actually decreased as nutrient concentrations increased. This response was unexpected and may be explained in one of two ways:

1. The relationships between benthic chlorophyll-a and total P and N could be spurious. Instead, benthic chlorophyll-a could be controlled by another unmeasured parameter such as river canopy cover or current velocity. Total N and P concentrations may also be correlated to this unmeasured parameter, which could create the spurious correlation observed here.
2. The relationship between benthic chlorophyll-a and total P and N is not spurious, but related to the sampling methodology where epilithic biofilms were specifically targeted and filamentous algae was intentionally avoided.

Because benthic chlorophyll-a was negatively correlated with both total N and total P, the spurious correlation scenario seems less likely. Furthermore, the strength of the relationship between benthic chlorophyll-a and total P suggests a direct connection. The second explanation seems more plausible but would only be valid if large filamentous algae occurring in more nutrient rich streams were intentionally avoided. This is often the design of some sampling schemes which are targeting diatoms specifically for bioassessment methods. If this assumption proves correct, then the total P threshold observed in this analysis may represent a critical P level above which the biomass of diatoms in epilithic biofilms rapidly decreases as the amount of filamentous algae in streams increases.

6. CONCLUSIONS

Changepoint Analysis, Regression Tree Analysis, and Categorical and Regression Tree Analysis are all useful techniques for identifying thresholds in environmental variables that result in biological or ecological changes in surface waters. These techniques allow scientists and regulators to extract more information from water quality databases that have large amounts of data collected over large spatial and temporal scales. We used these techniques to explore patterns between nutrient concentrations, biological responses, and covariates in the New Mexico Environment Department water quality database.

We found that thresholds total P and total N were often related to a number of biological responses in New Mexico surface waters. The nutrient thresholds at which biological changes were observed are summarized in Table 6.1.

Table 6.1. Total phosphorus and total nitrogen thresholds that at which biological shifts occurred in New Mexico Surface Waters.

Setting	Response Variable	Data Type	Total P threshold (mg/L)	Total P prediction strength (r^2)	Total N threshold (mg/L)	Total N prediction strength (r^2)
Lake	Secchi depth (m)	Median	0.035	0.14	0.27 - 0.55	0.27 (model)
Lake	Euphotic thickness (m)	Median	0.045	0.22	0.55	0.14
Lake	Chlorophyll-a (µg/L)	Median	0.045	0.14	0.93	0.26
Lake	Euphotic thickness (m)	All Data	0.044	0.16	--	--
Lake	Chlorophyll-a (µg/L)	All Data	--	--	1.54	0.17
Lake	% Cyanobacteria	All Data	0.045	0.19	0.84	0.22
Lake	% Cyanobacteria	Summer	0.045	0.36	0.73	0.23
Lake	Secchi depth (m)	WWAL	0.038	0.21	--	--
Lake	Euphotic thickness (m)	WWAL	0.038	0.21	--	--
Lake	Chlorophyll-a (µg/L)	WWAL	--	--	1.41	0.29
Lake	% Cyanobacteria	WWAL	0.046	0.19	--	--
Lake	Secchi depth (m)	CWAL	0.046	0.08	0.22	0.07
Lake	Euphotic thickness (m)	CWAL	0.044	0.14	0.39	0.09
Lake	Chlorophyll-a (µg/L)	CWAL	0.025	0.09	2.1	0.14
Lake	% Cyanobacteria	CWAL	0.043	0.17	0.84	0.31
River	Benthic Chlorophyll-a[1]	All Data	0.028	0.41	0.42	0.08
River	Trophic Diatom Index	All Data	0.020	0.12	0.34	0.08

[1]Units = µg cm^{-2}

7. REFERENCES

Chambers, P.A., C. Vis, R.B. Brua, M. Guy, J.M. Culp, and G.A. Benoy. 2008. Eutrophication of agricultural streams: Defining nutrient concentrations to protect ecological condition. Water Science and Technology 58:2203–2210.

Chambers, P.A., D.J. McGoldrick, R.B. Brura, C. Vis, J.M. Culp, and G.A. Benoy. 2011. Development of environmental thresholds for nitrogen and phosphorus in streams. Journal of Environmental Quality doi:10.2134/jeq2010.0273. Posted online 8 June 2011.

Clements, W.H., N.K.M. Vieira, and D.L. Sonderegger. 2010. Use of thresholds to assess ecological recovery in lotic ecosystems. Journal of the North American Benthological Society 29:1017-1023.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Dodds, W.K., and E.B. Welch. 2000. Establishing nutrient criteria in streams. Journal of the North American Benthological Society 19:186–196.

Dodds, W.K., W.H. Clements, K. Gido, R.H. Hilderbrand, and R.S. King. 2010. Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. Journal of the North American Benthological Society 29:988-997.

Hart, B.T., B. Maher, and I. Lawrence. 1999. New generation water quality guidelines for ecosystem protection. Freshwater Biology 41:347–359.

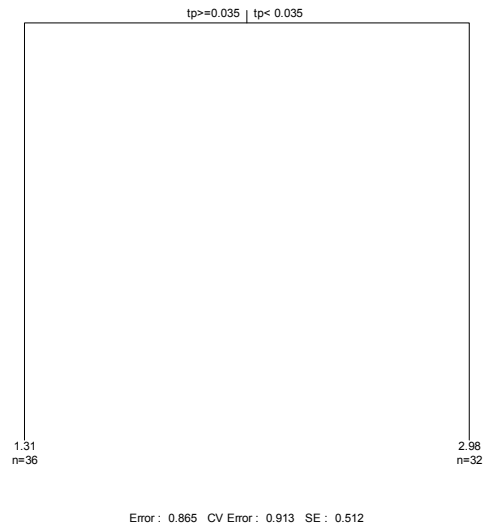
Kelly, M.G. 1998. Use of the trophic diatom index to monitor eutrophication in rivers. Water Research 32:236-242.

Richardson, C.J., R.S. King, S.S. Qian, P. Vaithiyanathan, R.G. Qualls, and C.A Stow. 2008. Estimating ecological thresholds for phosphorus in the Everglades. Environmental Science and Technology 41:8084–8091.

Scott, J.T., R.D. Doyle, S.J. Prochnow, and J.D. White. 2008. Are watershed and lacustrine controls on planktonic N fixation hierarchically structured? Ecological Applications 18:805-819.

APPENDIX I – R CODE & OUTPUT FOR LAKE AND RESERVOIR DATA

TP versus Secchi depth:



Code: mvpart(form = secchi ~ tp, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=68 (1 observation deleted due to missing data)

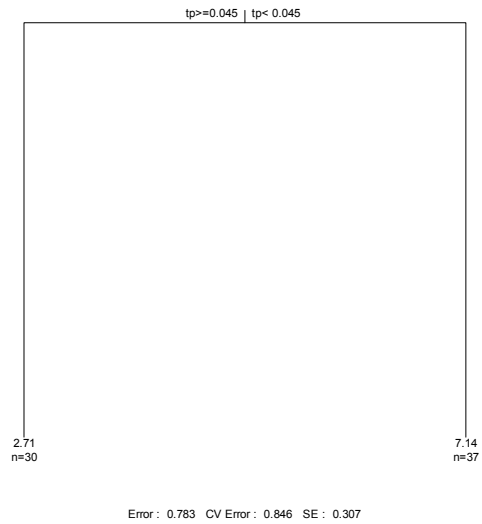
	CP	nsplit	rel error	xerror	xstd
1	0.13513033	0	1.0000000	1.0307395	0.5645933
2	0.01603747	1	0.8648697	0.9131482	0.5122514

Node number 1: 68 observations, complexity param=0.1351303
  mean=2.095221, MSE=5.153133
  left son=2 (36 obs) right son=3 (32 obs)
  Primary splits:
    tp < 0.035 to the right, improve=0.1351303, (0 missing)

Node number 2: 36 observations
  mean=1.308472, MSE=1.046637

Node number 3: 32 observations
  mean=2.980313, MSE=8.293209

TP versus euphotic depth:



Code: mvpart(form = zeu ~ tp, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=67 (2 observations deleted due to missing data)

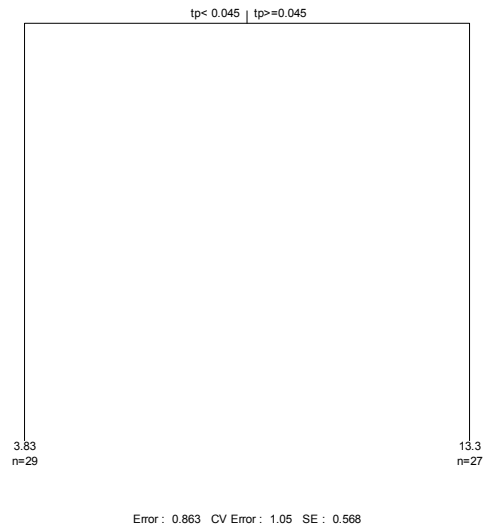
	CP	nsplit	rel error	xerror	xstd
1	0.21680710	0	1.0000000	1.0323398	0.3825372
2	0.01343076	1	0.7831929	0.8461148	0.3068462

Node number 1: 67 observations, complexity param=0.2168071
  mean=5.154925, MSE=22.44666
  left son=2 (30 obs) right son=3 (37 obs)
  Primary splits:
    tp < 0.045 to the right, improve=0.2168071, (0 missing)

Node number 2: 30 observations
  mean=2.705, MSE=2.315892

Node number 3: 37 observations
  mean=7.141351, MSE=29.95642

TP versus Chlorophyll-a:



Code: mvpart(form = chla ~ tp, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=56 (13 observations deleted due to missing data)

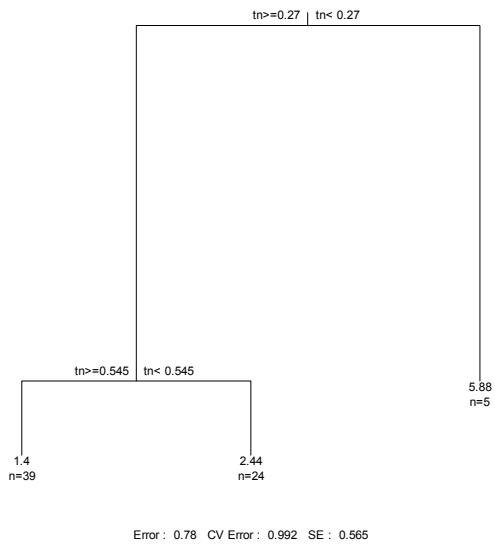
	CP	nsplit	rel error	xerror	xstd
1	0.13673045	0	1.0000000	1.025520	0.5259262
2	0.01641972	1	0.8632695	1.054781	0.5676047

Node number 1: 56 observations, complexity param=0.1367305
mean=8.393571, MSE=163.8208
left son=2 (29 obs) right son=3 (27 obs)
Primary splits:
tp < 0.045 to the left, improve=0.1367305, (0 missing)

Node number 2: 29 observations
mean=3.826897, MSE=16.62168

Node number 3: 27 observations
mean=13.29852, MSE=275.4659

TN versus Secchi depth:



Code: mvpart(form = secchi ~ tn, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=68 (1 observation deleted due to missingness)

	CP	nsplit	rel error	xerror	xstd
1	0.22026752	0	1.0000000	1.0441716	0.5692958
2	0.04589827	1	0.7797325	0.9537172	0.4687342

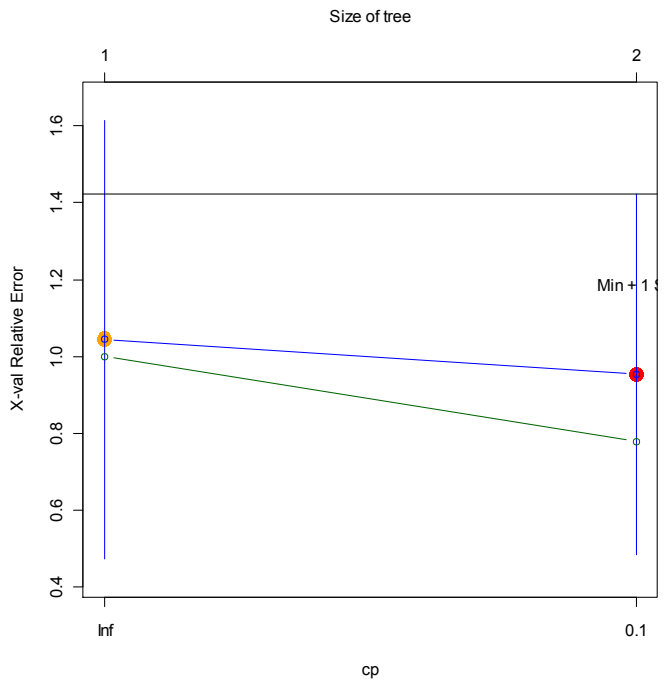
Node number 1: 68 observations, complexity param=0.2202675
mean=2.095221, MSE=5.153133
left son=2 (63 obs) right son=3 (5 obs)
Primary splits:
tn < 0.27 to the right, improve=0.2202675, (0 missing)

Node number 2: 63 observations, complexity param=0.04589827
mean=1.795079, MSE=2.281939
left son=4 (39 obs) right son=5 (24 obs)
Primary splits:
tn < 0.545 to the right, improve=0.1118747, (0 missing)

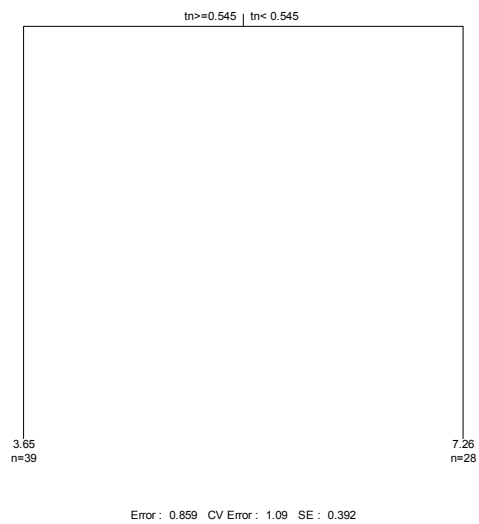
Node number 3: 5 observations
mean=5.877, MSE=25.89326

Node number 4: 39 observations
  mean=1.398718, MSE=1.558729

Node number 5: 24 observations
  mean=2.439167, MSE=2.787016



TN versus euphotic depth:



Code: mvpart(form = zeu ~ tn, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=67 (2 observations deleted due to missingness)

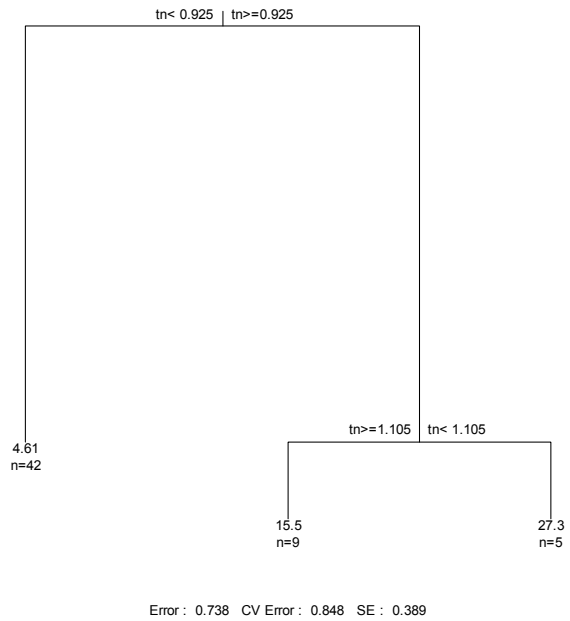
	CP	nsplit	rel error	xerror	xstd
1	0.14111034	0	1.0000000	1.016757	0.3777499
2	0.05827945	1	0.8588897	1.093769	0.3917636

Node number 1: 67 observations, complexity param=0.1411103
mean=5.154925, MSE=22.44666
left son=2 (39 obs) right son=3 (28 obs)
Primary splits:
tn < 0.545 to the right, improve=0.1411103, (0 missing)

Node number 2: 39 observations
mean=3.646923, MSE=11.40615

Node number 3: 28 observations
mean=7.255357, MSE=30.24524

TN versus Chlorophyll-a:



Code: mvpart(form = chla ~ tn, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=56 (13 observations deleted due to missing data)

	CP	nsplit	rel error	xerror	xstd
1	0.2624837	0	1.0000000	1.0183614	0.5256995
2	0.0488738	1	0.7375163	0.8482329	0.3889255

Node number 1: 56 observations, complexity param=0.2624837
mean=8.393571, MSE=163.8208
left son=2 (42 obs) right son=3 (14 obs)
Primary splits:
tn < 0.925 to the left, improve=0.2624837, (0 missing)

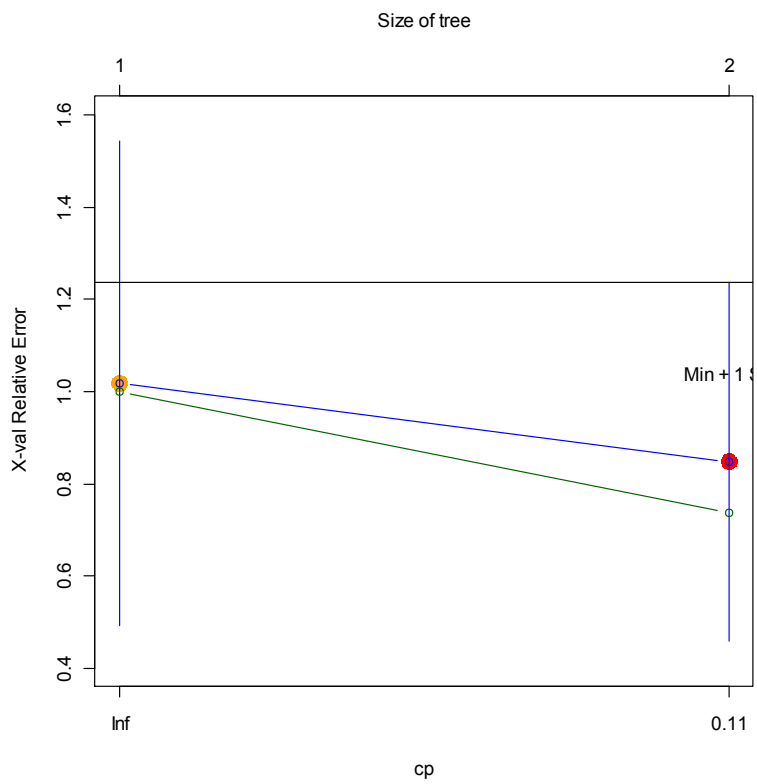
Node number 2: 42 observations
mean=4.607619, MSE=18.40699

Node number 3: 14 observations, complexity param=0.0488738
mean=19.75143, MSE=428.0612

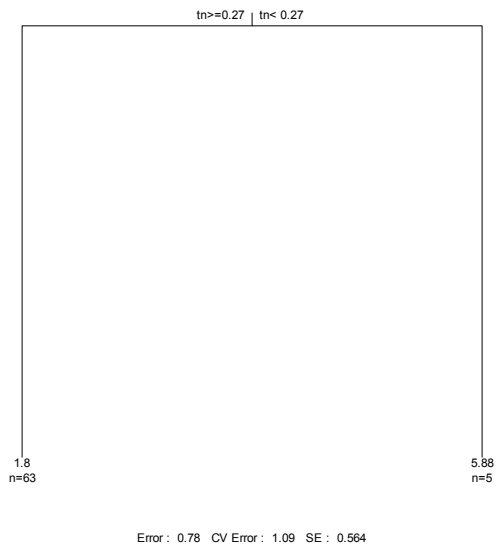
left son=6 (9 obs) right son=7 (5 obs)
Primary splits:
tn < 1.105 to the right, improve=0.07481685, (0 missing)

Node number 6: 9 observations
mean=15.53333, MSE=214.1451

Node number 7: 5 observations
mean=27.344, MSE=723.4367



Median Secchi modeled against Median TP and Median TN:



Code:mvpart(form = secchi ~ tp + tn, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=68 (1 observation deleted due to missingness)

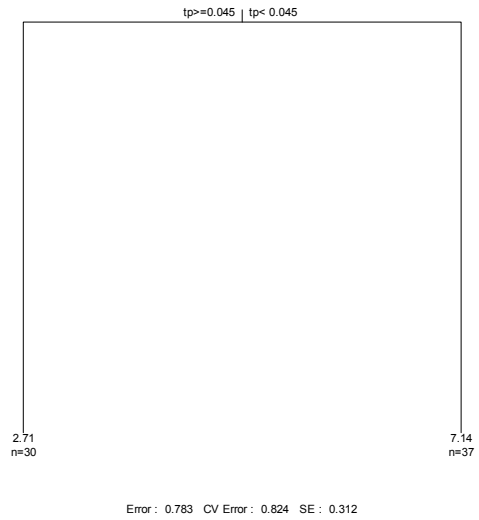
	CP	nsplit	rel error	xerror	xstd
1	0.2202675	0	1.0000000	1.021123	0.5605176
2	0.0755939	1	0.7797325	1.087709	0.5640912

Node number 1: 68 observations, complexity param=0.2202675
mean=2.095221, MSE=5.153133
left son=2 (63 obs) right son=3 (5 obs)
Primary splits:
tn < 0.27 to the right, improve=0.2202675, (0 missing)
tp < 0.035 to the right, improve=0.1351303, (0 missing)

Node number 2: 63 observations
mean=1.795079, MSE=2.281939

Node number 3: 5 observations
mean=5.877, MSE=25.89326

Median Euphotic thickness modeled against Median TP and Median TN:



Code: mvpart(form = zeu ~ tp + tn, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=67 (2 observations deleted due to missingness)

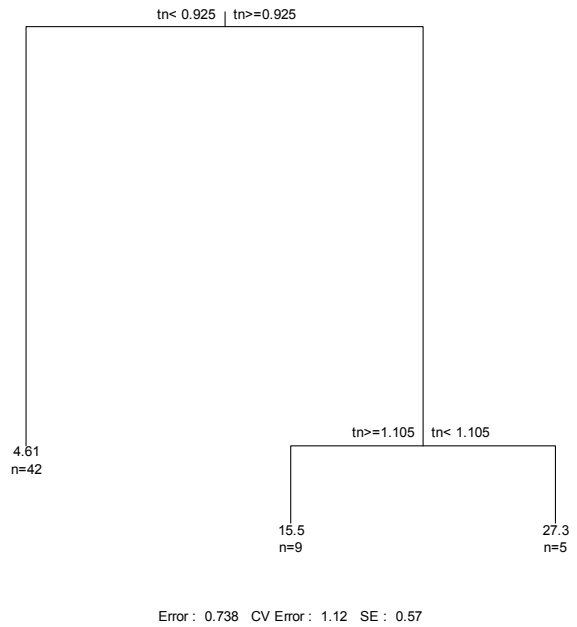
	CP	nsplit	rel error	xerror	xstd
1	0.21680710	0	1.0000000	1.025161	0.3758680
2	0.05287537	1	0.7831929	0.824346	0.3115756

Node number 1: 67 observations, complexity param=0.2168071
  mean=5.154925, MSE=22.44666
  left son=2 (30 obs) right son=3 (37 obs)
  Primary splits:
    tp < 0.045 to the right, improve=0.2168071, (0 missing)
    tn < 0.545 to the right, improve=0.1411103, (0 missing)

Node number 2: 30 observations
  mean=2.705, MSE=2.315892

Node number 3: 37 observations
  mean=7.141351, MSE=29.95642

Median Chlorophyll-a modeled against Median TP and Median TN:



Code:mvpart(form = chla ~ tp + tn, data = changepoint, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=56 (13 observations deleted due to missingness)

	CP	nsplit	rel error	xerror	xstd
1	0.2624837	0	1.0000000	1.035335	0.5339487
2	0.0488738	1	0.7375163	1.122106	0.5695687

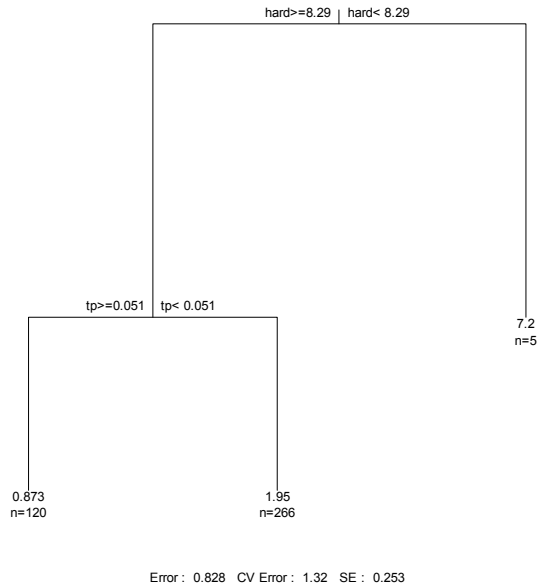
Node number 1: 56 observations, complexity param=0.2624837
mean=8.393571, MSE=163.8208
left son=2 (42 obs) right son=3 (14 obs)
Primary splits:
tn < 0.925 to the left, improve=0.2624837, (0 missing)
tp < 0.045 to the left, improve=0.1367305, (0 missing)

Node number 2: 42 observations
mean=4.607619, MSE=18.40699

Node number 3: 14 observations, complexity param=0.0488738
mean=19.75143, MSE=428.0612
left son=6 (9 obs) right son=7 (5 obs)

```
  Primary splits:
      tn < 1.105 to the right, improve=0.07481685, (0 missing)
      tp < 0.065 to the left,  improve=0.03467071, (0 missing)

Node number 6: 9 observations
  mean=15.53333, MSE=214.1451

Node number 7: 5 observations
  mean=27.344, MSE=723.4367
```

<u>Raw Data Regression Tree Model:</u> Secchi depth vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:



Code: mvpart(form = secchi ~ cond + alk + tss + tp + tn + hard + chloride + depth, data = regtree, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=399 (16 observations deleted due to missing data)

	CP	nsplit	rel error	xerror	xstd
1	0.1724401	0	1.00000	1.001055	0.2258289
2	0.1015200	1	0.82756	1.316157	0.2534981

Node number 1: 399 observations, complexity param=0.1724401
mean=1.678596, MSE=2.37394
left son=2 (387 obs) right son=3 (5 obs), 7 observations remain
Primary splits:
hard < 8.29 to the right, improve=0.16247840, (7 missing)
tp < 0.0515 to the right, improve=0.12296250, (2 missing)
depth < 21.5 to the left, improve=0.10374040, (6 missing)
tss < 3.5 to the right, improve=0.09239453, (66 missing)
tn < 0.405 to the right, improve=0.06626182, (6 missing)

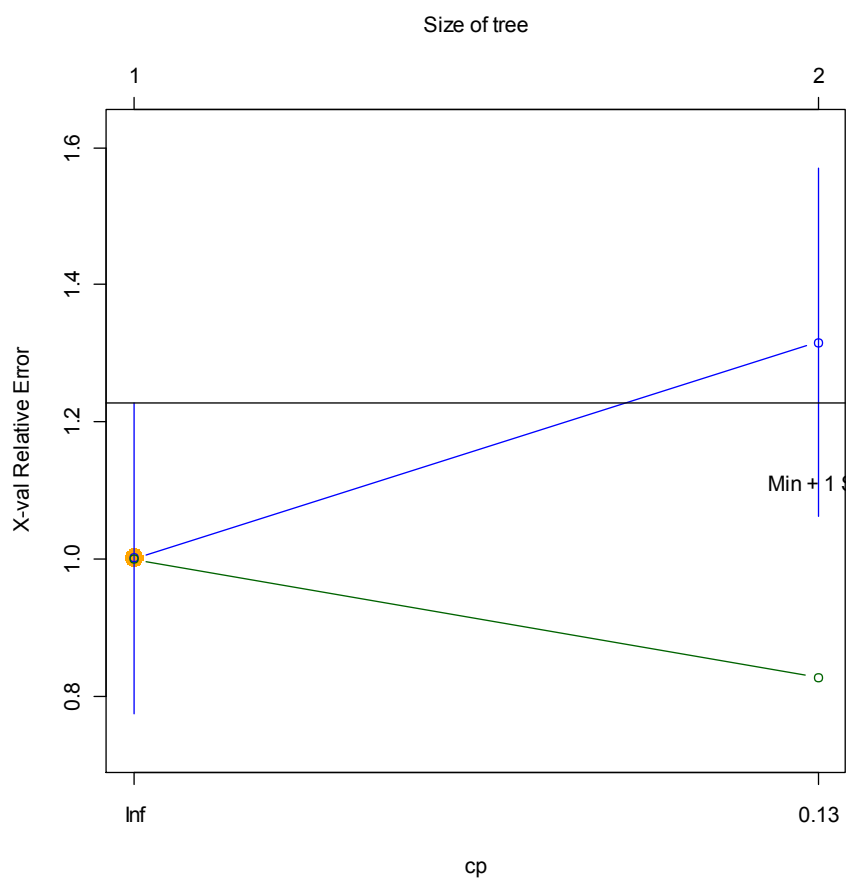
Node number 2: 387 observations, complexity param=0.10152
mean=1.616305, MSE=1.737692

left son=4 (120 obs) right son=5 (266 obs), 1 observation remains
Primary splits:
tp < 0.051 to the right, improve=0.14296500, (1 missing)
tss < 10.5 to the right, improve=0.14168000, (64 missing)
depth < 12.5 to the left, improve=0.10815250, (6 missing)
hard < 2445 to the left, improve=0.08345969, (0 missing)
tn < 0.385 to the right, improve=0.07836215, (5 missing)

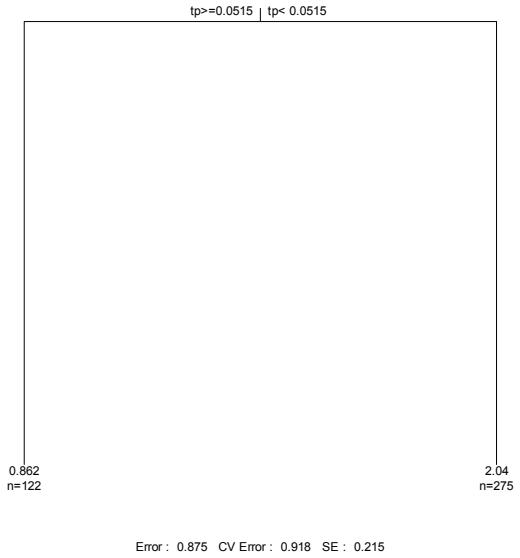
Node number 3: 5 observations
mean=7.2, MSE=22.276

Node number 4: 120 observations
mean=0.8729167, MSE=0.7356207

Node number 5: 266 observations
mean=1.951165, MSE=1.834782



<u>Raw Data Regression Tree Model:</u> Secchi depth modeled against specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, chloride, and depth:



Code: mvpart(form = secchi ~ cond + alk + tss + tp + tn + chloride + depth, data = regtree, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=399 (16 observations deleted due to missingness)

	CP	nsplit	rel error	xerror	xstd
1	0.12481651	0	1.0000000	1.001878	0.2261249
2	0.08573387	1	0.8751835	0.918266	0.2150948

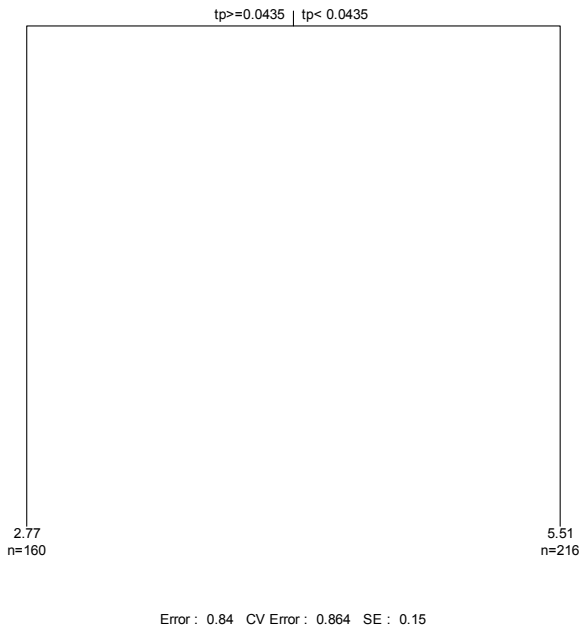
Node number 1: 399 observations, complexity param=0.1248165
mean=1.678596, MSE=2.37394
left son=2 (122 obs) right son=3 (275 obs), 2 observations remain
Primary splits:
tp < 0.0515 to the right, improve=0.12296250, (2 missing)
depth < 21.5 to the left, improve=0.10374040, (6 missing)
tss < 3.5 to the right, improve=0.09239453, (66 missing)
tn < 0.405 to the right, improve=0.06626182, (6 missing)
cond < 397.5 to the right, improve=0.04621581, (12 missing)

Node number 2: 122 observations

mean=0.8618852, MSE=0.7312317

Node number 3: 275 observations
mean=2.035855, MSE=2.690056

<u>Raw Data Regression Tree Model:</u> Euphotic thickness modeled against specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:



Call: mvpart(form = zeu ~ cond + alk + tss + tp + tn + hard + chloride + depth, data = regtree, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=378 (37 observations deleted due to missingness)

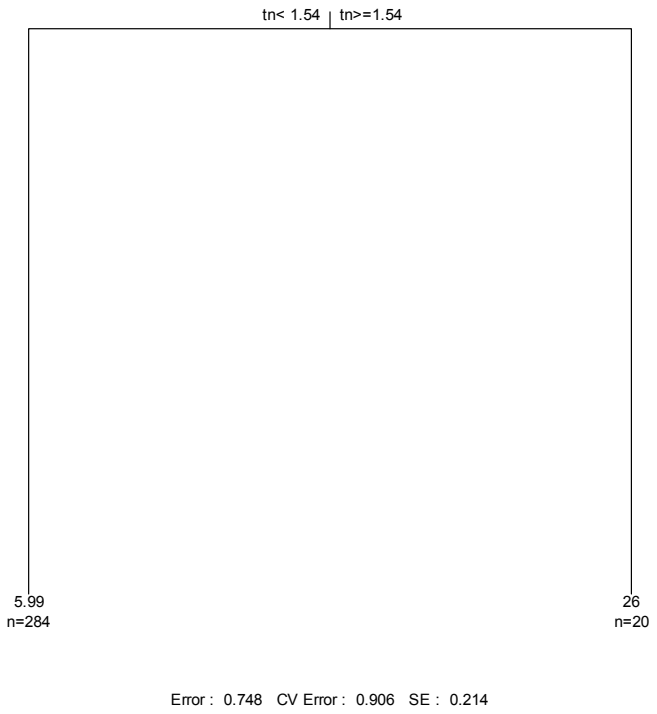
	CP	nsplit	rel error	xerror	xstd
1	0.16006557	0	1.0000000	1.0102287	0.1691944
2	0.09653995	1	0.8399344	0.8643109	0.1502659

Node number 1: 378 observations, complexity param=0.1600656
mean=4.346561, MSE=11.53598
left son=2 (160 obs) right son=3 (216 obs), 2 observations remain
Primary splits:
tp < 0.0435 to the right, improve=0.15886180, (2 missing)
depth < 12.5 to the left, improve=0.11768230, (2 missing)
hard < 2445 to the left, improve=0.08925156, (7 missing)
tss < 10.5 to the right, improve=0.08673668, (64 missing)
tn < 0.405 to the right, improve=0.06526280, (6 missing)

Node number 2: 160 observations
  mean=2.768125, MSE=4.463171

Node number 3: 216 observations
  mean=5.513426, MSE=13.65051

<u>Raw Data Regression Tree Model:</u> Chlorophyll-a modeled against specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:



Call: mvpart(form = chla ~ cond + alk + tss + tp + tn + hard + chloride + depth, data = regtree, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

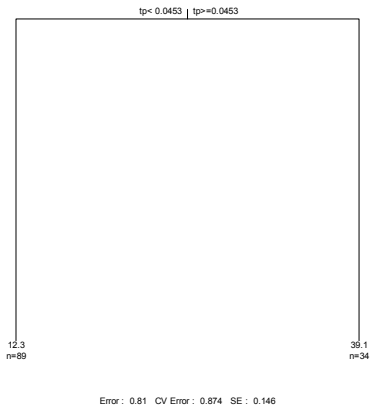
n=310 (105 observations deleted due to missingness)

	CP	nsplit	rel error	xerror	xstd
1	0.25221043	0	1.0000000	1.0110451	0.2238824
2	0.09964136	1	0.7477896	0.9063188	0.2142642

Node number 1: 310 observations, complexity param=0.2522104
mean=7.401968, MSE=143.2871
left son=2 (284 obs) right son=3 (20 obs), 6 observations remain
Primary splits:
tn < 1.54 to the left, improve=0.1684289, (6 missing)
tss < 3.5 to the left, improve=0.1178810, (5 missing)
hard < 1955 to the left, improve=0.1172366, (5 missing)
chloride < 1380 to the left, improve=0.1172037, (1 missing)

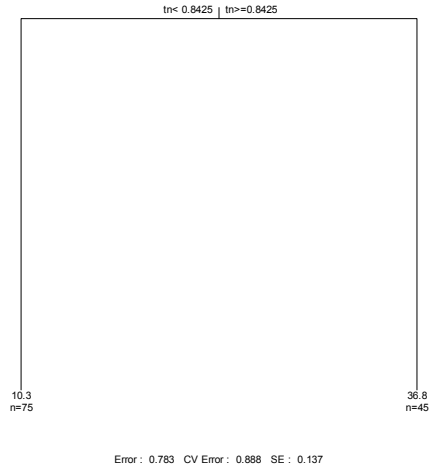
```
    cond     < 5660 to the left,  improve=0.1168433, (11 missing)

Node number 2: 284 observations
  mean=5.98662, MSE=101.7897

Node number 3: 20 observations
  mean=25.997, MSE=215.3886
```

<u>Raw Data Changepoint</u>: Percent Cyanobacertia with Total Phosphorus



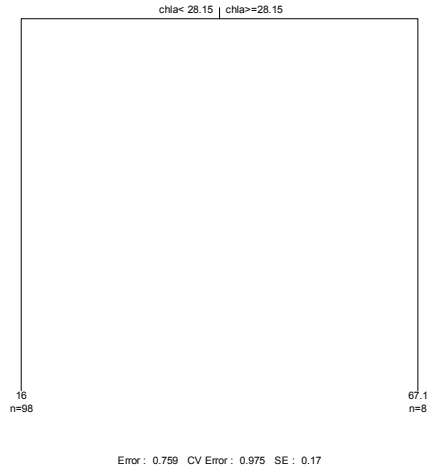

```
Call:
mvpart(form = percbg ~ tp, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n= 123

         CP nsplit rel error   xerror      xstd
1 0.19000627      0 1.0000000 1.012692 0.1609351
2 0.04066372      1 0.8099937 0.873988 0.1456467

Node number 1: 123 observations,    complexity param=0.1900063
  mean=19.72715, MSE=751.8374
  left son=2 (89 obs) right son=3 (34 obs)
  Primary splits:
      tp < 0.0453 to the left,  improve=0.1900063, (0 missing)

Node number 2: 89 observations
  mean=12.33978, MSE=409.4021

Node number 3: 34 observations
  mean=39.06471, MSE=1131.418
```

<u>Raw Data Changepoint</u>: Percent Cyanobacertia with Total Nitrogen



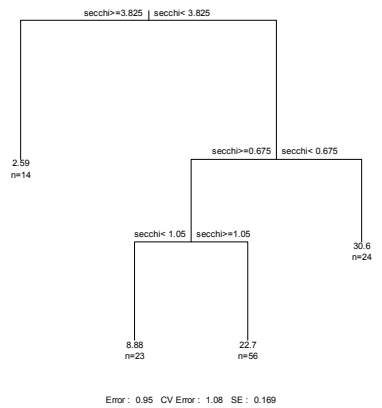

```
Call:
mvpart(form = percbg ~ tn, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=120 (3 observations deleted due to missingness)

         CP nsplit rel error    xerror      xstd
1 0.21663283      0 1.0000000 1.0092249 0.1589283
2 0.03592872      1 0.7833672 0.8876487 0.1370014

Node number 1: 120 observations,    complexity param=0.2166328
  mean=20.21533, MSE=760.8603
  left son=2 (75 obs) right son=3 (45 obs)
  Primary splits:
      tn < 0.8425 to the left,  improve=0.2166328, (0 missing)

Node number 2: 75 observations
  mean=10.27067, MSE=361.8927

Node number 3: 45 observations
  mean=36.78978, MSE=986.2667
```

<u>Raw Data Changepoint</u>: Percent Cyanobacertia with Chlorophyll-a




```
Call:
mvpart(form = percbg ~ chla, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=106 (17 observations deleted due to missingness)

         CP nsplit rel error    xerror      xstd
1 0.24078737      0 1.0000000 1.0129381 0.1648082
2 0.03458903      1 0.7592126 0.9179217 0.1659688

Node number 1: 106 observations,    complexity param=0.2407874
  mean=19.88528, MSE=754.9618
  left son=2 (98 obs) right son=3 (8 obs)
  Primary splits:
      chla < 28.15 to the left,  improve=0.2407874, (0 missing)

Node number 2: 98 observations
  mean=16.03306, MSE=545.0932

Node number 3: 8 observations
  mean=67.075, MSE=917.1969
```

<u>Raw Data Changepoint</u>: Percent Cyanobacertia with Secchi Depth



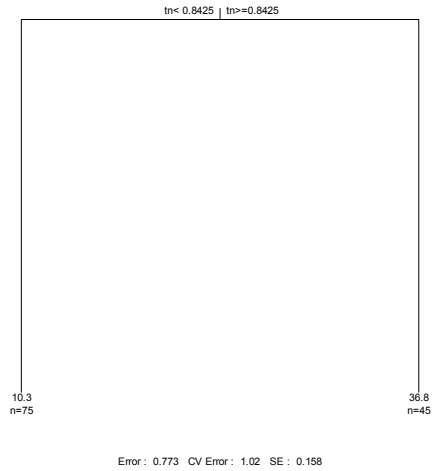

```
Call:
mvpart(form = percbg ~ secchi, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=117 (6 observations deleted due to missingness)

         CP nsplit rel error   xerror      xstd
1 0.05025757      0 1.0000000 1.022650 0.1682671
2 0.03281516      1 0.9497424 1.075018 0.1692252

Node number 1: 117 observations,    complexity param=0.05025757
  mean=19.17812, MSE=744.3179
  left son=2 (14 obs) right son=3 (103 obs)
  Primary splits:
      secchi < 3.825 to the right, improve=0.05025757, (0 missing)

Node number 2: 14 observations
  mean=2.588571, MSE=9.983698

Node number 3: 103 observations,    complexity param=0.03281516
  mean=21.43301, MSE=801.6381
  left son=6 (79 obs) right son=7 (24 obs)
  Primary splits:
      secchi < 0.675 to the right, improve=0.031644, (0 missing)

Node number 6: 79 observations,    complexity param=0.03281516
  mean=18.65696, MSE=652.3878
  left son=12 (23 obs) right son=13 (56 obs)
  Primary splits:
      secchi < 1.05  to the left,  improve=0.06019998, (0 missing)
```

Node number 7: 24 observations
  mean=30.57083, MSE=1184.054

Node number 12: 23 observations
  mean=8.878261, MSE=194.2104

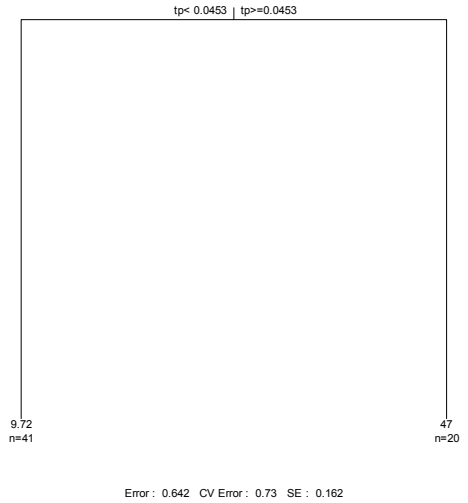
Node number 13: 56 observations
  mean=22.67321, MSE=785.1637

<u>Raw Data Categorical and Regression Tree Model</u>: Percent Cyanobacertia versus total phosphorus, total nitrogen, specific conductivity, alkalinity, total suspended solids, hardness, and chloride



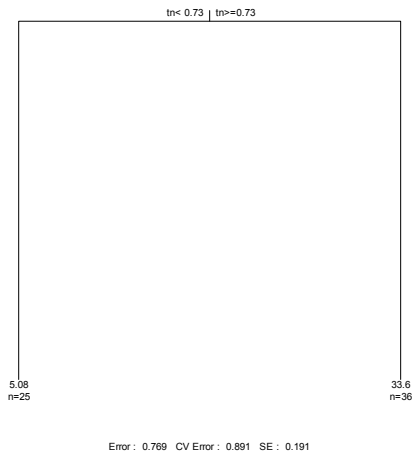

```
Call:
mvpart(form = percbg ~ tp + tn + spcond + alk + tss + hard +
    chlorid, data = bluegreen, xval = 10, method = "anova", minsplit = 10,
    minbucket = 5)
  n= 123

        CP nsplit rel error   xerror      xstd
1 0.2265674      0 1.0000000 1.007128 0.1598944
2 0.1785629      1 0.7734326 1.019618 0.1577588

Node number 1: 123 observations,    complexity param=0.2265674
  mean=19.72715, MSE=751.8374
  left son=2 (75 obs) right son=3 (45 obs), 3 observations remain
  Primary splits:
      tn     < 0.8425   to the left,  improve=0.21388550, (3 missing)
      tp     < 0.0453   to the left,  improve=0.19000630, (0 missing)
      alk    < 80.9     to the left,  improve=0.10983510, (0 missing)
      hard   < 113      to the left,  improve=0.06905282, (2 missing)
      spcond < 1862.333 to the right, improve=0.04534459, (9 missing)

Node number 2: 75 observations
  mean=10.27067, MSE=361.8927

Node number 3: 45 observations
  mean=36.78978, MSE=986.2667
```

<u>Summer Raw Data Only, Changepoint</u>: % Cyanobacteria with Total Phosphorus



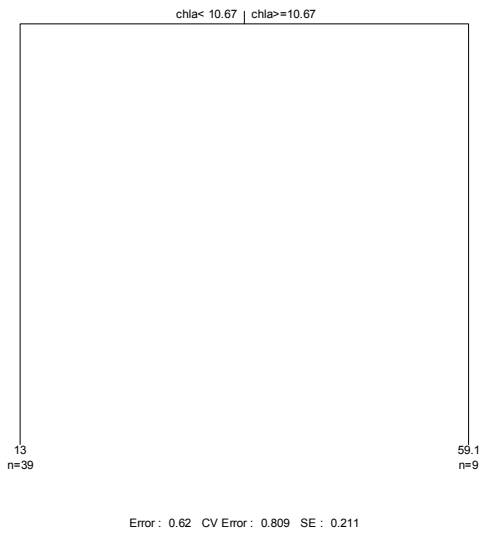

```
Call:
mvpart(form = percbg ~ tp, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n= 61

         CP nsplit rel error    xerror      xstd
1 0.35837761      0 1.0000000 1.0614352 0.2327669
2 0.08047535      1 0.6416224 0.7298032 0.1623148

Node number 1: 61 observations,    complexity param=0.3583776
  mean=21.92852, MSE=852.8608
  left son=2 (41 obs) right son=3 (20 obs)
  Primary splits:
      tp < 0.0453 to the left,  improve=0.3583776, (0 missing)

Node number 2: 41 observations
  mean=9.718049, MSE=247.6459

Node number 3: 20 observations
  mean=46.96, MSE=1161.33
```

<u>Summer Raw Data Only, Changepoint</u>: % Cyanobacteria with Total Nitrogen



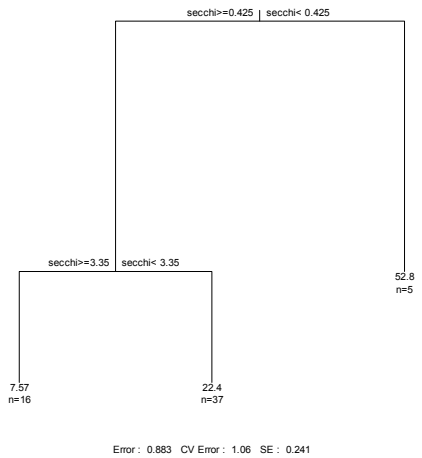

```
Call:
mvpart(form = percbg ~ tn, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n= 61

         CP nsplit rel error   xerror      xstd
1 0.23114425      0 1.0000000 1.038587 0.2247937
2 0.06726149      1 0.7688557 0.890629 0.1913638

Node number 1: 61 observations,    complexity param=0.2311443
  mean=21.92852, MSE=852.8608
  left son=2 (25 obs) right son=3 (36 obs)
  Primary splits:
      tn < 0.73 to the left,  improve=0.2311443, (0 missing)

Node number 2: 25 observations
  mean=5.08, MSE=53.2632

Node number 3: 36 observations
  mean=33.62889, MSE=1074.105
```

<u>Summer Raw Data Only, Changepoint</u>: % Cyanobacteria with Chlorophyll-a




```
Call:
mvpart(form = percbg ~ chla, data = bluegreen, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=48 (13 observations deleted due to missingness)

         CP nsplit rel error    xerror      xstd
1 0.38006463      0 1.0000000 1.0364768 0.2437454
2 0.04417035      1 0.6199354 0.8089295 0.2114517

Node number 1: 48 observations,    complexity param=0.3800646
  mean=21.61958, MSE=851.4447
  left son=2 (39 obs) right son=3 (9 obs)
  Primary splits:
     chla < 10.67 to the left,  improve=0.3800646, (0 missing)

Node number 2: 39 observations
  mean=12.97795, MSE=405.207

Node number 3: 9 observations
  mean=59.06667, MSE=1059.253
```

<u>Summer Raw Data Only, Changepoint</u>: % Cyanobacteria with Secchi depth



Call:
mvpart(form = percbg ~ secchi, data = bluegreen, xval = 10, method = "anova",
  minsplit = 10, minbucket = 5)
 n=58 (3 observations deleted due to missingness)

```
         CP nsplit rel error   xerror      xstd
1 0.11733941      0 1.0000000 1.033428 0.2409786
2 0.05182584      1 0.8826606 1.055918 0.2410674
```

Node number 1: 58 observations, complexity param=0.1173394
 mean=20.93517, MSE=818.3951
 left son=2 (53 obs) right son=3 (5 obs)
 Primary splits:
   secchi < 0.425 to the right, improve=0.1173394, (0 missing)

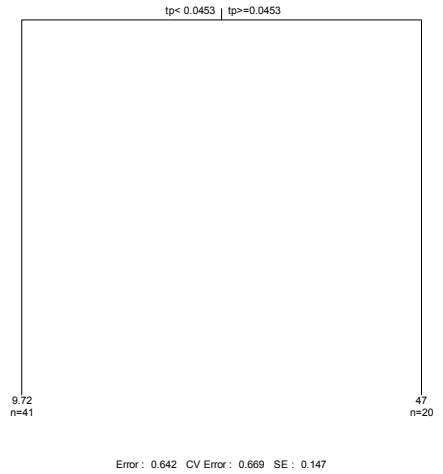
Node number 2: 53 observations, complexity param=0.05182584
 mean=17.92528, MSE=642.7239
 left son=4 (16 obs) right son=5 (37 obs)
 Primary splits:
   secchi < 3.35 to the right, improve=0.0722166, (0 missing)

Node number 3: 5 observations
 mean=52.84, MSE=1566.562

Node number 4: 16 observations
 mean=7.565, MSE=113.7734

Node number 5: 37 observations
  mean=22.40541, MSE=804.9724

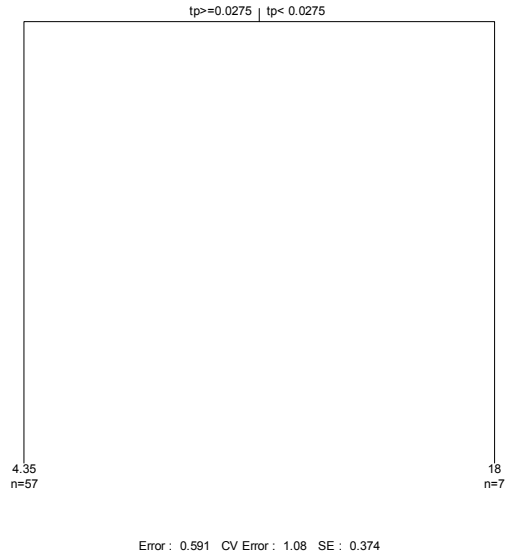
<u>Summer Raw Data only, Categorical and Regression Tree Model</u>: Percent Cyanobacertia versus total phosphorus, total nitrogen, specific conductivity, alkalinity, total suspended solids, hardness, and chloride




```
Call:
mvpart(form = percbg ~ tp + tn + spcond + alk + tss + hard +
   chlorid, data = bluegreen, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n= 61

        CP nsplit rel error    xerror      xstd
1 0.3583776      0 1.0000000 1.0337056 0.2244527
2 0.1054121      1 0.6416224 0.6693424 0.1465032

Node number 1: 61 observations,    complexity param=0.3583776
  mean=21.92852, MSE=852.8608
  left son=2 (41 obs) right son=3 (20 obs)
  Primary splits:
     tp      < 0.0453 to the left,  improve=0.35837760, (0 missing)
     tn      < 0.73   to the left,  improve=0.23114430, (0 missing)
     alk     < 86.1   to the left,  improve=0.11671460, (0 missing)
     hard    < 449.5  to the right, improve=0.08587177, (1 missing)
     chlorid < 13.65  to the right, improve=0.08343943, (0 missing)

Node number 2: 41 observations
  mean=9.718049, MSE=247.6459

Node number 3: 20 observations
  mean=46.96, MSE=1161.33
```

APPENDIX II – R CODE & OUTPUT FOR RIVER DATA

TP versus benthic chlorophyll-a:



Code: mvpart(form = chla ~ tp, data = river, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=64 (72 observations deleted due to missingness)

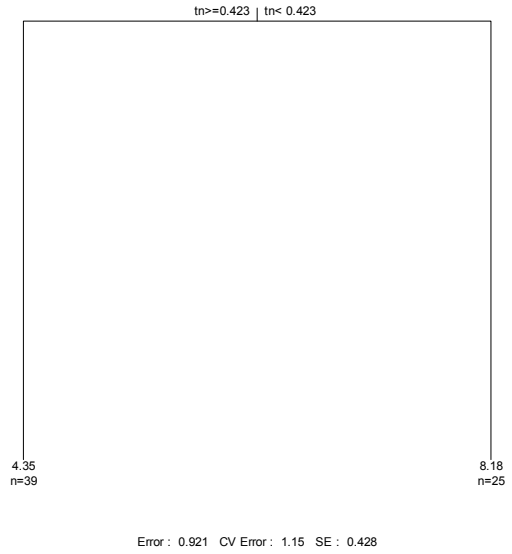
	CP	nsplit	rel error	xerror	xstd
1	0.40930685	0	1.0000000	1.043794	0.3804436
2	0.02672565	1	0.5906932	1.078793	0.3737727

Node number 1: 64 observations, complexity param=0.4093068
mean=5.846155, MSE=44.43232
left son=2 (57 obs) right son=3 (7 obs)
Primary splits:
tp < 0.0275 to the right, improve=0.4093068, (0 missing)

Node number 2: 57 observations
mean=4.351691, MSE=14.93078

Node number 3: 7 observations
mean=18.01537, MSE=118.383

TN versus benthic chlorophyll-a :



Code: mvpart(form = chla ~ tn, data = river, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

n=64 (72 observations deleted due to missingness)

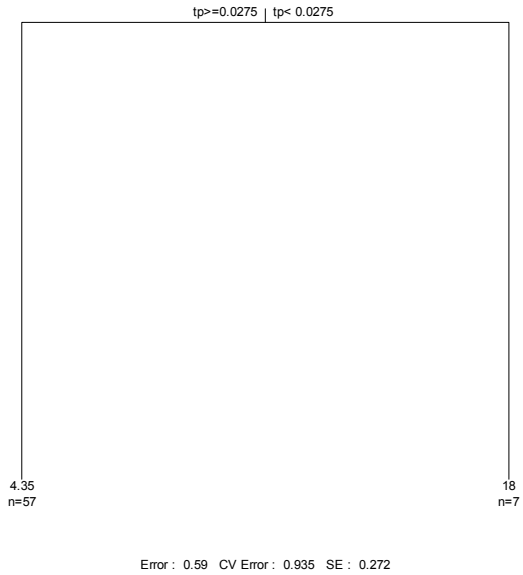
	CP	nsplit	rel error	xerror	xstd
1	0.0788103	0	1.0000000	1.025748	0.3694171
2	0.0611923	1	0.9211897	1.147163	0.4282900

Node number 1: 64 observations, complexity param=0.07881031
  mean=5.846155, MSE=44.43232
  left son=2 (39 obs) right son=3 (25 obs)
  Primary splits:
    tn < 0.423 to the right, improve=0.07881031, (0 missing)

Node number 2: 39 observations
  mean=4.347924, MSE=37.32589

Node number 3: 25 observations
  mean=8.183395, MSE=46.55393

<u>Regression Tree Model:</u> Benthic chlorophyll-a modeled against pH, electrical conductivity, temperature, dissolved oxygen, turbidity, total phosphorus, total nitrogen, total dissolved solids, total suspended solids, alkalinity, and substrate type:



Code: mvpart(form = chla ~ ph + ec + temp + do + turb + tp + tn + tds + tss + alk + substrate, data = river, xval = 10, method = "anova", minsplit = 10, minbucket = 5)

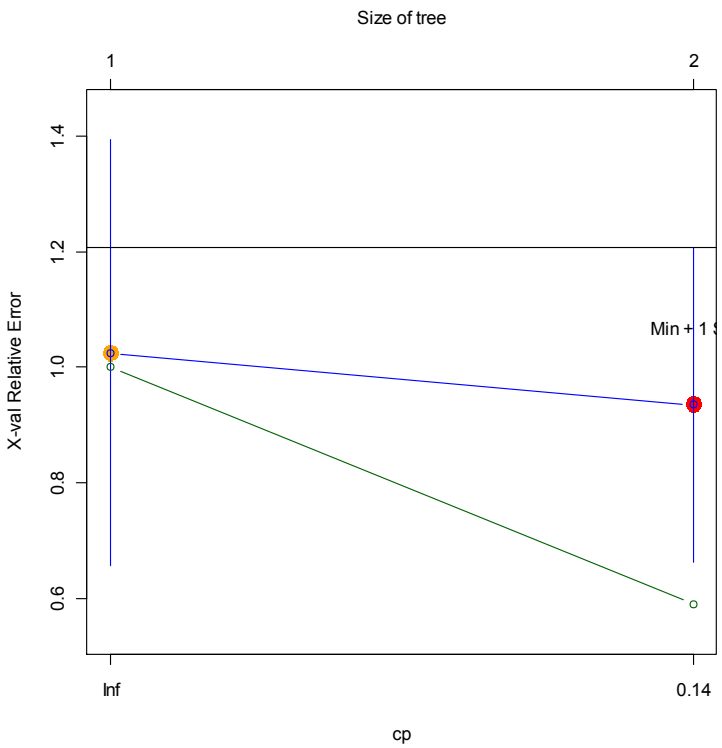
n=66 (70 observations deleted due to missingness)

	CP	nsplit	rel error	xerror	xstd
1	0.41012984	0	1.0000000	1.0245909	0.3692391
2	0.04616386	1	0.5898702	0.9347497	0.2721919

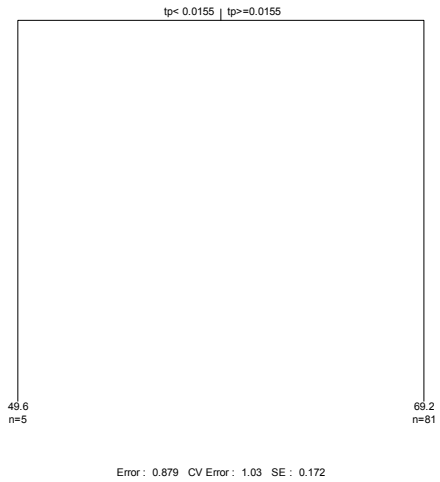
Node number 1: 66 observations, complexity param=0.4101298
mean=5.878143, MSE=43.146
left son=2 (57 obs) right son=3 (7 obs), 2 observations remain
Primary splits:
tp < 0.0275 to the right, improve=0.40873660, (2 missing)
temp < 8.82 to the right, improve=0.19254670, (2 missing)
do < 10.955 to the left, improve=0.19081790, (2 missing)
turb < 75.9 to the right, improve=0.08486004, (8 missing)
tn < 0.423 to the right, improve=0.07870050, (2 missing)

Node number 2: 57 observations
mean=4.351691, MSE=14.93078

Node number 3: 7 observations
mean=18.01537, MSE=118.383

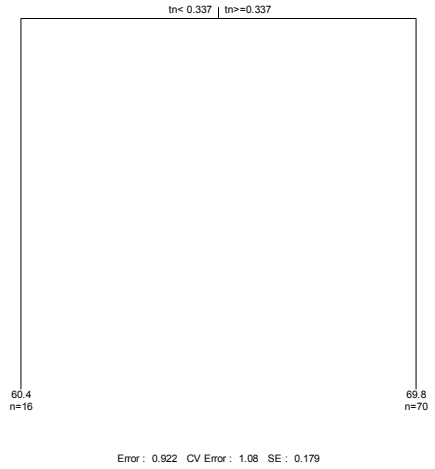


TP versus Tropic Diatom Index:



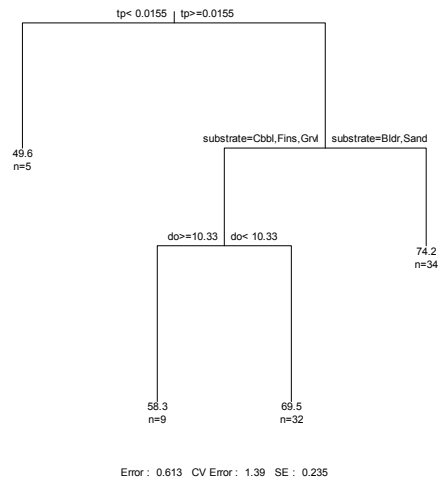

```
Call:
mvpart(form = tdi ~ tp, data = river, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=86 (50 observations deleted due to missingness)

         CP nsplit rel error   xerror      xstd
1 0.12107576      0 1.0000000 1.044711 0.1718751
2 0.06283041      1 0.8789242 1.028111 0.1715458

Node number 1: 86 observations,    complexity param=0.1210758
  mean=68.03179, MSE=172.5158
  left son=2 (5 obs) right son=3 (81 obs)
  Primary splits:
      tp < 0.0155 to the left,  improve=0.1210758, (0 missing)

Node number 2: 5 observations
  mean=49.63676, MSE=86.09167

Node number 3: 81 observations
  mean=69.16728, MSE=155.6738
```

Total Nitrogen versus Trophic Diatom Index:




```
Call:
mvpart(form = tdi ~ tn, data = river, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=86 (50 observations deleted due to missingness)

         CP nsplit rel error   xerror      xstd
1 0.07815056      0 1.0000000 1.045913 0.1737801
2 0.04337979      1 0.9218494 1.078204 0.1786331

Node number 1: 86 observations,    complexity param=0.07815056
  mean=68.03179, MSE=172.5158
  left son=2 (16 obs) right son=3 (70 obs)
  Primary splits:
      tn < 0.337 to the left,  improve=0.07815056, (0 missing)

Node number 2: 16 observations
  mean=60.35164, MSE=132.9509

Node number 3: 70 observations
  mean=69.78725, MSE=164.9953
```

<u>Categorical and Regression Tree Model</u>: Trophic Diatom Index predicted from pH, specific conductance, temperature, dissolved oxygen, turbidity, total phosphorus, total nitrogen, total dissolved solids, total suspended solids, alkalinity, and substrate type:

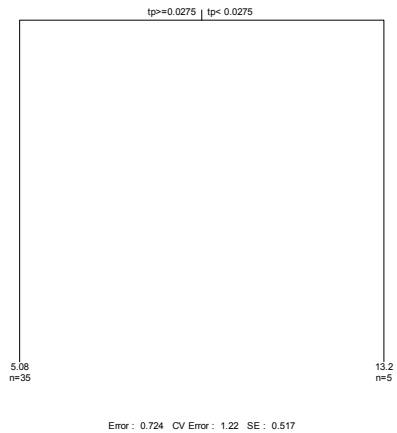



```
Call:
mvpart(form = tdi ~ ph + ec + temp + do + turb + tp + tn + tds +
   tss + alk + substrate, data = river, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=89 (47 observations deleted due to missingness)

         CP nsplit rel error   xerror      xstd
1 0.12912684      0 1.0000000 1.026760 0.1683181
2 0.06749652      3 0.6126195 1.386006 0.2353084

Node number 1: 89 observations,    complexity param=0.1291268
  mean=67.96994, MSE=168.0248
  left son=2 (5 obs) right son=3 (81 obs), 3 observations remain
  Primary splits:
     tp        < 0.0155 to the left,  improve=0.12012160, (3 missing)
     substrate splits as  -RLLLR,     improve=0.09567901, (0 missing)
     ph        < 7.935  to the right, improve=0.08319322, (7 missing)
     tn        < 0.337  to the left,  improve=0.07753467, (3 missing)
     tds       < 228    to the left,  improve=0.07288387, (24 missing)

Node number 2: 5 observations
  mean=49.63676, MSE=86.09167

Node number 3: 81 observations,    complexity param=0.1291268
```

```
  mean=69.16728, MSE=155.6738
  left son=6 (47 obs) right son=7 (34 obs)
  Primary splits:
     substrate splits as  -RLLLR,     improve=0.11850260, (0 missing)
     ph        < 7.935  to the right, improve=0.07305775, (7 missing)
     do        < 11.355 to the right, improve=0.06304562, (6 missing)
     alk       < 161    to the left,  improve=0.04978058, (27 missing)
     temp      < 10.93  to the left,  improve=0.03229882, (10 missing)

Node number 6: 47 observations,    complexity param=0.1291268
  mean=65.51418, MSE=169.5214
  left son=12 (9 obs) right son=13 (32 obs), 6 observations remain
  Primary splits:
     do   < 10.33  to the right, improve=0.11069020, (6 missing)
     turb < 10.45  to the left,  improve=0.09382598, (10 missing)
     temp < 19.955 to the right, improve=0.08325087, (7 missing)
     tp   < 0.166  to the right, improve=0.07448020, (0 missing)
     ph   < 7.88   to the right, improve=0.06891077, (7 missing)

Node number 7: 34 observations
  mean=74.21717, MSE=92.58248

Node number 12: 9 observations
  mean=58.33181, MSE=249.6311

Node number 13: 32 observations
  mean=69.53679, MSE=104.2593
```

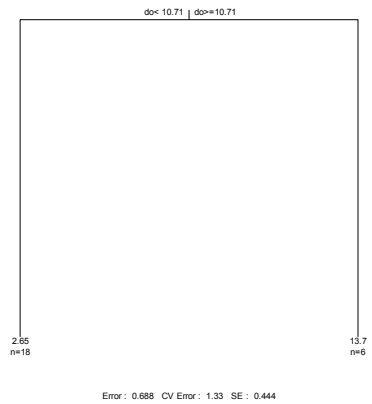
APPENDIX III – R CODE & OUTPUT FOR REVISIONS

<u>Regression Tree Model on River excluding Sandy Substrate:</u> Benthic chlorophyll-a modeled against pH, electrical conductivity, temperature, dissolved oxygen, turbidity, total phosphorus, total nitrogen, total dissolved solids, total suspended solids, alkalinity, and substrate type:



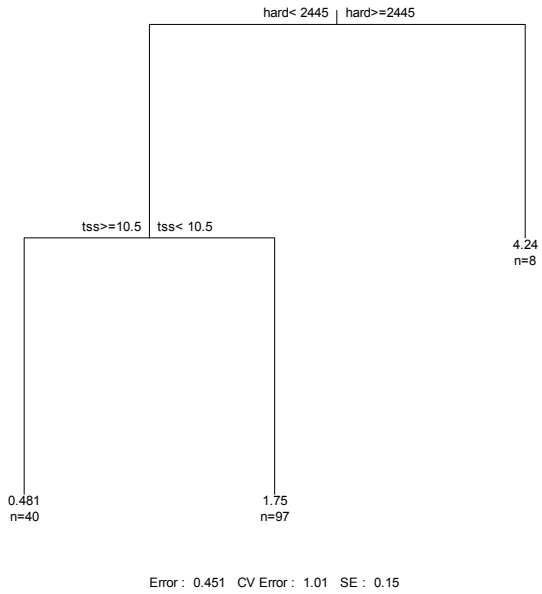

```
Call:
mvpart(form = chla ~ ph + ec + temp + do + turb + tp + tn + tds +
   tss + alk + substrate, data = river_nosand, xval = 10, method = "anova",
   minsplit = 10, minbucket = 5)
  n=42 (34 observations deleted due to missingness)

        CP nsplit rel error   xerror      xstd
1 0.2759100      0   1.00000 1.025509 0.5132811
2 0.1183210      1   0.72409 1.219681 0.5169798

Node number 1: 42 observations,    complexity param=0.27591
  mean=6.139981, MSE=25.45669
  left son=2 (35 obs) right son=3 (5 obs), 2 observations remain
  Primary splits:
      tp   < 0.0275 to the right, improve=0.27308060, (2 missing)
      temp < 8.82   to the right, improve=0.24592090, (2 missing)
      tn   < 0.435  to the right, improve=0.14658660, (2 missing)
      do   < 9.99   to the left,  improve=0.13264150, (2 missing)
      turb < 102    to the right, improve=0.09210584, (7 missing)

Node number 2: 35 observations
  mean=5.080737, MSE=11.01324

Node number 3: 5 observations
  mean=13.24998, MSE=77.74398
```

<u>Regression Tree Model on River with only Sandy Substrate:</u> Benthic chlorophyll-a modeled against pH, electrical conductivity, temperature, dissolved oxygen, turbidity, total phosphorus, total nitrogen, total dissolved solids, total suspended solids, alkalinity, and substrate type:




```
Call:
mvpart(form = chla ~ ph + ec + temp + do + turb + tp + tn + tds +
   tss + alk + substrate, data = river_sandonly, xval = 10,
   method = "anova", minsplit = 10, minbucket = 5)
  n=24 (29 observations deleted due to missingness)

         CP nsplit rel error   xerror      xstd
1 0.31190680      0 1.0000000 1.045734 0.5291646
2 0.03448058      1 0.6880932 1.325068 0.4435624

Node number 1: 24 observations,    complexity param=0.3119068
  mean=5.419926, MSE=73.77234
  left son=2 (18 obs) right son=3 (6 obs)
  Primary splits:
     do   < 10.71 to the left,  improve=0.31190680, (0 missing)
     turb < 7.7   to the right, improve=0.26871110, (1 missing)
     alk  < 196.5 to the left,  improve=0.16772190, (6 missing)
     temp < 16.08 to the right, improve=0.10012200, (0 missing)
     tss  < 4.5   to the right, improve=0.09257918, (2 missing)

Node number 2: 18 observations
  mean=2.650444, MSE=11.22933

Node number 3: 6 observations
  mean=13.72837, MSE=169.361
```

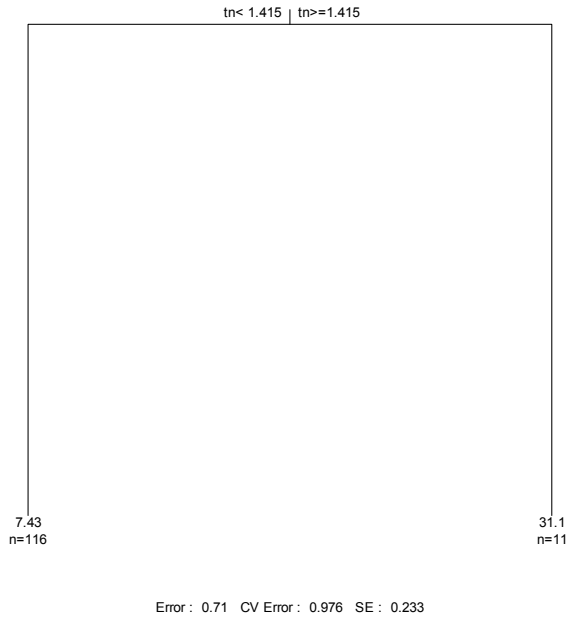
<u>Raw Data Regression Tree Model - WWAL:</u> Secchi depth vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:




```
Call:
mvpart(form = secchi ~ cond + alk + tss + tp + tn + hard + chloride +
   depth, data = wwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n=207 (10 observations deleted due to missingness)

         CP nsplit rel error   xerror      xstd
1 0.27458780      0 1.0000000 1.008569 0.1774353
2 0.07552026      2 0.4508244 1.009323 0.1496708

Node number 1: 207 observations,    complexity param=0.2745878
  mean=1.272947, MSE=1.469172
  left son=2 (196 obs) right son=3 (8 obs), 3 observations remain
  Primary splits:
     hard     < 2445   to the left,  improve=0.2394967, (3 missing)
     tp       < 0.0375 to the right, improve=0.2046210, (1 missing)
     chloride < 2185   to the left,  improve=0.1848603, (57 missing)
     cond     < 8720   to the left,  improve=0.1472689, (7 missing)
     tss      < 10.5   to the right, improve=0.1398083, (61 missing)
```

```
Node number 2: 196 observations,    complexity param=0.2745878
  mean=1.159184, MSE=1.064686
  left son=4 (40 obs) right son=5 (97 obs), 59 observations remain
  Primary splits:
      tss   < 10.5   to the right, improve=0.21758020, (59 missing)
      tp    < 0.0375 to the right, improve=0.17868050, (1 missing)
      depth < 6.5    to the left,  improve=0.11244470, (5 missing)
      tn    < 0.66   to the right, improve=0.06681980, (1 missing)
      hard  < 110    to the right, improve=0.05733389, (0 missing)

Node number 3: 8 observations
  mean=4.2375, MSE=2.464844

Node number 4: 40 observations
  mean=0.48125, MSE=0.04871094

Node number 5: 97 observations
  mean=1.747423, MSE=1.190071
```

<u>Raw Data Regression Tree Model - WWAL:</u> chlorophyll-a vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:

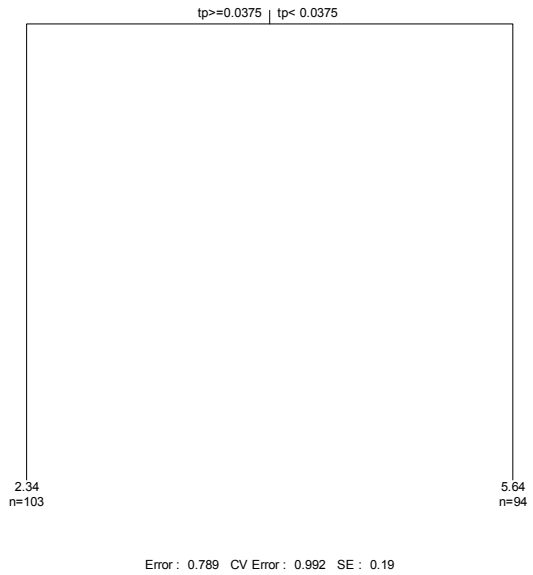



```
Call:
mvpart(form = chla ~ cond + alk + tss + tp + tn + hard + chloride +
   depth, data = wwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n=128 (89 observations deleted due to missingness)

        CP nsplit rel error    xerror      xstd
1 0.2904629      0 1.0000000 1.0091916 0.2281888
2 0.1484697      1 0.7095371 0.9755175 0.2334818

Node number 1: 128 observations,    complexity param=0.2904629
  mean=9.412891, MSE=153.2172
  left son=2 (116 obs) right son=3 (11 obs), 1 observation remains
  Primary splits:
     tn       < 1.415 to the left,  improve=0.2869711, (1 missing)
     tss      < 16.5  to the left,  improve=0.2526697, (2 missing)
     chloride < 1380  to the left,  improve=0.2386395, (0 missing)
     hard     < 1955  to the left,  improve=0.2377126, (1 missing)
     cond     < 5660  to the left,  improve=0.2368617, (5 missing)
```

Node number 2: 116 observations
  mean=7.427845, MSE=105.8439

Node number 3: 11 observations
  mean=31.09545, MSE=148.8559

<u>Raw Data Regression Tree Model - WWAL:</u> euphotic depth vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:

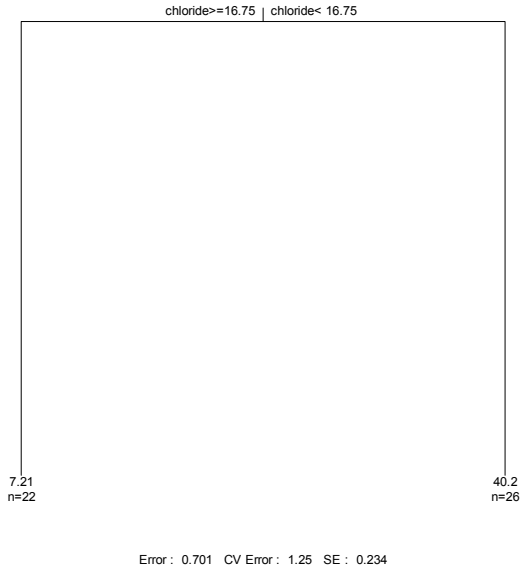



```
Call:
mvpart(form = zeu ~ cond + alk + tss + tp + tn + hard + chloride +
   depth, data = wwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n=198 (19 observations deleted due to missingness)

        CP nsplit rel error    xerror      xstd
1 0.2105021      0  1.000000 1.0078299 0.2089996
2 0.1547484      1  0.789498 0.9924744 0.1897389

Node number 1: 198 observations,    complexity param=0.2105021
  mean=3.927273, MSE=12.98713
  left son=2 (103 obs) right son=3 (94 obs), 1 observation remains
  Primary splits:
     tp    < 0.0375 to the right, improve=0.2084832, (1 missing)
     hard  < 2445   to the left,  improve=0.1737394, (3 missing)
     tss   < 10.5   to the right, improve=0.1666931, (59 missing)
     cond  < 8720   to the left,  improve=0.1216621, (5 missing)
     depth < 3.5    to the left,  improve=0.1206643, (1 missing)

Node number 2: 103 observations
```

mean=2.339806, MSE=4.268998

Node number 3: 94 observations
mean=5.642553, MSE=16.91968

<u>Raw Data Regression Tree Model - WWAL:</u> % cyanobacteria vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, and chloride:

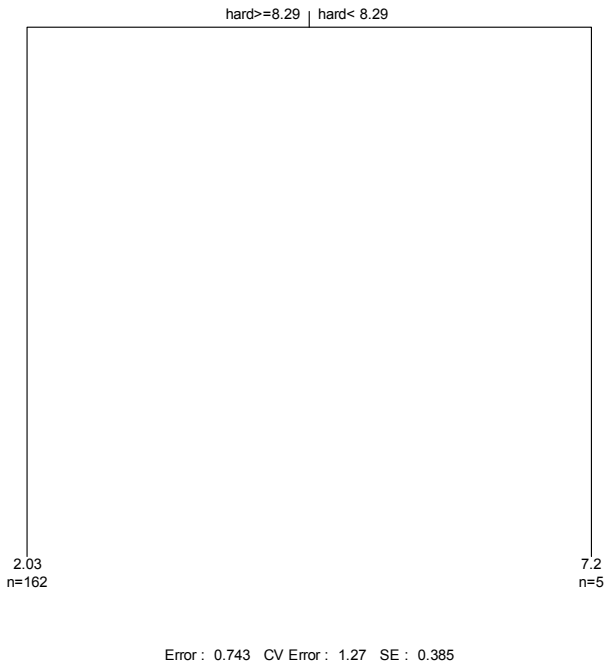



```
Call:
mvpart(form = cyano ~ cond + alk + tss + tp + tn + hard + chloride,
   data = phyto_wwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n= 48

        CP nsplit rel error   xerror      xstd
1 0.2991922      0 1.0000000 1.034392 0.2116424
2 0.2818948      1 0.7008078 1.246526 0.2341095

Node number 1: 48 observations,    complexity param=0.2991922
  mean=25.05625, MSE=900.8183
  left son=2 (22 obs) right son=3 (26 obs)
  Primary splits:
     chloride < 16.75    to the right, improve=0.2991922, (0 missing)
     hard     < 441      to the right, improve=0.2945640, (2 missing)
     alk      < 165.5    to the left,  improve=0.2840769, (0 missing)
     cond     < 1317.333 to the right, improve=0.2617664, (3 missing)
     tp       < 0.0455   to the left,  improve=0.1943421, (0 missing)
```

Node number 2: 22 observations
  mean=7.209091, MSE=116.9317

Node number 3: 26 observations
  mean=40.15769, MSE=1066.536

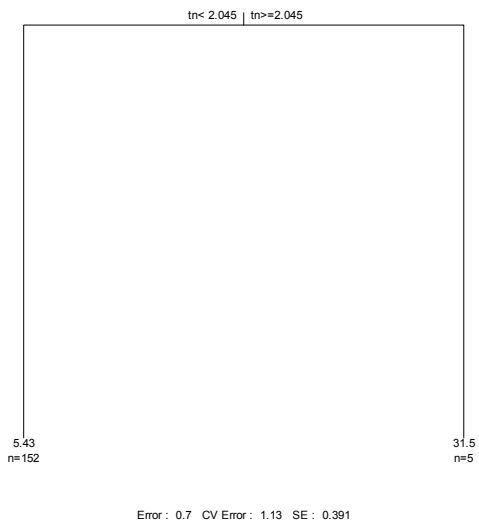
<u>Raw Data Regression Tree Model - CWAL:</u> secchi depth vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:




```
Call:
mvpart(form = secchi ~ cond + alk + tss + tp + tn + hard + chloride +
   depth, data = cwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n=171 (12 observations deleted due to missingness)

        CP nsplit rel error   xerror      xstd
1 0.2569505      0 1.0000000 1.018383 0.3679875
2 0.1491342      1 0.7430495 1.274931 0.3851839

Node number 1: 171 observations,    complexity param=0.2569505
  mean=2.169942, MSE=3.10164
  left son=2 (162 obs) right son=3 (5 obs), 4 observations remain
  Primary splits:
     hard  < 8.29   to the right, improve=0.24423000, (4 missing)
     tss   < 3.5    to the right, improve=0.12394890, (3 missing)
     tp    < 0.0455 to the right, improve=0.08267465, (1 missing)
     depth < 14.5   to the left,  improve=0.07784807, (1 missing)
```

```
    tn    < 0.215  to the right, improve=0.06672882, (5 missing)

Node number 2: 162 observations
  mean=2.03216, MSE=1.745179

Node number 3: 5 observations
  mean=7.2, MSE=22.276
```

<u>Raw Data Regression Tree Model - CWAL:</u> chlorophyll-a vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:

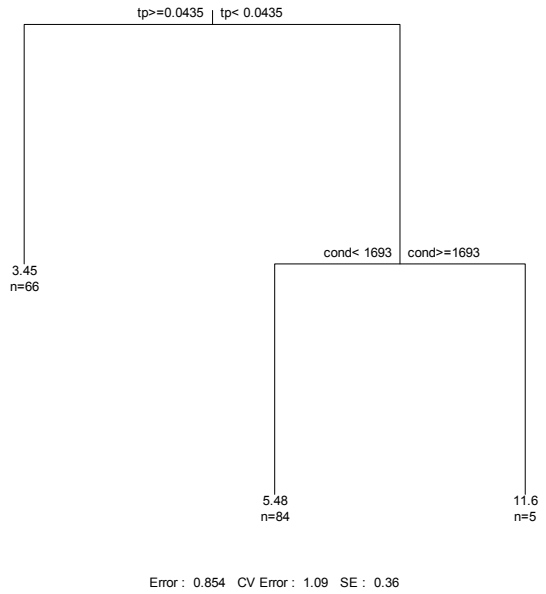



```
Call:
mvpart(form = chla ~ cond + alk + tss + tp + tn + hard + chloride +
   depth, data = cwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n=162 (21 observations deleted due to missingness)

        CP nsplit rel error   xerror      xstd
1 0.3002423      0 1.0000000 1.010663 0.3766302
2 0.1085285      1 0.6997577 1.130796 0.3913185

Node number 1: 162 observations,    complexity param=0.3002423
  mean=6.518765, MSE=145.0412
  left son=2 (152 obs) right son=3 (5 obs), 5 observations remain
  Primary splits:
     tn    < 2.045  to the left,  improve=0.14029470, (5 missing)
     tss   < 3.5    to the left,  improve=0.11850010, (1 missing)
     tp    < 0.0245 to the left,  improve=0.08488622, (1 missing)
     depth < 9.5    to the right, improve=0.05898549, (3 missing)
     cond  < 3946.5 to the left,  improve=0.04146903, (6 missing)

Node number 2: 152 observations
```

mean=5.428421, MSE=104.539

Node number 3: 5 observations
mean=31.524, MSE=110.4111

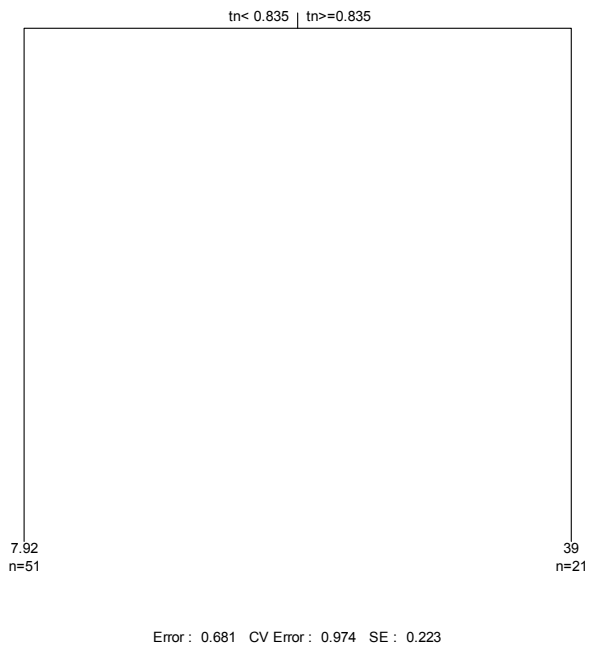
<u>Raw Data Regression Tree Model - CWAL:</u> euphotic depth vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:




```
Call:
mvpart(form = zeu ~ cond + alk + tss + tp + tn + hard + chloride +
   depth, data = cwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n=159 (24 observations deleted due to missingness)

        CP nsplit rel error   xerror      xstd
1 0.1461130      0  1.000000 1.011366 0.3178421
2 0.1407290      1  0.853887 1.085788 0.3599464

Node number 1: 159 observations,    complexity param=0.146113
  mean=4.863522, MSE=9.999047
  left son=2 (66 obs) right son=3 (92 obs), 1 observation remains
  Primary splits:
     tp    < 0.0435 to the right, improve=0.14391490, (1 missing)
     depth < 14.5   to the left,  improve=0.13490070, (1 missing)
     tss   < 116    to the left,  improve=0.12419870, (3 missing)
     tn    < 0.385  to the right, improve=0.08629971, (5 missing)
     cond  < 2101.5 to the left,  improve=0.06527229, (5 missing)
```

```
Node number 2: 66 observations
  mean=3.454545, MSE=3.747934

Node number 3: 92 observations,    complexity param=0.140729
  mean=5.894565, MSE=12.06725
  left son=6 (84 obs) right son=7 (5 obs), 3 observations remain
  Primary splits:
      cond     < 1692.5 to the left,  improve=0.16032520, (3 missing)
      chloride < 40.4   to the left,  improve=0.15612030, (0 missing)
      tss      < 20.5   to the left,  improve=0.12762880, (3 missing)
      depth    < 14.5   to the left,  improve=0.08123499, (0 missing)
      alk      < 78.7   to the left,  improve=0.08061531, (0 missing)

Node number 6: 84 observations
  mean=5.478571, MSE=6.530493

Node number 7: 5 observations
  mean=11.62, MSE=67.5776
```

<u>Raw Data Regression Tree Model - CWAL:</u> %cyanobacteria vs specific conductance, alkalinity, total suspended solids, total phosphorus, total nitrogen, hardness, chloride, and depth:




```
Call:
mvpart(form = cyano ~ cond + alk + tss + tp + tn + hard + chloride,
   data = phyto_cwal, xval = 10, method = "anova", minsplit = 10,
   minbucket = 5)
  n= 74

         CP nsplit rel error    xerror      xstd
1 0.31917730      0 1.0000000 1.0262877 0.2379537
2 0.06687619      1 0.6808227 0.9738805 0.2232619

Node number 1: 74 observations,    complexity param=0.3191773
  mean=16.53703, MSE=631.5042
  left son=2 (51 obs) right son=3 (21 obs), 2 observations remain
  Primary splits:
     tn   < 0.835    to the left,  improve=0.3075767, (2 missing)
     tp   < 0.0425   to the left,  improve=0.1647702, (0 missing)
     cond < 921.8333 to the left,  improve=0.1481133, (6 missing)
     alk  < 89.1     to the left,  improve=0.1471797, (0 missing)
```

```
    hard < 259.5    to the left,  improve=0.1092909, (0 missing)

Node number 2: 51 observations
  mean=7.921569, MSE=270.5782

Node number 3: 21 observations
  mean=39.00667, MSE=857.9167
```