Abstract:

Attempts to determine the factors that influence an individual's decision to visit a lake are of increasing importance as more detailed statistics are collected for a large number of lakes across the United States. The present research seeks to use the recently compiled LAGOS dataset to create a functional model lake visitation, with a special attention paid to water clarity and the number of lakes nearby a given lake. This modeling found that both clarity and nearby lakes were significant predictors in determining whether a given lake is visited, as well as being significant factors in estimating total lake visitation. In addition, three models of water clarity were constructed with a focus placed on lake chemistry. The first focused directly on modeling lake clarity, using a linear mixed effects model to account for multiple measurements taken at a single lake. This model found significant negative effects for chlorophyll and phosphorus, as well as an effect for region in the US. The second model sought to determine factors that can be used to predict eutrophic status in 5 years for a given lake. This model found significant negative effects for initial water clarity, and positive effects for chlorophyll change and phosphorus change. The last model attempts to determine what features of a lake can serve as predictors as to whether or not the lake is measured for clarity. This model found that region of the United States and population surrounding the lake were significant predictors.

1. Introduction:

Lakes, rivers, and other publicly available bodies of water are important staples of society for many reasons. These water features present many opportunities for recreation, and also can be used as indicators of the health of the surrounding region. Despite these benefits, it is difficult to quantify the overall value that lakes provide. Recently, attempts have been made to extensively gather information about lakes across the United States in order to analyze various facets of lake value. One such initiative, the CSI-Limnology Project funded by the National Sanitation Foundation, have created the comprehensive LAGOS dataset on 141,271 lakes spanning 17 different states across the Midwestern and Northeastern United States[1]. The present research uses the LAGOS data to model two main indicators of lake value: visitation and water clarity.

Lake visitation is one of the more apparent determinants of lake value, because it directly measures the public valuation of a lake and can be translated into monetary benefit. On the surface, lake visitation seems rather easy to calculate. One possible approach is to conduct surveys at various lakes in order to physically count the number of visitors and also to determine factors that influence an individual's decision to visit the lake. However, there are several immediate problems with survey data. The first major issue is that surveys are difficult to administer from a resource perspective. Data must be physically acquired which would necessitate a researcher being at the lake in person to collect the data. Because of this, decisions must be made as to which lakes will be surveyed, which makes data privy to selection

[1] Patricia A. Soranno, Edward G. Bissell, Kendra S. Cheruvelil, Samuel T. Christel, Sarah M. Collins, C. Emi Fergus, Christopher T. Filstrup, Jean-Francois Lapierre, Noah R. Lottig, Samantha K. Oliver, Caren E. Scott, Nicole J. Smith, Scott Stopyak, Shuai Yuan, Mary Tate Bremigan, John A. Downing, Corinna Gries, Emily N. Henry, Nick K. Skaff, Emily H. Stanley, Craig A. Stow, Pang-Ning Tan, Tyler Wagner, Katherine E. Webster. 2015. **Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse.** GigaScience. 4:28 DOI: 10.1186/s13742-015-0067-4 (https://gigascience.biomedcentral.com/articles/10.1186/s13742-015-0067-4)

biases. Lakes that experience greater degrees of visitation would be more likely to be surveyed, which may affect survey results.

Recent research has attempted to ameliorate these issues with surveys by using data collected from social media. Keeler, et. al.[2] used a measure of lake visitation known as photo-user-days (PUD). PUD data draws on geotagged photos shared to the social media website Flickr. When a person takes any number of photos at a given lake on one day, that is one photo-user-day. Comparison between PUD data and recorded survey data for lakes in Iowa and Minnesota found that PUD is a good measure for lake visitation (Wood, et. al., 2013)[3]. This research uses PUD data collected for 51,107 lakes in the LAGOS dataset in order to determine what factors of a lake affect the probability that a particular lake is visited, as well as which lake features influence total visitation to the lake. The two main variables of interest are water clarity and the number of nearby lakes. Water clarity is the most notable feature of a lake, and can serve as an indicator of any immediate risks recreation at a particular lake might possess. The interest in nearby lakes is in determining whether people living in regions that are densely populated with lakes have different preferences than individuals in lake scarce regions who may not have as many options for lake recreation. One hypothesis is that people with a large number of nearby lakes would be more selective in the lakes they choose to visit. A consequence of this would be that preference of water clarity would be a more defining factor in lake dense regions, where recreational users have a wider selection of lakes to choose from.

The second metric for lake value this study seeks to evaluate is the clarity of the water. The LAGOS data uses secchi depth, the depth at which a black and white secchi disk cannot be seen when submerged in water, as a metric of clarity. In order to parse the various factors that influence water clarity, this research attempts to use data from the LAGOS dataset to model the various elements of a lake that can alter the clarity of the lake's water. The most widely studied factor of lake clarity is the chemistry of the water. Often, the presence of chemicals in the water can serve as indicators of potential pollution that has occurred. This study considers how the presence of various chemicals affects water clarity, after accounting for a variety of other factors including lake features and usage of the area surrounding the lake.

One manner of evaluating water clarity is to identify lakes that are classified as eutrophic. Eutrophic lakes are those that have a clarity of less than 1.83 meters (approx. 6 feet). The current study seeks to determine what factors influence a change in eutrophic status. As with the overall clarity model, our main variables of interest are the change in chemicals present in the lake.

A possible issue with this clarity data is the somewhat sporadic nature of the taken measurements. Many of the lakes do not have a recorded clarity measurement, while some lakes are measured many times each month. It is possible that there are qualitative differences between lakes that are measured and lakes that are not. The present research looks to determine these differences by modeling factors of a lake that can be used to predict whether or not a lake's clarity has been measured. Significant differences in measured lakes as compared

---

[2] Keeler, Bonnie L et al. "Recreational Demand For Clean Water: Evidence From Geotagged Photographs By Visitors To Lakes." *Frontiers In Ecology And The Environment*, vol 13, no. 2, 2015, pp. 76-81. *Wiley-Blackwell*, doi:10.1890/140124.

[3] Wood SA, Guerry AD, Silver JM, and Lacayo M. 2013. Using social media to quantify nature-based tourism and recreation. Scientific Reports 3: 2976.

to non-measured lakes may affect the ability of models that rely solely on observed clarity measurements to impute any missing clarity values.

2. Data:

Our primary index of lakes for analysis was lagoslakes_10400, which contained data on 141,271 lakes spanning 17 different states. This dataset possessed information for each lake such as lake area and perimeter, total number and area of upstream lakes, maximum depth, latitude and longitude, and location statistics such as state name, HUC4, and HUC12. Using the HUC4, an agricultural region was determined for each lake[4]. Low Agriculture is defined as the region where agricultural land usage is less than 10%. Lakes with HUC4 of 1,2,3,4,6,7,8,9,10, and 11 comprise this region. High Agriculture is defined as the region where agricultural land usage is greater than 75%. Lakes with HUC4 of 34, 50, 53, 56, 57, 61, and 63 comprise this region.

Lake visitation data was taken from lagos_osm_flickrpud_051817, which provides us with the information of total PUD, summer PUD, and lake amenities such as boat launch, toilets, hotels, marinas, bbqs, beaches, and shelters. Amenities data was obtained on 10/8/2016 using OpenStreetMap. Total PUD was calculated by taking the average yearly PUD for each lake from 2005 to 2014. Summer PUD is restricted to photo-user-days that occurred between June 15 and September 15 in the same years.

Population statistics and demographics were taken from HUC12pop. These data included total population, population living below the poverty line, population with at least a bachelor's degree, mean/median income, total households, Hispanic population, non-Hispanic white population, non-Hispanic black population, and median age. We calculated population percentages for population with a bachelor's degree, poor population, Hispanic population, white population, and black population by dividing the individual population statistic by the total population for the specific HUC12 region.Some HUC12s had multiple measurements in this dataset. To account for this, values for those particular HUC12s were averaged across all measurements.

Chemical and clarity measurements were collected from lagos_epi_nutr_10541. Chemistry data was selected based on chemicals that had more than 40,000 observations, which were chlorophyll (chla), phosphorus (tp), ammonium (nh4), nitrogen (tkn), and nitrogen oxide (no2no3). Clarity was measured in terms of secchi depth.

Surrounding land usage statistics were taken from lakes4ha_buffer500m_lulc. These data describe the categorization of land cover in a 500 meter buffer around the lake. The categories reported in this data set are Canopy, Open Water, Low Intensity Residential, Medium Intensity Residential, High Intensity Residential, Commercial/Industrial/Transportation, Barren (Rock/Sand/Clay), Quarries/Strip Mines/Gravel Pits, Transitional (Barren), Deciduous Forest, Evergreen Forest, Mixed Forest, Scrub/Shrub, Orchards/Vineyards/Other, Grasslands/Herbaceous, Pasture/Hay, Raw/Crops, Small Grains, Cultivated Crops, Urban/Recreational Grasses, Woody Wetlands, and Emergent Herbaceous Wetlands. The land usage statistics were reported for each of the years 1992, 2001, 2006, and 2011.

---

[4] Nelson, Eric, and Jesse Chung. "Using The LAGOS Database And Summer Photo User Day (PUD) Counts To Estimate The Impact Of Lake Quality And Lake Characteristics On Lake Recreation.."

3. Predicting Lake Visitation

3.1 Data

Lake visitation was measured in the form of Photo-User-Days (PUD) taken during the summer (June 15th to September 15th) during the period between 2005 and 2014. In order to retain a similar time frame for all variables, clarity measurements were restricted to this summer period from 2005 to 2014 as well. We used the average clarity measurement for each lake over this time period in our model. Additionally, surrounding land usage measurements were averaged for the years 2001, 2006, and 2011.

Lake amenities (hotels, beaches, boat launches, marinas, bbqs, toilets, and shelters) were coded as binary indicators, with "1" denoting lakes that possessed any number of a particular feature and "0" denoting lakes without the feature.

A "nearby lakes" statistic was calculated for all lakes in our dataset by applying an 80 km buffer around the lake and determining the number of lakes that fall within the buffer area. Distances were calculated by applying a Haversine transformation to the latitude and longitude data for each lake. Eighty km was chosen as the size for the buffer because it represents the range of travel for a day trip to visit a lake[5].

After consideration for the covariance of our predictor variables, mean income, maxdepth, white population percentage, lake perimeter, upstream lake count, total number of households, and median age variables were removed from analysis due to high multicollinearity with other variables ($r > 0.5$).

3.2 Methods

Previous literature has used linear models with a logarithmic transformation on non-zero PUD values to estimate lake visitation[6]. In our analysis, we compared this log-linear approach to a two-stage hurdle model in order determine which method better fits the PUD data.

In building the two-stage model, we separated the zero model component from the count model component in order to determine the type of model that best predicts our data. For the zero component, a binomial logit model was fit for the data, using the presence of PUD as the response (0 indicates no PUD, 1 indicates PUD of at least 1). Variable selection occurred during this stage of building the model in order to better identify significant variables in the zero/non-zero portion of the full hurdle model later. For the count component, the dataset was truncated to only lakes with a PUD value of at least 1. This reduced dataset was initially fit to a poisson model, which revealed a significant degree of overdispersion amongst the data. In order to correct for this, the truncated data was fit to a negative binomial model.

To determine the best two-stage model to use for our data, both a hurdle model and a zero inflated (ZINB) model were fit, using the reduced zero model constructed earlier and a full negative binomial model that used all of our predictor variables. Comparison of these two models using Akaike Information Criterion (AIC) found the hurdle model to be a better fit for the data ($Z = 4.207$, $p < .001$).

---

[5] Keeler, Bonnie L, et. al., 2015

[6] Keeler, Bonnie L, et. al., 2015

Both the two-stage hurdle model and a log-linear model were fit to the data using backwards variable selection using AIC and likelihood ratio comparisons. After both models had been created, the two were compared using leave-one-out cross validation, using sum of the absolute error as a measure. This cross validation found that the hurdle model was a better fit for the PUD data.

3.3 Results

Our model of lake visitation considered physical lake features, surrounding land usage, population demographics, lake amenities, and region as predictor variables. Model estimates for the zero component are given in Table 1 and estimates for the count component are given in Table 2.

	Log Odds of Lake Visitation	
	Estimate	(Standard Error)
Constant	1.7404	(1.337)
Latitude	-0.1653***	(0.026)
Longitude	0.0144	(0.010)
Clarity (m)	0.5748***	(0.135)
Low Agriculture	−0.3196	(0.377)
"Other" Agriculture	0.2800	(0.238)
log(Lake Area)	0.9965***	(0.040)
log(Area of Upstream Lakes)	0.0328**	(0.008)
log(Total Population)	0.3060***	(0.025)
Percentage of Population with Bachelor's Degree	0.0322***	(0.004)
Nearby Lakes	−0.0024***	(0.0004)
Presence of Boat Launch Features	0.5353***	(0.131)
Presence of Beaches	1.1361***	(0.280)
Percent Land Use: Pasture/Hay	−0.0179***	(0.004)
Percent Land Use: Cultivated Crops	−0.0238***	(0.003)
Percent Land Use: Emergent Herbaceous Wetlands	−0.0305***	(0.007)
Clarity * Low Agriculture	−0.5159***	(0.140)
Clarity * "Other" Agriculture	−0.4669***	(0.137)
Observations	4540	
Note:	Significance Levels: *p<0.05; **p<0.01; ***p<0.001	

Table 1: Model estimates for the binomial portion of the final hurdle model. One standard error is given in parentheses

	log-Lake Visitation (Photo-User-Days)	
	Estimate	(Standard Error)
Constant	1.3562	(0.956)
Latitude	−0.0722**	(0.024)
Longitude	0.0194*	(0.008)
Clarity (m)	0.2643***	(0.039)
Low Agriculture	−1.0865**	(0.377)
"Other" Agriculture	−0.3327	(0.236)
log(Lake Area)	0.5573***	(0.024)
log(Total Population)	0.0886***	(0.021)
Percentage of Population with Bachelor's Degree	0.0292***	(0.003)
Percentage of Population Below Poverty Line	0.0236***	(0.006)
Nearby Lakes	−0.0007	(0.001)
Presence of Beaches	0.3833***	(0.116)
Presence of Hotels	1.5903**	(0.485)
Presence of Boat Launch Features	0.2450**	(0.086)
Presence of Toilets	1.2022**	(0.234)
Percent Land Use: Developed, Medium Intensity	0.0481***	(0.007)
Clarity * Nearby Lakes	−0.0007***	(0.0001)
Nearby Lakes * Low Agriculture	0.0024	(0.002)
Nearby Lakes * "Other" Agriculture	0.0022*	(0.001)
log(Theta)	−0.7791***	(0.089)
Observations	4540	
Note:	Significance Levels: *p<0.05; **p<0.01; ***p<0.001	

Table 2: Model estimates for the truncated negative binomial count portion of the final hurdle model. One standard error is given in parentheses. The nearby lakes variable is measured by tens.

Overall, we see regional effects for agricultural region as well as significant effects for latitude and longitude. Analysis of model residuals using Moran's I found significant spatial correlation, (Observed I = 0.013, $p < .001$) indicating that these location variables do not account for all of the geographical influence in the data. However, the observed Moran's I is relatively small, indicating that the effect of any spatial correlation is weak and may not have any influence on the interpretation of our model.

Our two primary variables of interest, water clarity and nearby lakes, were significant in both components of the model, as well as interacting with each other in the count component. The effects of these two predictors are given in Figure 1 and Figure 2. We also see interactions between these variables and agricultural region, so these figures also identify any regional effects. Additionally, the effect of the interaction between water clarity and nearby lakes is given in Figure 3.

Additionally, total lake area and HUC12 population were both significant factors in the zero and count components of the model. From this we get that larger lakes are more likely to be visited and experience a greater degree of visitation than smaller lakes, and lakes in densely populated areas are more likely to be visited and will experience a greater degree of visitation than lakes in less populated regions.

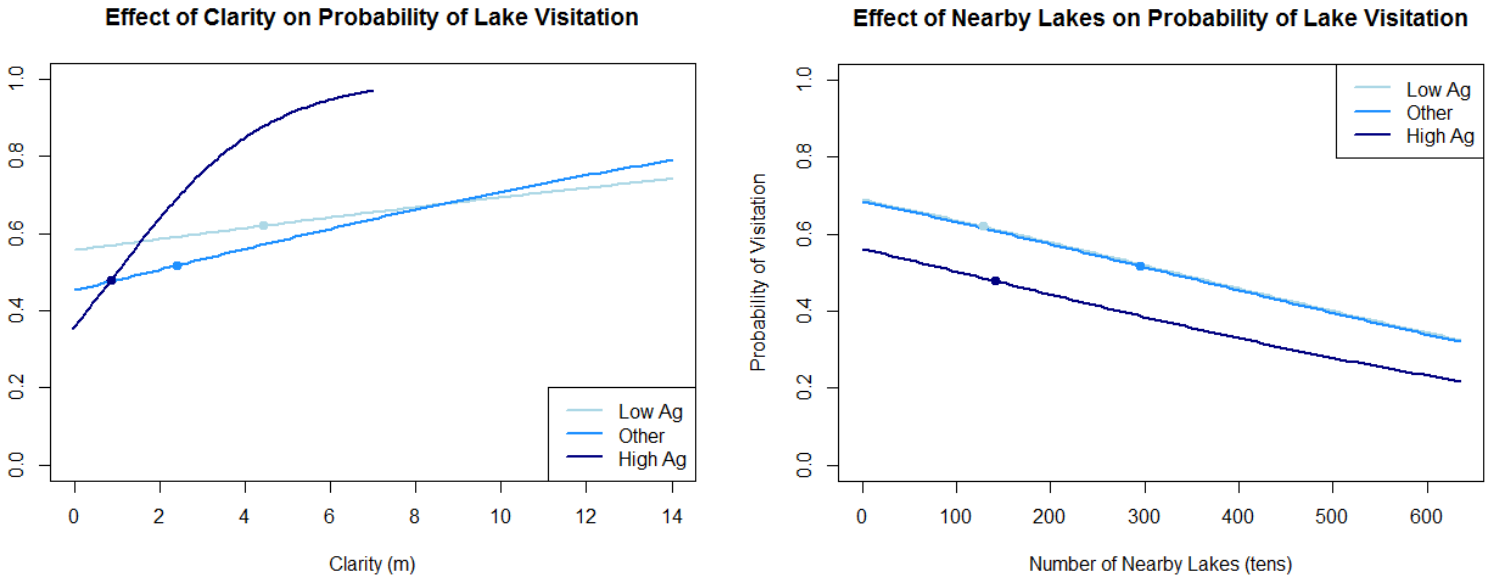


Figure 1: Effects plots for clarity and nearby lakes for the zero component of the hurdle model, grouped by region. Each line assumes median estimates for all other model predictors, calculated for each agricultural region. Points indicate the median observed value for the plotted variable, clarity on the left and nearby lakes on the right.

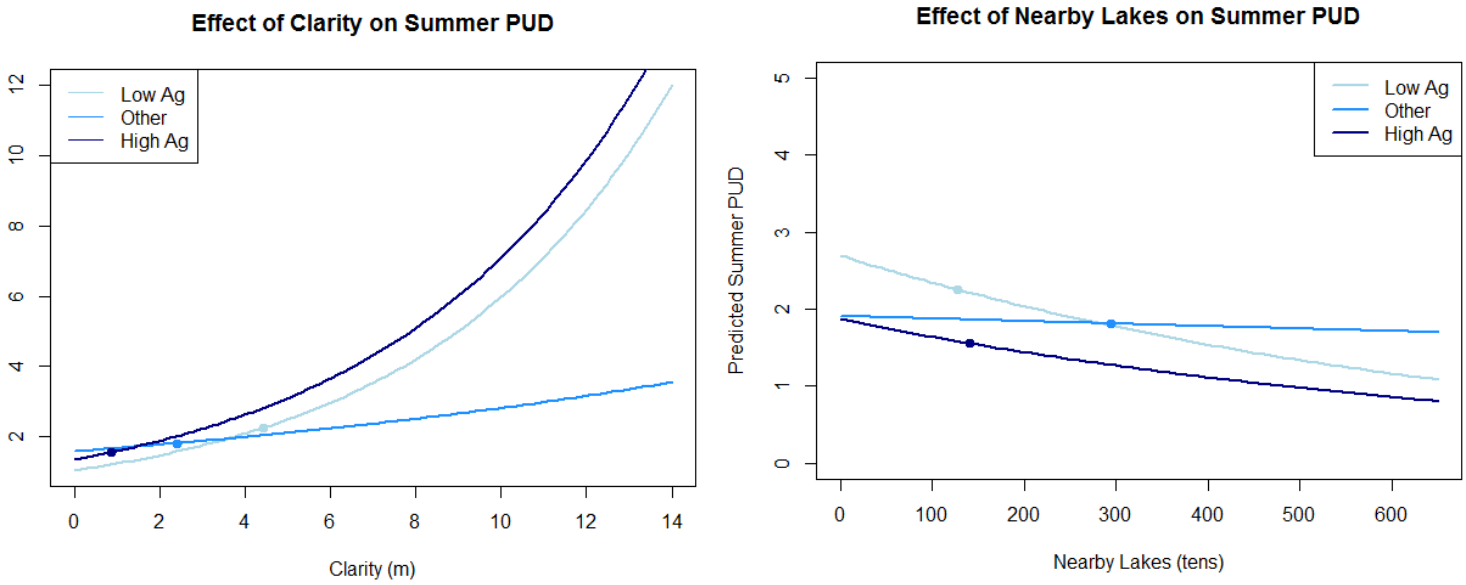


Figure 2: Effects plots for clarity and nearby lakes for the count component of the hurdle model, grouped by region. Each line assumes median estimates for other model predictors, calculated for each agricultural region. Points indicate the median observed value for the plotted variable, clarity on the left and nearby lakes on the right.

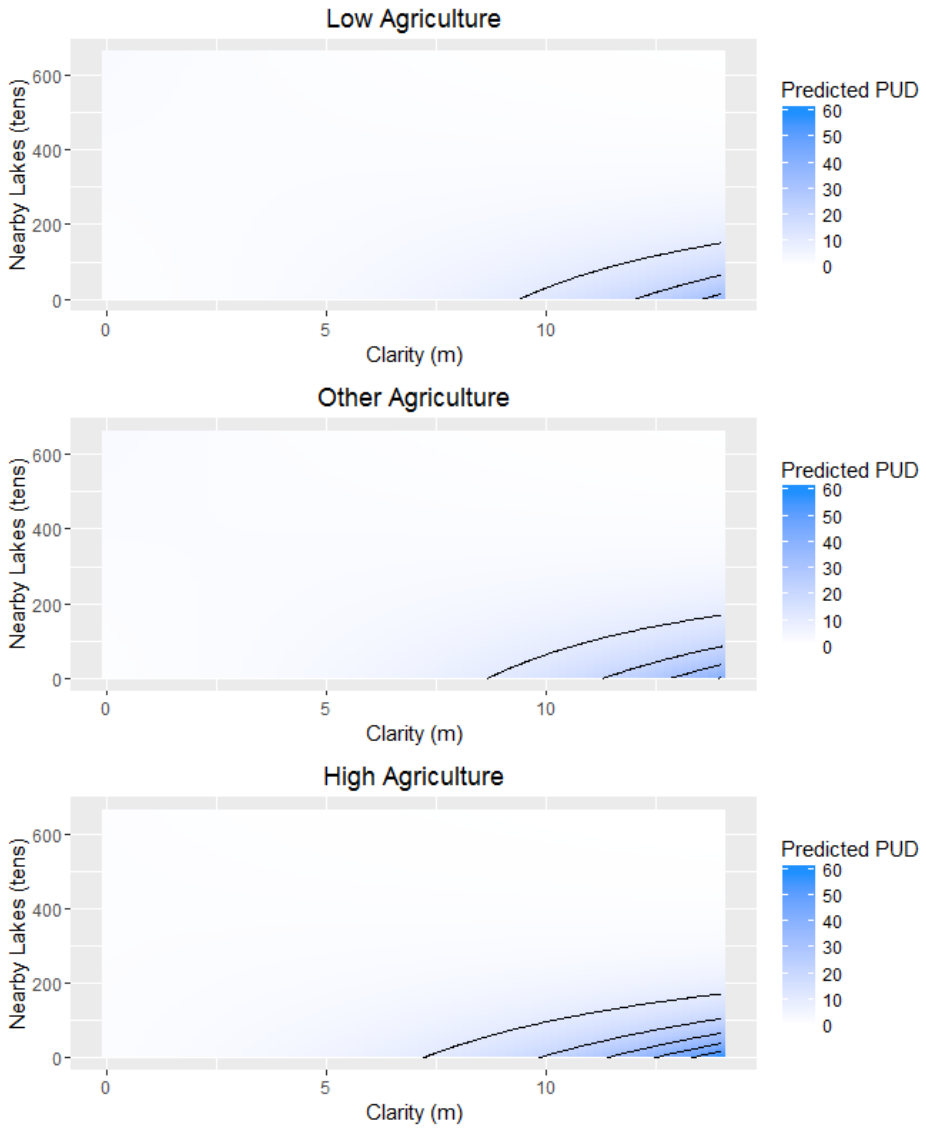


Figure 3: Effect of clarity and nearby lakes on predicted PUD, using the count component of the hurdle model. Contour lines are drawn every 10 PUD. Each plot assumes median estimates for all other model predictors.

3.4 Discussion

Our model found that there is a significant positive effect of water clarity on lake visitation. We would expect that a 1 meter increase in water clarity would increase the odds that a lake is visited by 6.07% for a low agriculture lake, 11.39% for an “other” agriculture lake, and 77.68% for a high agriculture lake (Figure 1, left). In terms of lake preference, people are more likely to visit a lake that has clear water as compared to a lake with murky water (Figure 2, right). When looking at lakes that are visited, this 1 meter increase in water clarity would increase overall lake visitation by 19.11% for a low agriculture lake, 5.97% for an “other” agriculture lake, and 18.03% for a high agriculture lake, assuming median number of nearby lakes for each agricultural region (Figure 2, left). Again, we see this positive effect for water

clarity. Not only are clearer lakes more likely to be visited that murky lakes, but of lakes that are visited, clear lakes are visited the most.

One important factor that affects water clarity is the number of nearby lakes. There is a negative interaction between nearby lakes and clarity for lakes that have at least one visit, suggesting that as the number of nearby lakes increases, the impact of clarity decreases, which we can see in Figure 3. This would seem to suggest that when there are many lakes to choose from, people don't rely as strongly on water clarity to guide lake preference.

We also see a negative effect for the nearby lakes variable by itself, indicating that when there are many lakes near to each other, the visitation to any one lake decreases. Although this seems intuitive, it is interesting to note that this effect is present in both stages of the model. This means that lakes in regions with a high lake density are less likely to be visited, and among lakes that are visited, high density lakes are visited less.

It should be noted that visits to lakes in high agriculture regions are more affected by changes in clarity than lakes in other regions, as evidenced by Figures 1 and 2. We can see in these figures that equal changes in water clarity will increase the probability that a lake is visited to a greater extent in high agriculture regions than either of the other two regions. This finding reveals an interesting difference in lake preference: people living in high agriculture regions place a greater importance on lake clarity than people in other regions. One possible explanation for this is that the median water clarity in the high ag region is much lower than either of the other two regions, making clear lakes more scarce and therefore more valuable to lake visitors.

Other significant factors that influence lake visitation are select lake amenities such as beaches, toilets, hotels, and boat launches. Beaches and boat launches were significant predictors of visitation, indicating that lakes that possess beaches and boat launches are more likely to be visited than those lakes without. All four amenities had positive effects on overall visitation, which tells us that lakes with any of these features are visited more than those without.

4. Predicting Water Clarity

4.1 Data

Land usage statistics were calculated for each lake by averaging the land use measurements collected in 1992, 2001, 2006, and 2011. Additionally, a land usage change statistic was calculated by taking the difference in land use between 1992 and 2011.

No lake amenities or HUC12 population demographics were used in this model of water clarity.

4.2 Methods

A pilot model was constructed for individual lake clarity measurements using chemistry data as predictors, in order to identify which chemicals are influential when predicting water clarity. Since multiple measurements were taken for many of the lakes, a mixed effects model was used with a random effect for lake id. This model found that chlorophyll (chla) and phosphorus (tp) were significant.

Exploratory data analysis revealed heavy right skews on the clarity, chlorophyll, and phosphorus variables, so a logarithmic transformation on all three variables was performed.

Our model of lake clarity considered physical lake features, location and region identifiers, surrounding land usage, and water chemistry as predictor variables. Model estimates are given in Table 3.

A full linear mixed effects model was constructed using a log-transformed water clarity as the response variable. The dataset used was restricted to only lakes that had values for every variable initially considered. This reduced the size of the dataset from 224,759 observations to 62,268 observations. We initially considered random effects for lake id, chlorophyll, and phosphorus. However, there were not enough repeated measurements of chlorophyll and phosphorus within each lake to adequately model a lake-specific chlorophyll or phosphorus effect, so we ended up only using a random effect for lake id. An interaction between each of the chemicals and agricultural region was considered for this model.

After constructing the full linear mixed effects model and eliminating insignificant predictors, we obtained a model with the predictor set shown in Table 3. In order to assess the validity of model estimates, our final model was refit to the original dataset of 224,759 observations.

4.3 Results

Analysis revealed significant effects for agricultural region, as well as latitude and longitude. A calculation of spatial correlation was conducted using Moran's I for the model residuals. We found significant spatial correlation between residuals, Observed I = 0.05320, $p < .001$. However, the observed Moran's I is relatively small, indicating a weak effect for spatial correlation, which may not have any practical implications on the interpretation of our model.

We have significant effects for our two chemicals of interest, chlorophyll and phosphorus, as well as a significant interaction with agricultural region for both chemicals. Effects of these variables are given in Figure 4 and Figure 5.

To test the validity of our model, the final model was fit onto both the restricted dataset used during the initial modeling process and a full dataset containing all observations for every

lake. Confidence intervals were constructed for each variable for both models, which are given in Table 4. Comparison of the two models revealed that the estimates are not significantly different from each other, indicating that our model is a good fit for the data.

	Dependent variable:
	log(Secchi)
Random Effects:	
Lake Id:	Var = 0.1041 (0.3184)
Residual:	Var = 0.1159 (0.3404)
Fixed Effects:	
Latitude	0.0788 (0.0031)
Longitude	0.0109 (0.0012)
log(Chlorophyll)	−0.2171 (0.0066)
log(Phosphorus)	−0.3778 (0.0109)
agLow Ag	−0.5078 (0.0611)
agOther	−0.2543 (0.0523)
log(Upstream Lakes Count)	−0.0115 (0.0016)
log(Lake Area)	0.0301 (0.0045)
Percent Canopy Coverage	−0.0017 (0.0004)
Percent Land Use Commercial/Industrial	−0.0130 (0.0023)
Percent Land Use Pasture/Hay	−0.0046 (0.0010)
Percent Land Use Woody Wetlands	−0.0060 (0.0006)
Percent Land Use Emergent Herbaceous Wetlands	−0.0067 (0.0010)
log(Chlorophyll):agLow Ag	0.0525 (0.0085)
log(Chlorophyll):agOther	0.0874 (0.0069)
log(Phosphorus):agLow Ag	0.1872 (0.0139)
log(Phosphorus):agOther	0.0273 (0.0115)
Constant	−0.1555 (0.154)
Observations	74,429
Note:	Coefficients are given with one standard error in parentheses

Table 3: Estimates for the linear mixed effects model of lake clarity. All variables in the model were significant at the

	small 2.5%	small 97.5%	big 2.5%	big 97.5%
nhd_lat	0.071	0.085	0.076	0.087
nhd_long	0.009	0.014	0.009	0.013
log(chla + 0.001)	-0.226	-0.206	-0.228	-0.203
log(tp + 0.001)	-0.397	-0.360	-0.405	-0.369
agLow Ag	-0.632	-0.391	-0.674	-0.417
agOther	-0.360	-0.161	-0.378	-0.187
log(lakes4ha_upstreamlakes_4ha_count + 0.001)	-0.015	-0.008	-0.014	-0.007
mean23	-0.018	-0.008	-0.015	-0.009
mean81	-0.006	-0.003	-0.005	-0.001
mean91	-0.007	-0.005	-0.007	-0.005
mean95	-0.009	-0.004	-0.008	-0.004
log(chla + 0.001):agLow Ag	0.037	0.068	0.037	0.063
log(chla + 0.001):agOther	0.075	0.099	0.069	0.094
log(tp + 0.001):agLow Ag	0.159	0.213	0.170	0.222
log(tp + 0.001):agOther	0.006	0.050	0.020	0.057

p < .05 level. This model uses the restricted dataset mentioned in

Table 4: Confidence intervals for model estimates using a restricted dataset (small) and a full dataset (big). Only variables that were found to be significant during modeling are included

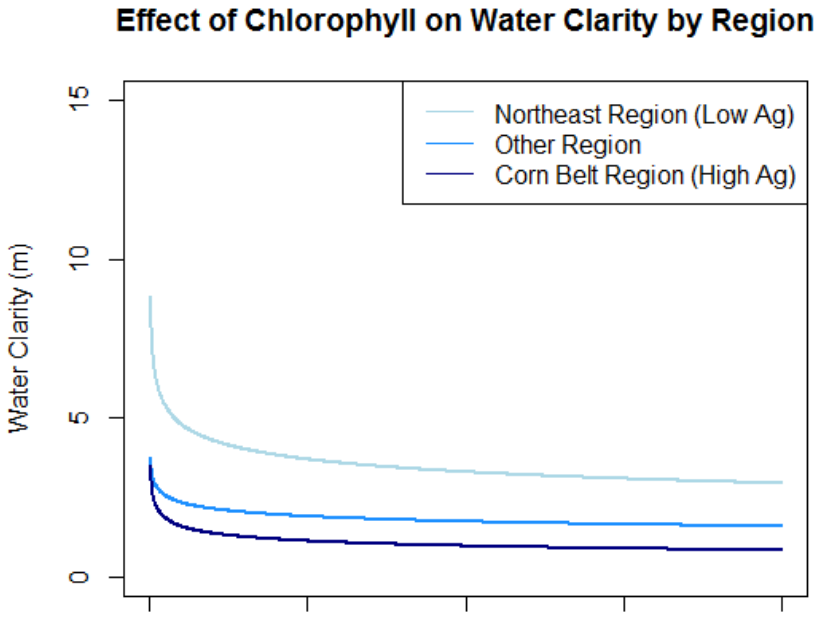


Figure 4: Effect plot of chlorophyll on water clarity, grouped by agricultural region. Each line assumed median levels of all model variables, calculated for each region independently.

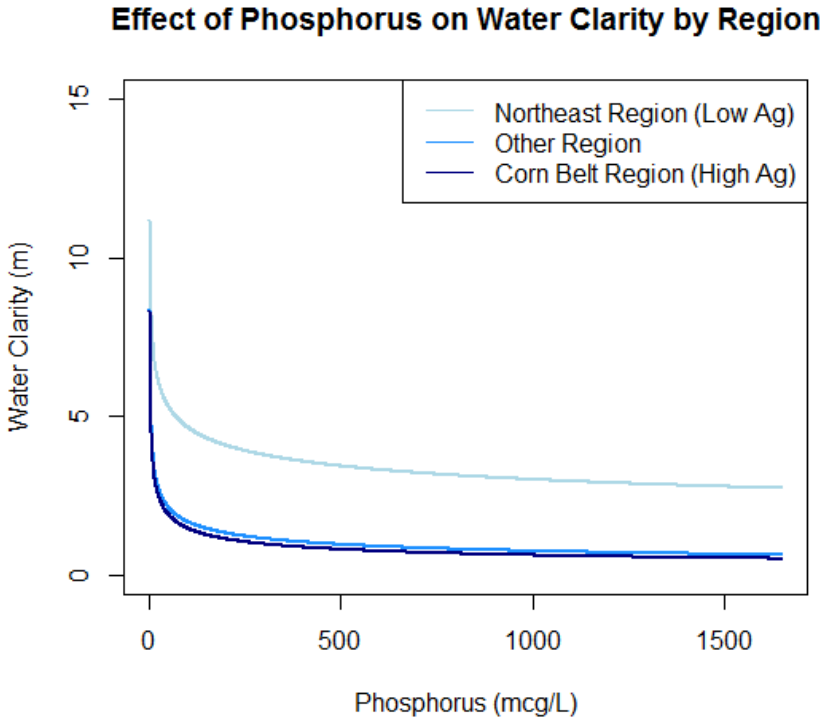


Figure 5: Effect plot of phosphorus on water clarity, grouped by agricultural region. Each line assumed median levels of all model variables, calculated for each region independently.

4.4 Discussion

Our model found significant negative effects for both chlorophyll and phosphorus, indicating that an increase in levels of either chemical result in a decrease of water clarity. Additionally, there was a significant positive interaction for both these chemicals with agricultural

region. This finding suggests that, for lakes in Low and “Other” agricultural regions, chlorophyll and phosphorus do not have as strong of an effect as in high agricultural regions. We can see this in both Figure 4 and Figure 5, where the estimates for high agriculture lakes appear to have a lesser slope than the estimates for either low or other agriculture lakes. One possible explanation for this is that lakes in high agricultural regions have a lower water clarity than either of the other two regions, which would make it more difficult for an increase in either chlorophyll or phosphorus to have an impact on the already poor lake clarity.

Additionally, we found significant negative effects for four different types of land usage: commercial/industrial, pasture/hay, woody wetlands and emergent herbaceous wetlands. These findings seem straightforward. Land used for both commercial/industrial and pasture/hay would increase the amount of pollutants introduced into the lake, while the wetlands would deposit tannins into the water. Tannins occur naturally in the bark of trees found in wetlands, and are colored brown, which would decrease the clarity of the lake water.

5. Predicting Change of Eutrophic Status

5.1 Data

We considered average clarity for lake measurements collected for the summer months (June 15th - Sept 15th) from 2000 to 2002, and again for summer months (June 15th - Sept 15th) from 2005 to 2007. A binary indicator for lake eutrophic status in 2001 and 2006 was created using these average clarity measurements; a 1 indicates a eutrophic lake (average clarity less than 1.83m) and a 0 indicates a non-eutrophic lake.

All variable measurements were restricted to the same time frame as our clarity measurements. Chlorophyll and phosphorus values were averaged for the 2000-2002 and 2005-2007 periods. A change value was calculated for each chemical by taking the difference between the observed levels for each period (2005-2007 minus 2000-2002).

Land usage statistics were measured in 2001 and 2006. A change of land usage variable was constructed for each classification of land usage by taking the difference between the two measurements. One classification, Evergreen Forest, was found to have a change value of 0 for every lake in the dataset and so the variable was removed from our analysis.

5.2 Methods

The data were divided into two groups based on eutrophic status in 2000-2002. This resulted in a dataset of 618 lakes that were eutrophic during the initial period and 305 lakes that were non-eutrophic. Lakes in the eutrophic dataset were restricted to having a depth of at least 1.83 meters.

A binomial logistic model was created for each dataset, using eutrophic status in 2006 as the response variable. Both models used the same initial set of predictor variables, but variable selection occurred separately.

Exploratory data analysis of empirical log odds for each predictor variable revealed that a quadratic term for both change in chlorophyll and change in phosphorus was appropriate for the models. Variables included in the initial models are change in chlorophyll, change in phosphorus, initial mean clarity in 2000-2002, latitude/longitude and agricultural region, change in land usage statistics, total area of the lake, total number of upstream lakes, and maximum water depth.

5.3 Results

Table 5 gives model estimates for lakes that were not eutrophic in 2000-2002. This model found that initial clarity in 2001 and change in chlorophyll were significant predictors of eutrophic status in 2006. These effects are shown in Figure 6.

There were no significant effects for agricultural region or latitude/longitude. Model residuals were analyzed for spatial correlation using Moran's I, revealing a significant correlation (Observed I = 0.0262, $p < .001$). Although the spatial correlation was significant, the observed Moran's I had a small effect size indicating a weak spatial correlation, which may have little practical influence on the variance of model estimates.

	Log Odds of Eutrophic Status in 2005-2007:	
	Estimate	Standard Error
Initial Clarity	−1.365	(0.298)
Change in Chlorophyll	0.162	(0.032)
Constant	1.131	(0.822)
Observations	618	
Note:	Lakes used for this model were not eutrophic in 2000-2002	

Table 5: Estimates for the binomial logistic model for non-eutrophic lakes in 2000-2002. One standard error is given in parentheses

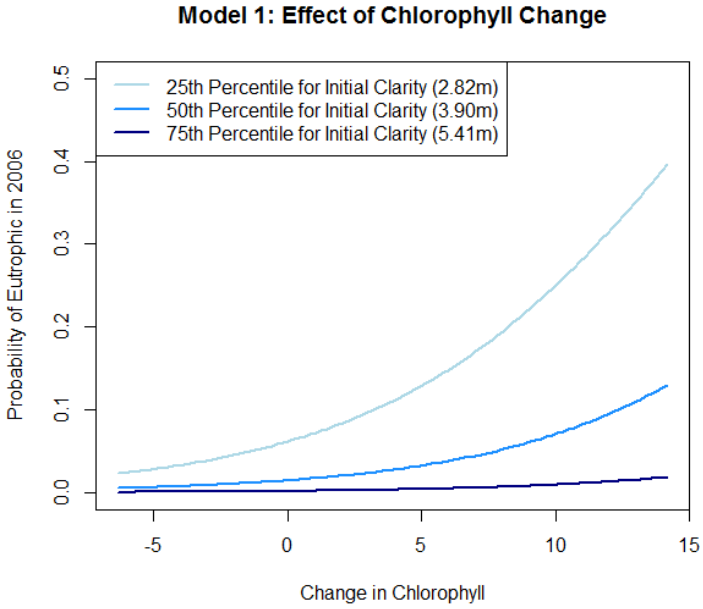


Figure 6: Effect of chlorophyll change on the probability of eutrophic status in 2006, using non-eutrophic lakes in 2001. Effects are grouped by quartiles of initial clarity in 2001.

Model estimates for lakes that were eutrophic in 2001 are given in Table 6. This model found significant effects for initial clarity, change in chlorophyll and change in phosphorus, as well as a significant quadratic effect for both chemical change variables. Effects these variables are given in Figure 7.

No significant effects for latitude/longitude or agricultural region were found. Analysis of spatial correlation within model residuals revealed significant correlation (Observed I = 0.0322, p = .002). However, the observed Moran's I is relatively small which suggests a weak effect of spatial correlation. As such, there may be no substantial influence of this correlation on model interpretations.

	Log Odds of Eutrophic Status in 2005-2007:	
	Estimate	Standard Error
Initial Clarity	−4.406	(0.770)
Change in Chlorophyll	0.103	(0.031)
(Change in Chlorophyll)^2	0.002	(0.001)
Change in Phosphorus	0.045	(0.020)
(Change in Phosphorus)^2	0.0002	(0.0002)
Constant	7.411	(1.196)
Observations	305	
Note:	Lakes used for this model were eutrophic in 2000-2002	

Table 6: Estimates for the binomial logistic model for eutrophic lakes in 2000-2002. One standard error is given in parentheses.

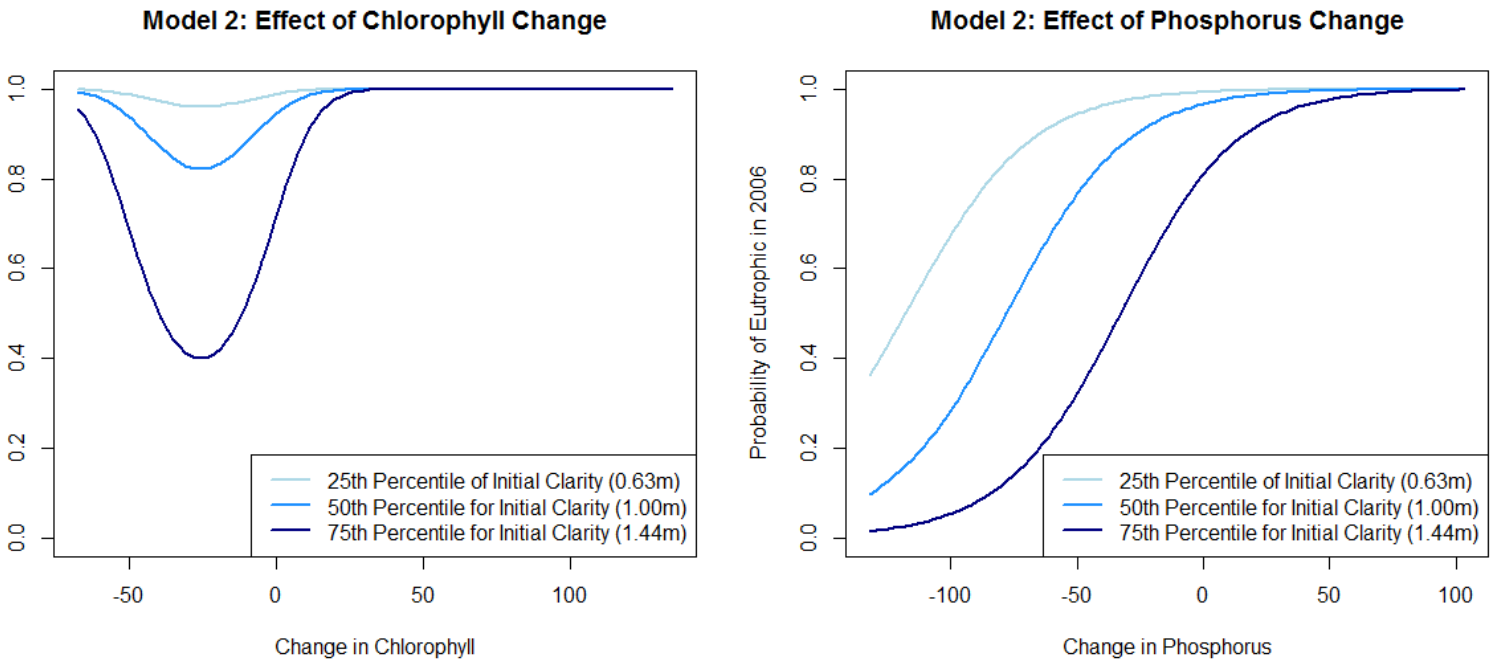


Figure 7: Effects of change in chlorophyll (left) and change in phosphorus (right) on the probability of being eutrophic in 2006, using eutrophic lakes in 2001. Effects are grouped by quartile of initial clarity in 2001.

5.4 Discussion

Both models found significant negative effects for initial clarity in 2001. For non-eutrophic lakes in 2001, a 1 meter increase in initial clarity results in a 74.5% decrease in the odds that a given lake will be eutrophic in 2006. For eutrophic lakes in 2001, this 1 meter increase in initial clarity results in a 36.6% decrease in the odds that a given lake will be eutrophic in 2006. This finding is unsurprising, since lakes that have initial clarity levels closer to the eutrophic cutoff of 1.83 meter would be more likely to experience a change in eutrophic status. We can see this effect in both Figure 6 and Figure 7, where lakes at higher quartiles of initial clarity have the lowest probability of being eutrophic in 2006.

We also have significant positive effects for change in chlorophyll in both models, and a positive effect for change in phosphorus when lakes were eutrophic in 2001. From Figure 7, we can see that large changes in chlorophyll, regardless of the change being positive or negative, result in high probabilities of being eutrophic in 2006. This occurs because we have a significant quadratic effect for chlorophyll change. It is important to note that there is also a significant quadratic effect for change in phosphorus, but it has a much smaller effect size. Because of this, we do not see an increase in the probability that a lake will be eutrophic in 2006 at large negative changes in phosphorus.

The finding that chemical changes have significant effect on the probability of a lake being eutrophic is supported by our earlier model of water clarity, where both chlorophyll and phosphorus were significant factors for predicting water clarity. Using this earlier model, we would also predict that both changes in chlorophyll and phosphorus would have a positive effect on the probability of being eutrophic in 2006 because both of these chemicals were found to have a negative effect on water clarity.

6. Predicting Missing Clarity Measurements

6.1 Data

In order to better predict whether or not a lake is measured for water clarity, we restricted our data to only the 51,107 lakes that were recorded with PUD data. Additionally, any lakes that were missing population data or land usage statistics were removed. This left us with a total of 48,873 lakes for analysis.

In an attempt to correspond with the time frame of the PUD data, clarity measurements were limited to only those take between June 15 and September 15, since the year 2005. A binary indicator was created for whether a lake was missing clarity measurements. A 1 indicates that a lake was not measured for clarity during this time period (missing); a 0 indicates that the lake had at least one measurement. Out of all lakes in the PUD data set, only 5,609 (11%) of these lakes have at least one clarity measurement during the summer since 2005.

Lake perimeter, upstream lake count, and percent of land used by deciduous forest variables were removed from analysis due to high multicollinearity with other variables. Also, the max depth variable was removed from analysis due to large amount of NA's.

Dummy variables were created for three of the land use classifications due to a large proportion of observed 0 values: Developed Medium Intensity, Developed High Intensity, Barren (Rock/Sand/Clay) variables. For these variables, a 1 indicated that the surrounding area contained some portion of land dedicated to these classifications, and a 0 indicates no land used for each.

Exploratory data analysis revealed heavy right skews on the total population, lake area, and percentage of canopy variables, so a logarithmic transformation was performed.

6.2 Methods

A binomial logistic model was fit to our data, using the indicator of missing clarity as the response. This model physical lake features, location and region identifiers, surrounding land usage, and demographic info as predictor variables. An interaction with agricultural region was considered for the four main variables of interest: lake area, population, canopy, and upstream lake area. Backwards selection was then performed to eliminate all the insignificant variables.

6.3 Results

Model estimates for missing clarity are given in Table 7. This model found significant effects for three of our variables of interest: total population, lake area, and percentage of canopy cover. Moreover, there were significant interactions with agricultural region for all three of these variables. The effects of population, lake area, and canopy cover in the model are given in Figure 8, Figure 9, and Figure 10 respectively.

The model also found significant effects for many different types of land usage, such as all four types of developed land, evergreen and mixed forests, and woody and emergent herbaceous wetland.

	log-odds of missing clarity	
	Estimate	(Standard Error)
log(Total Population)	−0.357***	(0.038)
log(Lake Area)	−1.230***	(0.051)
log(Canopy)	−0.120***	(0.041)
Northeast Region	−7.462***	(0.594)
Other Region	−0.383	(0.395)
Percent White Population	−0.016***	(0.002)
Avg. of Med. Household Income	−9.206e-07**	(4.496e-07)
Percent Land: Open Water	0.016***	(0.003)
Percent Land: Developed, Open Space	−0.028***	(0.003)
Percent Land: Developed, Low Intensity	−0.006**	(0.003)
Contain Land: Developed, Medium Intensity	−0.192***	(0.048)
Contain Land: Developed, High Intensity	0.225***	(0.059)
Contain Land: Barren (Rock/Sand/Clay)	0.461***	(0.053)
Percent Land: Evergreen Forest	0.006***	(0.002)
Percent Land: Mixed Forest	−0.011***	(0.002)
Percent Land: Scrub/Shrub	0.014**	(0.006)
Percent Land: Pasture/Hay	−0.012***	(0.002)
Percent Land: Woody Wetlands	0.014***	(0.002)
Percent Land: Emergent Herbaceous Wetland	0.028***	(0.003)
Longitude	0.091***	(0.005)
Latitude	−0.065***	(0.011)
log(Total Population):Northeast Region	0.356***	(0.041)
log(Total Population):Other Region	0.087**	(0.038)
log(Lake Area):Northeast Region	0.380***	(0.061)
log(Lake Area):Other Region	0.006	(0.053)
log(Canopy):Northeast Region	0.534***	(0.102)
log(Canopy):Other Region	−0.182***	(0.053)
Constant	21.731***	(0.610)
Observations	48,158	
Note:	Significance Levels: *p<0.05; **p<0.01; ***p<0.001	

Table 7: Model estimates for the binomial model for missing clarity. One standard error is given in parentheses.

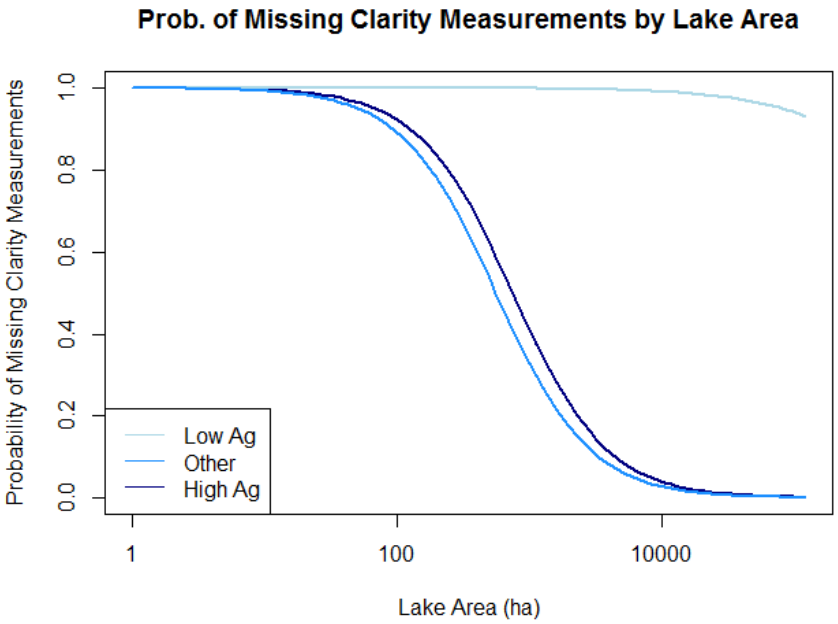


Figure 8: Effect of lake area change on the probability of whether or not a lake misses clarity measurement. Effects are grouped by different agricultural classifications. Each line assumed median levels of all model variables, calculated for each region independently.

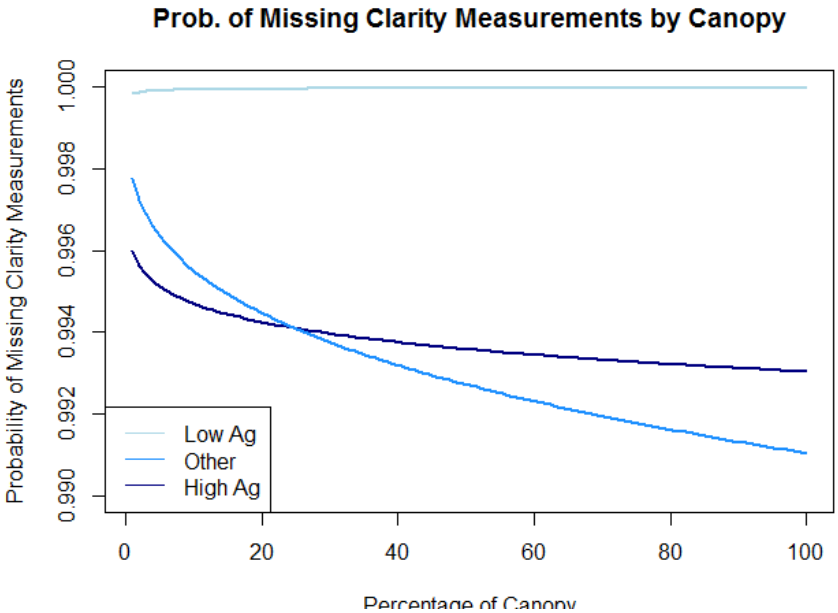


Figure 9: Effect of percentage of canopy cover change on the probability of whether or not a lake misses clarity measurement. Effects are grouped by different agricultural classifications.Each line assumed median levels of all model variables, calculated for each region independently.

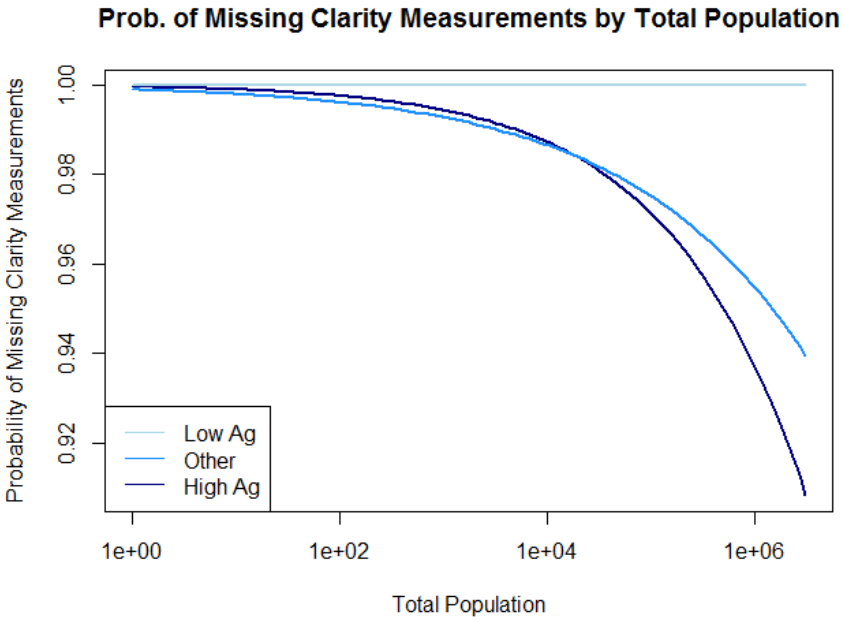


Figure 10: Effect of total population change on the probability of whether or not a lake misses clarity measurement. Effects are grouped by different agricultural classifications. Each line assumed median levels of all model variables, calculated for each region independently.

6.4 Discussion

According to our model, there appears to be an increased probability of missing clarity measurement for northern/western lakes, based on the significance of latitude and longitude. Figure 11 shows the distributions of lakes that have been measured and lakes that are missing clarity measurements during the time period under consideration.

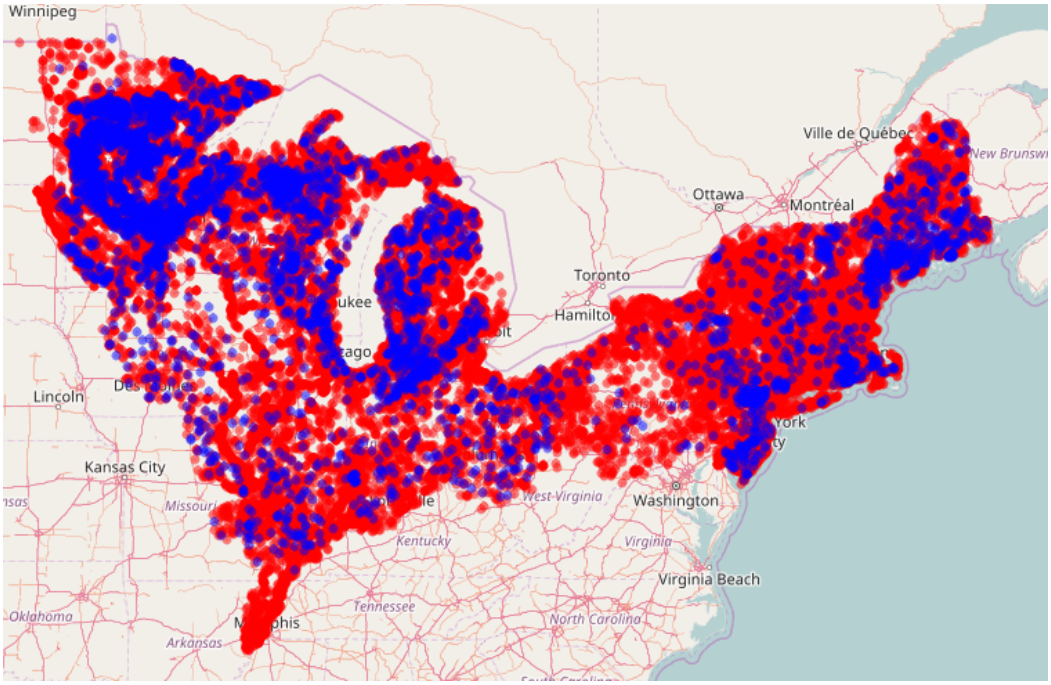


Figure 11: Map of all lakes used during modeling. Red indicates lakes that are missing clarity measurements; blue indicates lakes that have been measured.

Our model found a significant negative effect for lake area, which indicates that larger lakes are more likely to have clarity measurements. However, due to the presence of an interaction between area and agricultural region, the effect of lake size varies across the regions: for high agricultural region, doubling lake area will result in 57% decrease in the odds the lake is missing a clarity measurements; for low agricultural region, this doubling of lake area decreases the odds of missing clarity by 45%; for other agricultural region, a doubling of area decreases the odds of missing clarity by 57%.

Additionally,the model suggests that total population has a negative effect on missing clarity. According to our findings, lakes in densely populated areas are less likely to miss clarity measurements. We would expect that doubling the HUC12 population size around a lake would decrease the odds that a lake is missing clarity measurement by 22% for a high agricultural region lake, 0.1% for a low agricultural region lake, and 17% for an “other” agricultural region lake.

The third variable of interest that was significant in the model was percentage of canopy coverage for a lake. Canopy was found to have a negative effect on missing clarity measurements, which indicates that lakes with a larger degree of canopy are less likely to be missing clarity measurements. It should be noted that, compared to lake area and population, canopy had a relatively low effect size, resulting in a smaller impact on the probability that a lake is missing a clarity measurement. We can see this effect in Figure 9 above.

An interesting result from this model was that a variety of land usage classifications were significant, with both positive and negative effects. We see positive effects for Open Water, Evergreen Forest, Scrub/Shrub, Woody Wetlands, and Emergent Herbaceous Wetland, or containing High Intensity Residential Area or Barren (Rock/Sand/Clay) which suggest that higher percentages of land used for these categories would increase the odds that a lake would

not be measured for clarity. Alternatively, we have negative effects for Open Space, Low Intensity Residential Area, Mixed Forest, Pasture/Hay, or containing Medium Intensity Residential Area, which would indicate that higher percentages of land used by these categories would decrease the likelihood that a lake would be missing a clarity measurement.

Appendix A: Exploratory Data Analysis for Lake Visitation Model

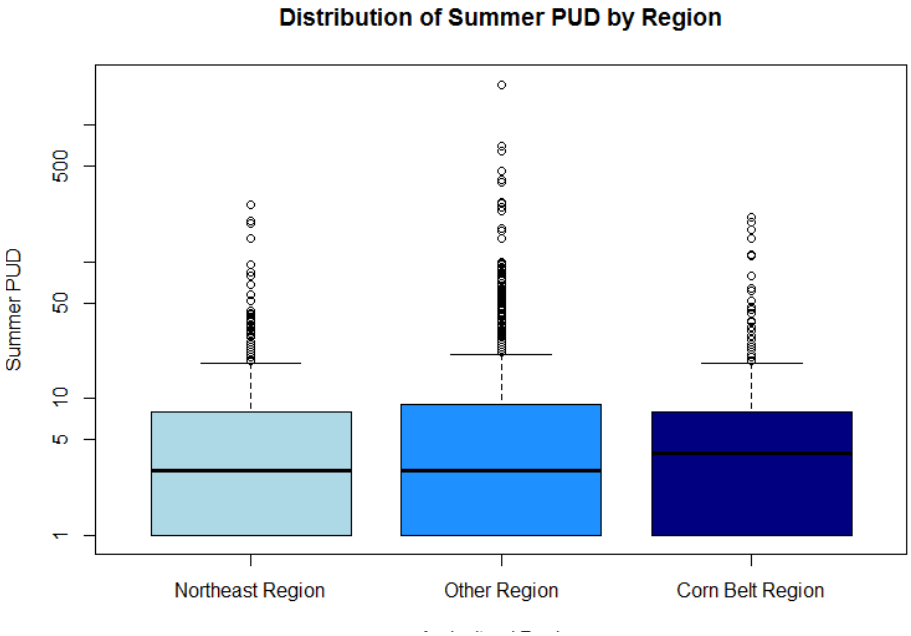


Figure A-1: Boxplot of Summer PUD grouped by agricultural region. Data for this plot was taken from the truncated dataset (Lakes with PUD > 0).

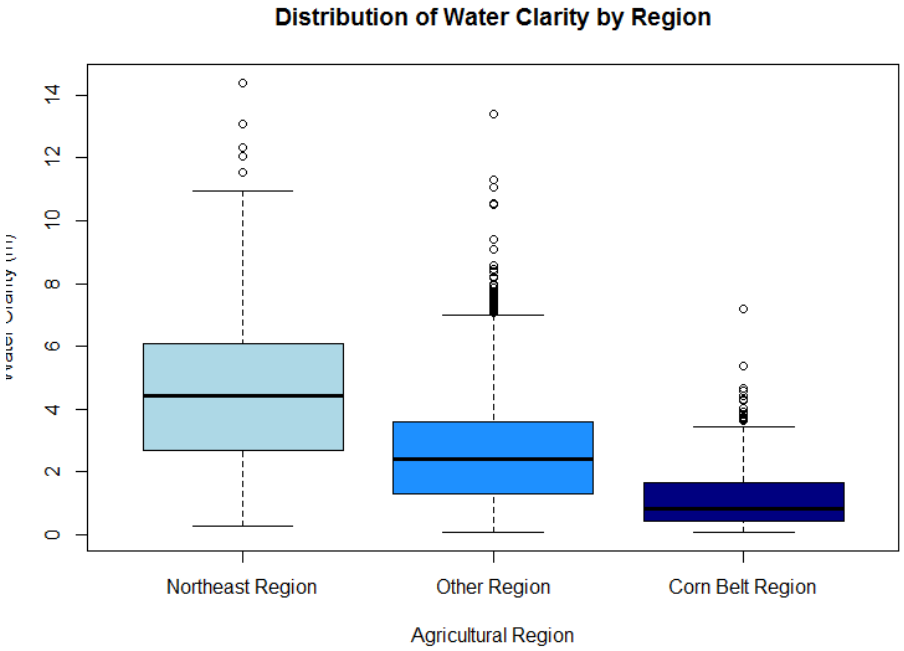


Figure A-2: Boxplot of water clarity grouped by agricultural region. Data for this plot was taken from the full dataset.

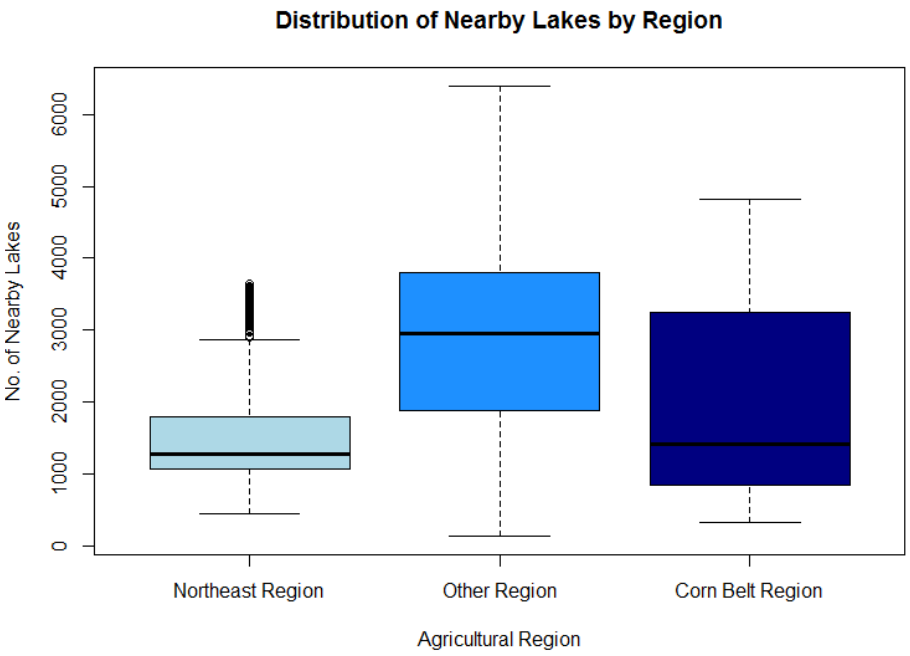


Figure A-3: Boxplot of nearby lakes grouped by agricultural region . Data for this plot was taken from the full dataset.

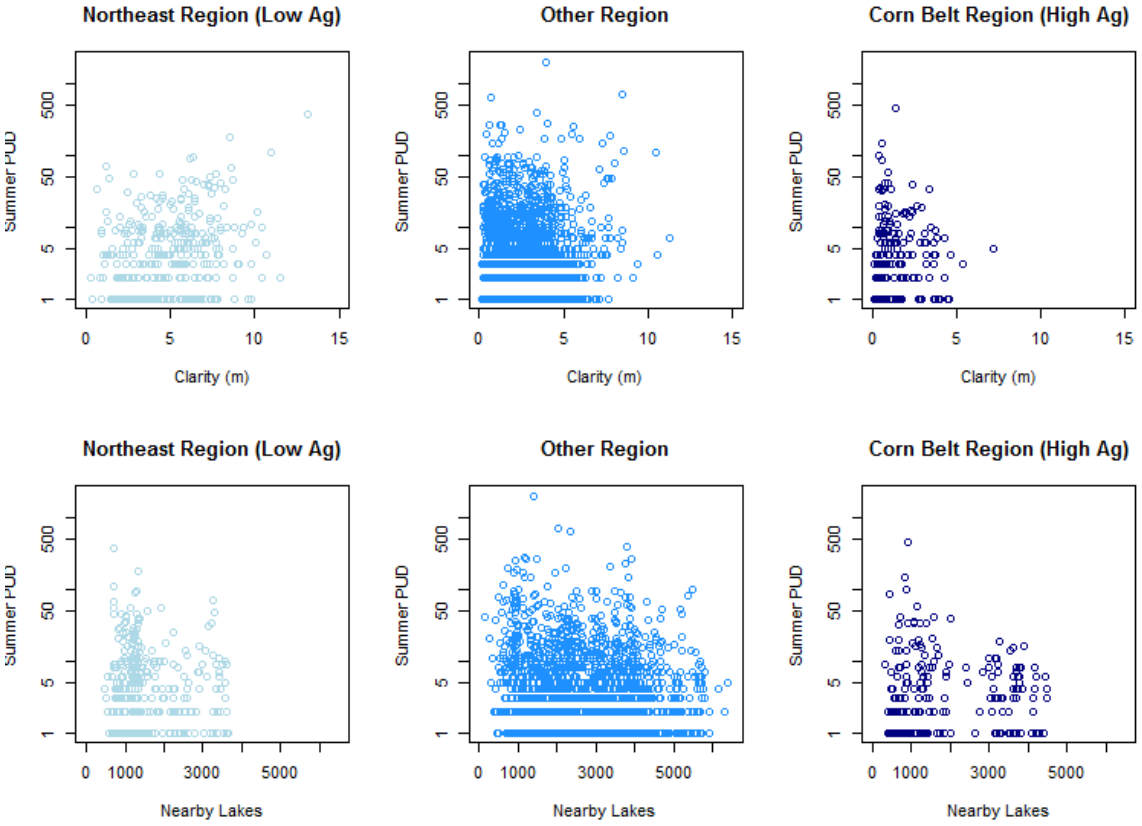
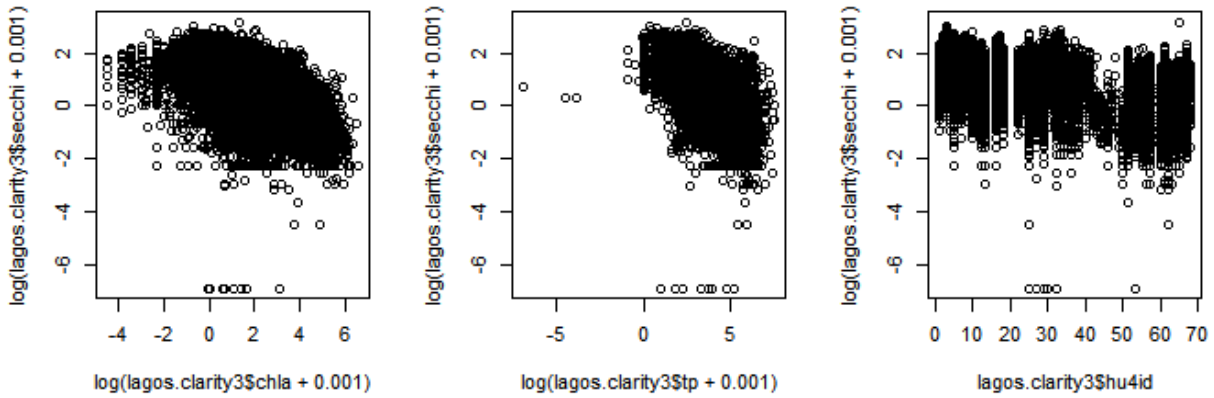


Figure A-4: Scatterplots comparing clarity and nearby lakes with Summer PUD, grouped by agricultural region. Data used for these plots was taken from the truncated dataset (lakes with summer PUD > 0).

Appendix B: Exploratory Data Analysis for the Model of Water Clarity



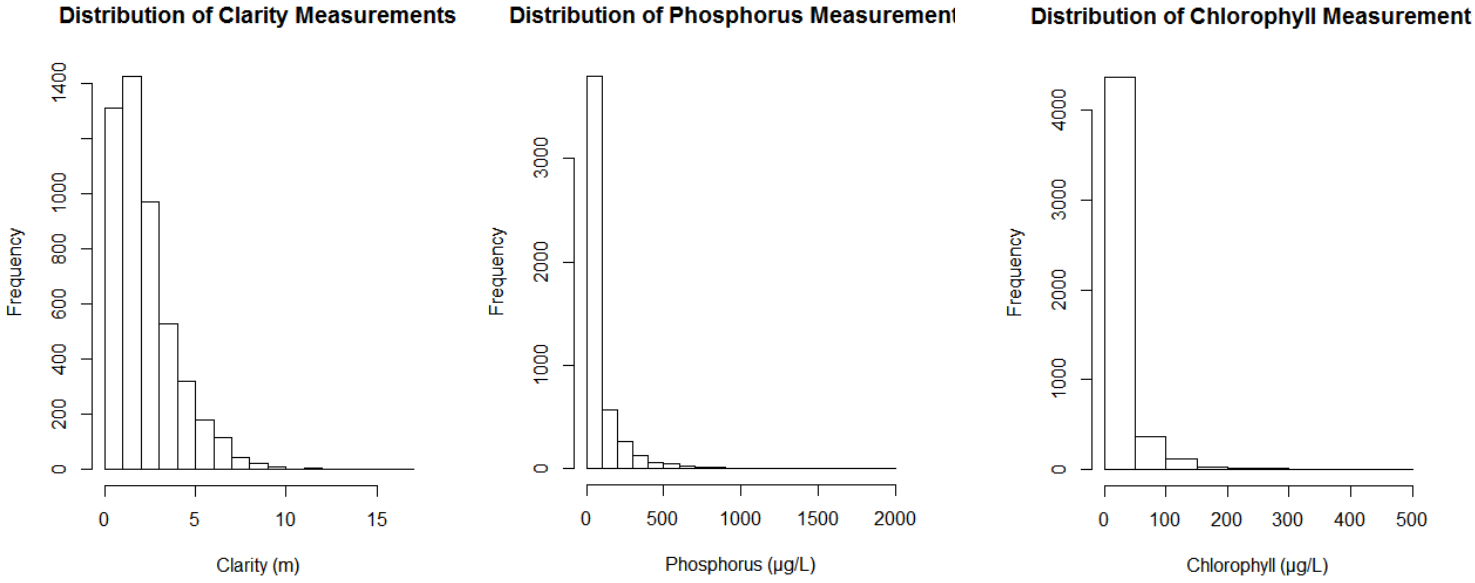


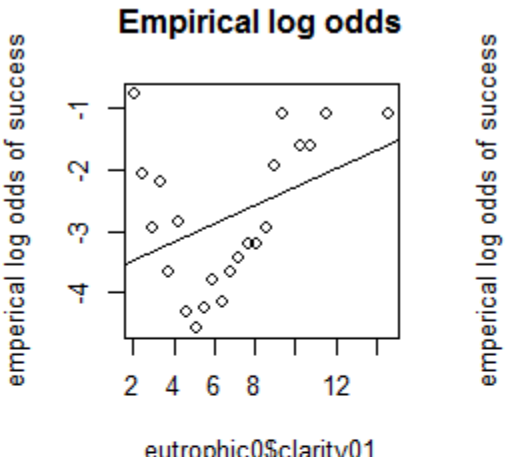
Figure B-1: Scatterplots of clarity with chlorophyll (left), phosphorus (middle), and hu4id indicator (right).

Figure B-2: Histogram distributions of water clarity (left), phosphorus (middle), and chlorophyll (right)

WebLink 1: https://statscon.shinyapps.io/ClarityMap/

This link leads to an interactive map showing all lakes used in model of water clarity. Map features include descriptions of each lake using measured values for all variables in the model, effects plots for both chlorophyll and phosphorus using lake specific values, and predicted clarity levels for all lakes.

Appendix C: Exploratory Data Analysis for the Model of Eutrophic Status



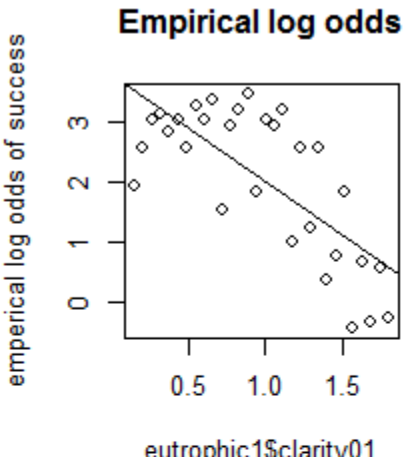
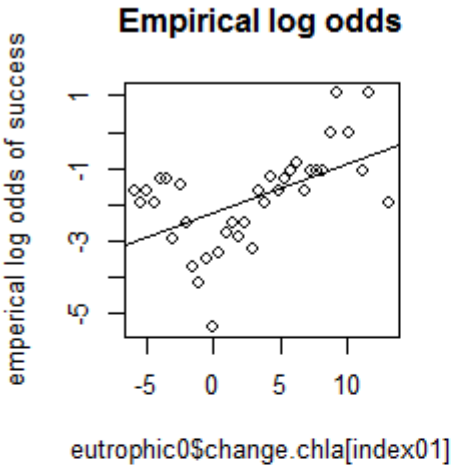


Figure C-1: Empirical log odds plots for clarity, using lakes that were non-eutrophic in 2001 (left) and lakes that were eutrophic in 2001 (right)



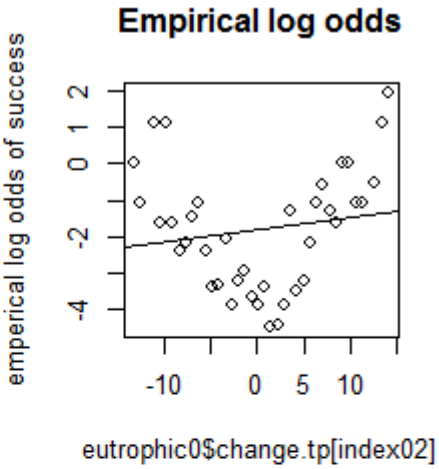
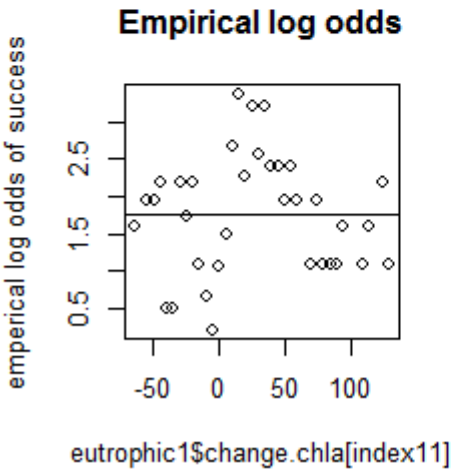
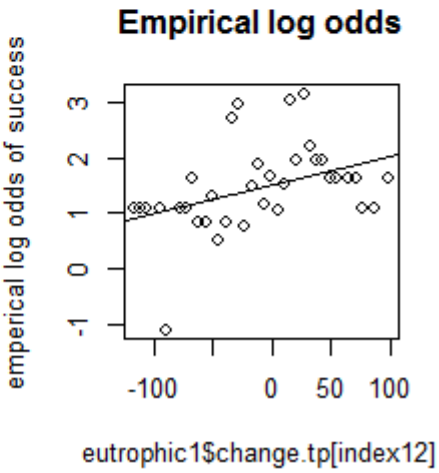


Figure C-2: Empirical log odds plots for both chlorophyll and phosphorus, using the middle 95% of observed values. The top plots are for lakes that were non-eutrophic in 2001, and the bottom plots are for lakes that were eutrophic in 2001.





WebLink 2: https://statscon.shinyapps.io/EutrophicMap/
This link leads to an interactive map modeling eutrophic status of lakes in 2006. Map features include descriptions of all lakes in 2001 and 2006, a map showing lakes that changes eutrophic status between the two years, and the probability that a selected lake was eutrophic in 2006.

Appendix D: Exploratory Data Analysis for the Model of Missing Clarity

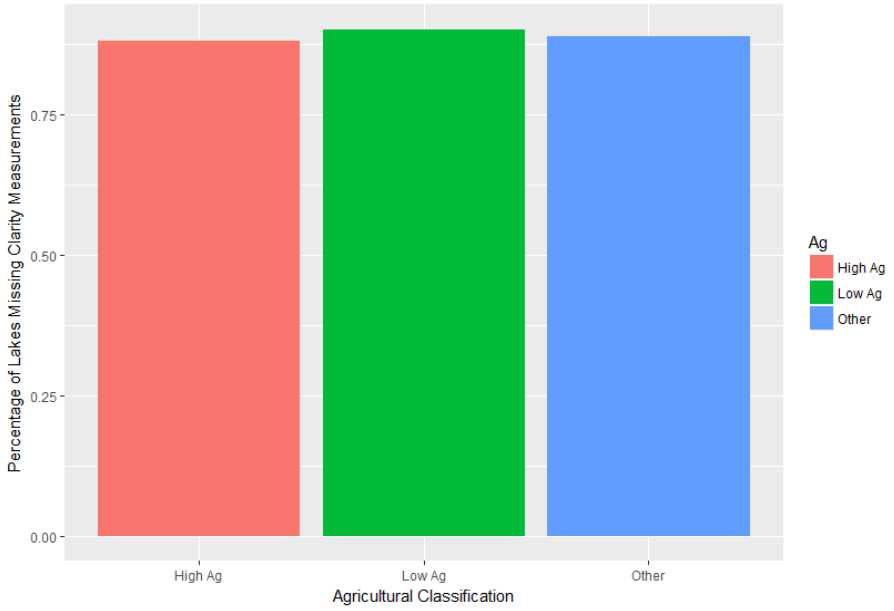


Figure D-1: Barplot of percentage of lakes missing clarity measurements grouped by agricultural region. Data for this plot was taken from the full PUD dataset.

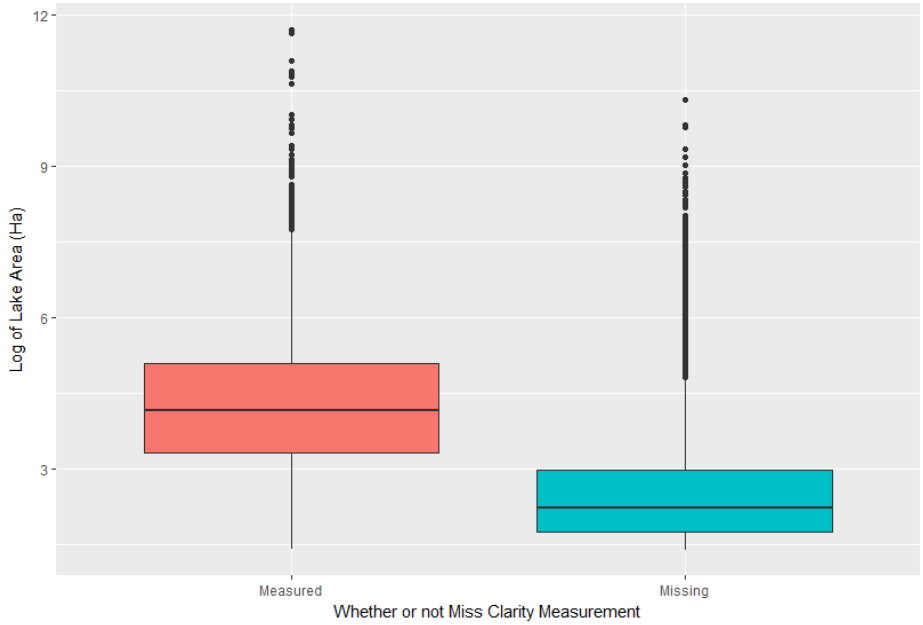


Figure D-2: Boxplot of lake area by whether or not misses clarity measurements. Data for this plot was taken from the full PUD dataset.