Modeling Lake Trophic State: A Data Mining Approach

Jeffrey W. Hollister * [1] W. Bryan Milstead [1] Betty J. Kreakie [1]

[1] *US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, 27 Tarzwell Drive Narragansett, RI, 02882, USA*

* *corresponding author: hollister.jeff@epa.gov*

Abstract

Productivity of lentic ecosystems has been well studied and it is widely accepted that as nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. eutrophic). These broad trophic state classifications are good predictors of ecosystem health and ecosystem services/disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). While the relationship between nutrients and trophic state provides reliable predictions, it requires *in situ* water quality data in order to paramterize the model. This limits the application of these models to lakes with existing and, more importanly, available water quality data. To expand our ability to predict in lakes without water quality data, we take advantage of the availability of a large national lakes water quality database, land use/land cover data, lake morphometry data, other universally available data, and modern data mining approaches to build and assess models of lake tropic state that may be more universally applied. We use random forests and random forest variable selection to identify variables to be used for predicting trophic state and we compare the performance of two models of trophic state (as determined by chlorophyll *a* concentration). The first model estimates trophic state with *in situ* as well as universally available data and the second model uses universally available data only. For each of these models we used three separate trophic state categories, for a total of six models. Overall accuracy for models buils from *in situ* and universal data ranged from 0.669% to 0.867%. For the universal data only models, Overall accuraccy ranged from 0.489% to 0.757%. Lastly, it is believed that the presence and abundance of cyanobacteria is strongly associated with trophic state. To test this we examine the association between estimates of cyanobacteria biovolume and the measured and predicted trophic state and find a positive relationship. Expanding these preliminary results to include cyanobacteria taxa indicates that cyanobacteria are significantly more likely to be found in highly eutrophic lakes. These results suggest that predictive models of lake trophic state may be improved with additional information on the landscape surrounding lakes and that those models provide additional information on the presence of potentially harmful cyanobacteria taxa.

1 Introduction

Productivity in lentic systems is often categorized across a range of tropic states (e.g. the tropic continuum) from early succesional (i.e. oligotrophic)to late successional lakes (i.e. hypereutrophic) (Carlson 1977) and lakes naturally occur across this range. Naturally oligotrophic lakes occur in nutrient poor areas or have a more recent geologic history. These lakes are often found in higher elevations, have clear water, and are often favored for drinking water or direct contact recreation (e.g. swimming).

Lakes with higher productivity (e.g. eutrophic lakes) have greater nutrient loads, tend to be less clear, have greater density of aquatic plants, and often support more diverse and abundant fish communities. Higher primary productivity is not necessarily a predictor of poor ecological condition.

It is natural for lakes to shift from lower to higher trophic states but this is a slow process. Given this fact, monitoring trophic state allows the identification of rapid shifts in trophic state or locating lakes with unusually high productivity (e.g. hypereutrophic). These cases are indicative of lakes under greater anthropogenic nutrient loads, also known as cultural eutrophication, and are more likely to be at risk of fish kills, fouling, and harmful algal blooms(Smith 1998, Smith et al. 1999, 2006). Given the association between trophic state and many ecosystem services and disservices, being able to model trophic state could allow for estimating trophic state in unmonitored lakes and provide a first cut at identifying lakes with the potential for harmful algal blooms and other problems associated with cultural eutrophication.

As trophic state can be defined by a number of *in situ* water quality measurements, most models have used this information as predictors. This leads to accurate models, but also requires data that is sparse and not always availble, limiting the population of lakes that can be modeled. Landscape data is ubiqutious . . .

We have three goals for this preliminary research. First, we build and assess multiple models of lake trophic state using a full suite of data including *in situ* water quality and universally available data (e.g. landscape data). Second, we assess the accuracy of predicted trophic state in lakes with only the universally available data. Lastly, we explore associations between trophic state and cyanobacteria to explore.

2 Methods

2.1 Data and Study Area

We utilize four primary sources of data for this study,the National Lakes Assessment (NLA), the National Lake Cover Dataset (NLCD), modeled lake morphometery, and estimated cyanobacteria biovolumes (Homer et al. 2004, USEPA 2009, Xian et al. 2009, Hollister and Milstead 2010, Hollister et al. 2011, Beaulieu et al. 2013, Hollister 2014). All datasets are national in scale and provide a unique snapshot view of the condition of lakes in the United States'.

The NLA data were collected during the summer of 2007 and the final data were released in 2009. With consistent methods and metrics collected at 1056 locations across the conterminous United States (Figure **??**), the NLA provides a unique opportunity to examine broad scale patterns in lake productivity. The NLA collected data on biophysical measures of lake water quality and habitat. For this analysis we primarily examined the water quality measurements from the NLA (USEPA 2009). Adding to the monitoring data collected via the NLA, we use the 2006 NLCD data to examine the possible landscape-level drivers of trophic status in lakes. The NLCD is a nationally collected land use land cover dataset that also provides estimates of impervious surface. We collected total land use land cover and total percent impervious surface within a 3 kilometer buffer surrounding the lake to examine larger landscape-level effect (Homer et al. 2004, Xian et al. 2009). We also used various measures of lake morphometry (i.e. depth, volume, fetch, etc.) as they are important in understanding lake productivity, yet many of these data are difficult to obtain for large numbers of lakes over broad regions. To add this information we modeled lake morphometry (Hollister and Milstead 2010, **???**, Hollister et al. 2011, Hollister 2014). Lastly, to explore associations between trophic state and cyanobacteria, we used estimates of cyanobacterial biovoulme caluclated by Beaulieu *et al.* (2013). Cyanobacteria biovolumes are a truer measure of cyanobacteria dominance than abundance as there is great variability in the size within and between species. We have consolidated the taxa level estimates from Beaulieu *et al.* (2013) and summed that information on a per-lake basis.

2.2 Predicting Trophic State with Random Forests

Random forest is a machine learning algorithm that aggregates numerous decision trees in order to obtain a consensus prediction of the response categories (Breiman 2001). Bootstrapped sample data is recursively partitioned according to a given random subset of predictor variables and completely grown without pruning. With each new tree, both the sample data and predictor variable subset is randomly selected.

While random forests are able to handle numerous correlated variables without a decrease in prediction accuracy, unusually large numbers of related variables can reduce accuracy and increase the chances of over-fitting the model. This is a problem often faced in gene selection and in that field, a variable selection method based on random forest has been succesfully applied (Díaz-Uriarte and De Andres 2006). We use varselRF in R to initially examine the importance of the water quality and GIS derived variables and select a subset, the reduced model, to then pass to random forest(Diaz-Uriarte 2010).

Using Râ€™s randomForest package, we pass the reduced models selected with varSelRF and calculate confusion matrices, overall accuracy and kappa coeffecient (Liaw and Wiener 2002). From the reduced model random forests we collect a consensus prediction and calculate a confusion matrix and summary stats.

2.3 Model Details

Using a combination of the `varSelRF` and `randomForest` we ran models for six combinations of variables and trophic state classifications. These combinations included different combinations of the Chlorphyll *a* trophic states (Table 1) along with all variables and the GIS only variables (i.e. no *in situ* infromation). The six model combinations were:

1. Chlorophyll *a* trophic state - 4 class = All variables (*in situ* water quality, lake morphometry, and landscape)
2. Chlorophyll *a* trophic state - 3 class = All variables (*in situ* water quality, lake morphometry, and landscape)
3. Chlorophyll *a* trophic state - 2 class = All variables (*in situ* water quality, lake morphometry, and landscape)
4. Chlorophyll *a* trophic state - 4 class = All variables (lake morphometry, and landscape)
5. Chlorophyll *a* trophic state - 3 class = All variables (lake morphometry, and landscape)
6. Chlorophyll *a* trophic state - 2 class = All variables (lake morphometry, and landscape)

3 Results and Discussion

3.1 Model 1: 4 Trophic States ~ All Variables

The selected variables that made up Model 1 were Potassium, Nitrogen:Phosphorus, Total Nitrogen, Total Phosphorus, Total Organic Carbon, Turbidity, Ecoregion, Organic Ions, Dissolved Organic Carbon, and Maximum Lake Depth (Table 2). Total accuracy for Model 1 is 0.669% and the Cohen's Kappa is 0.549 (Table 3).

Lastly, tubidity, total phosphorus, total nitrogen, and total organic carbon were the most important predictors of the 4 classes of trophic state (Figure **??**).

3.2 Model 2: 3 Trophic States ~ All Variables

Total accuracy for Model 2 is 0.796% and the Cohen's Kappa is 0.613.

3.3 Model 3: 2 Trophic States ~ All Variables

Total accuracy for Model 3 is 0.867% and the Cohen's Kappa is 0.734.

3.4 Model 4: 4 Trophic States ~ GIS Only Variables

Total accuracy for Model 4 is 0.489% and the Cohen's Kappa is 0.302.

3.5 Model 5: 3 Trophic States ~ GIS Only Variables

Total accuracy for Model 5 is 0.676% and the Cohen's Kappa is 0.347.

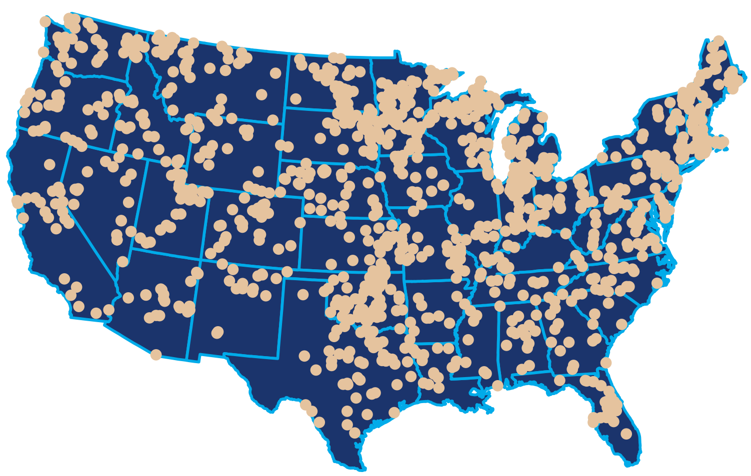
3.6 Model 6: 2 Trophic States ~ GIS Only Variables

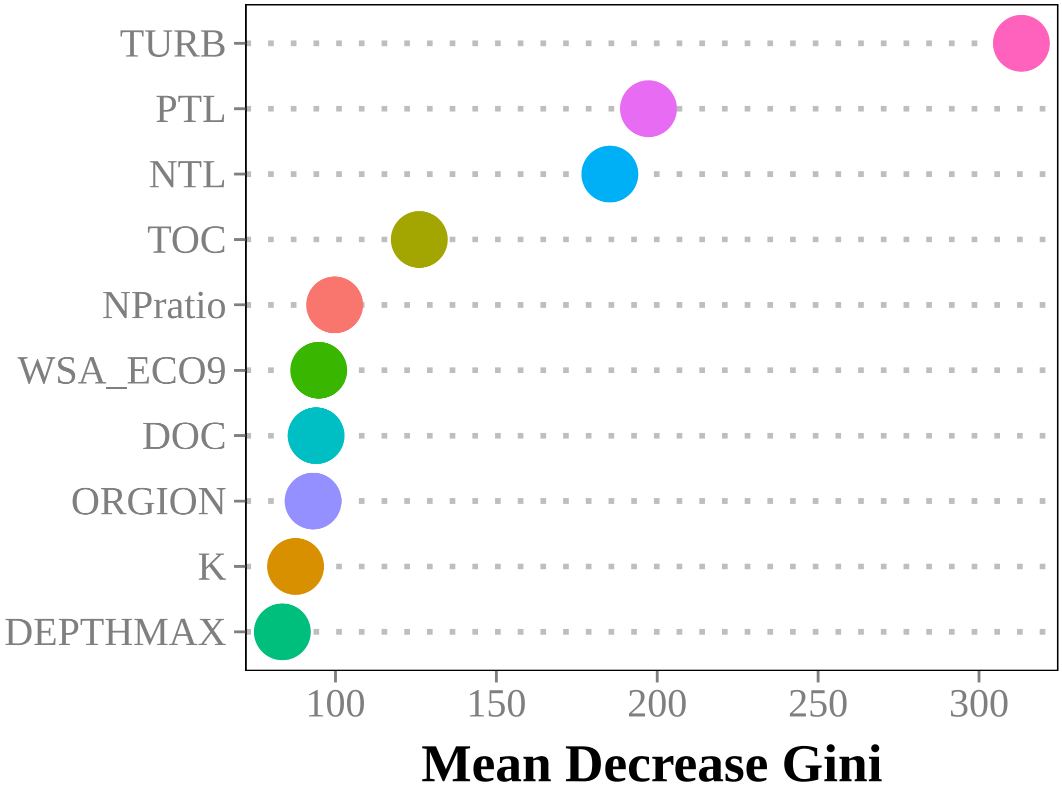
Total accuracy for Model 6 0.757% and the Cohen's Kappa is 0.515.

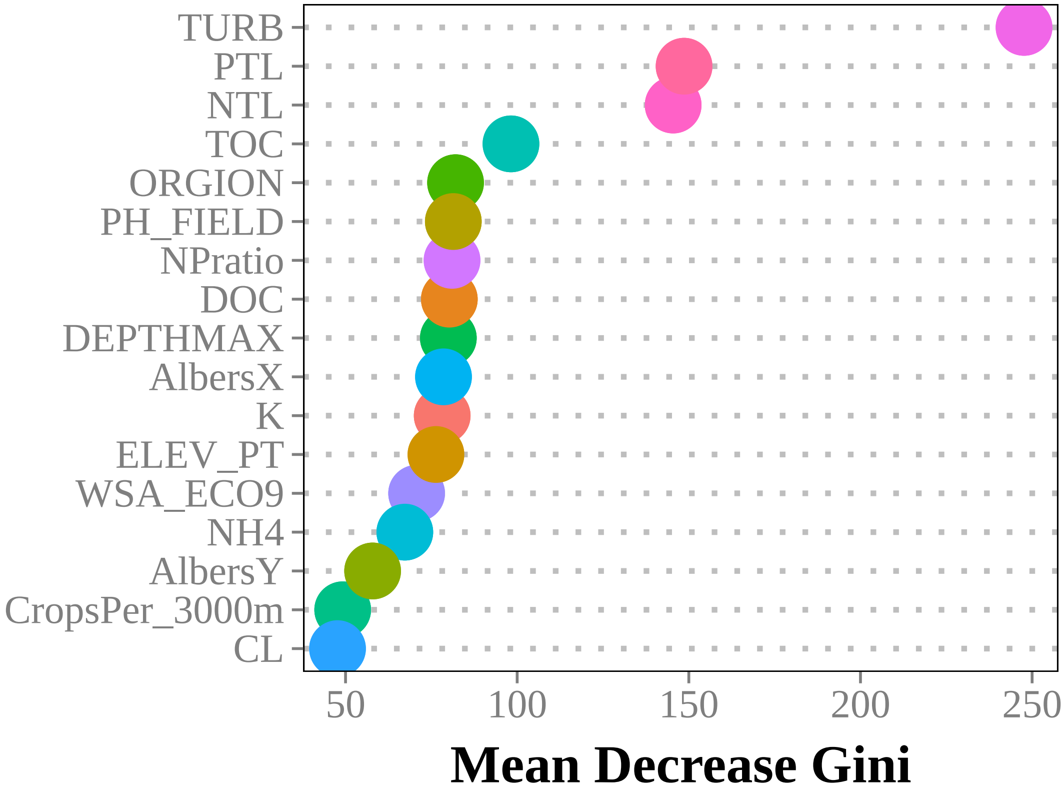
3.7 Associating Trophic State and Cyanobacteria

Arm waving goes here.

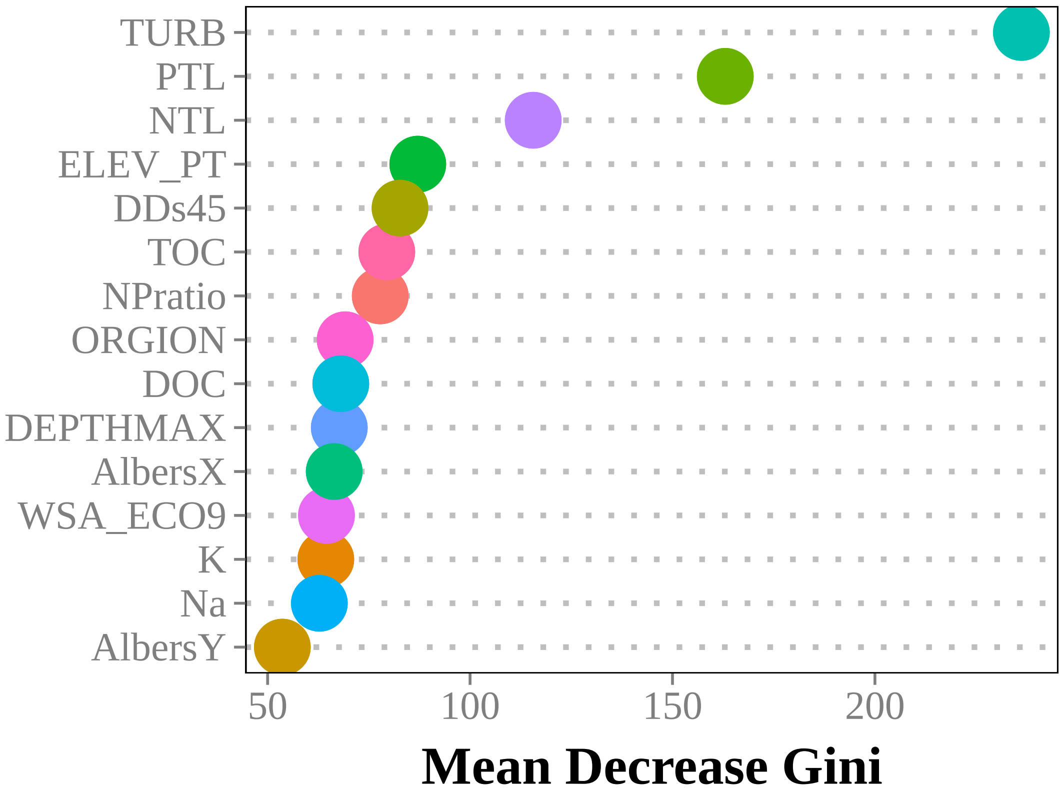
4 Figures

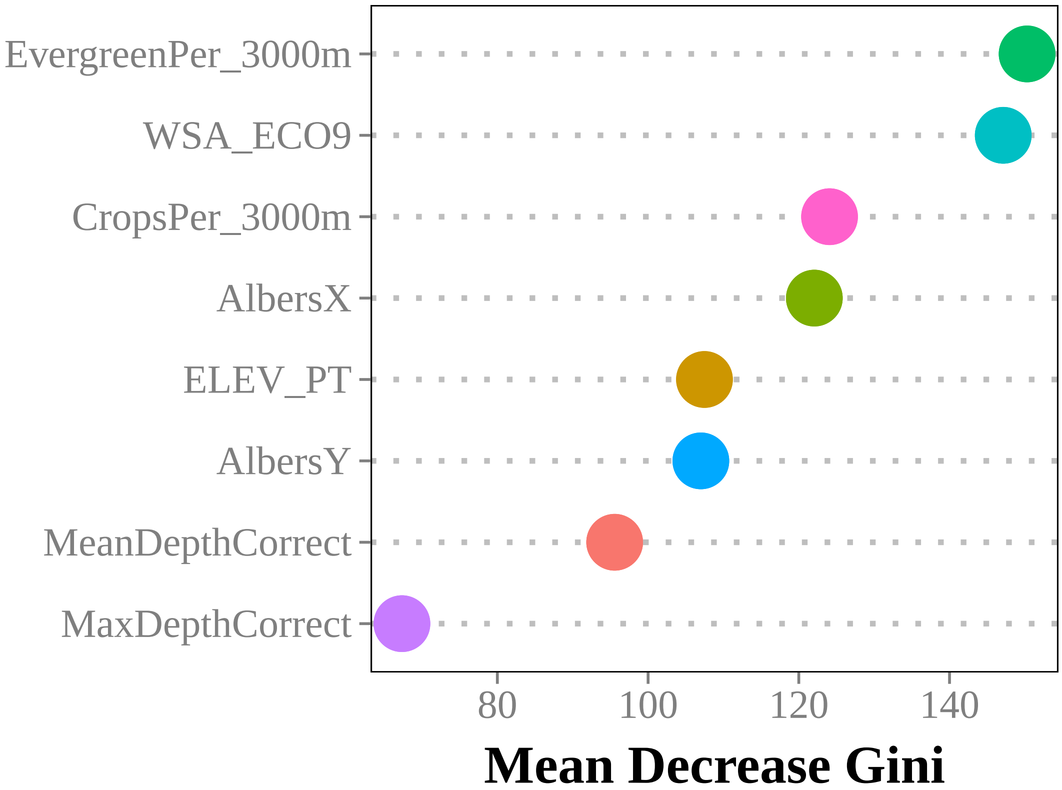



TURB
PTL
NTL
TOC
NPratio
WSA_ECO9
DOC
ORGION
K
DEPTHMAX
100
150
200
250
300
Mean Decrease Gini

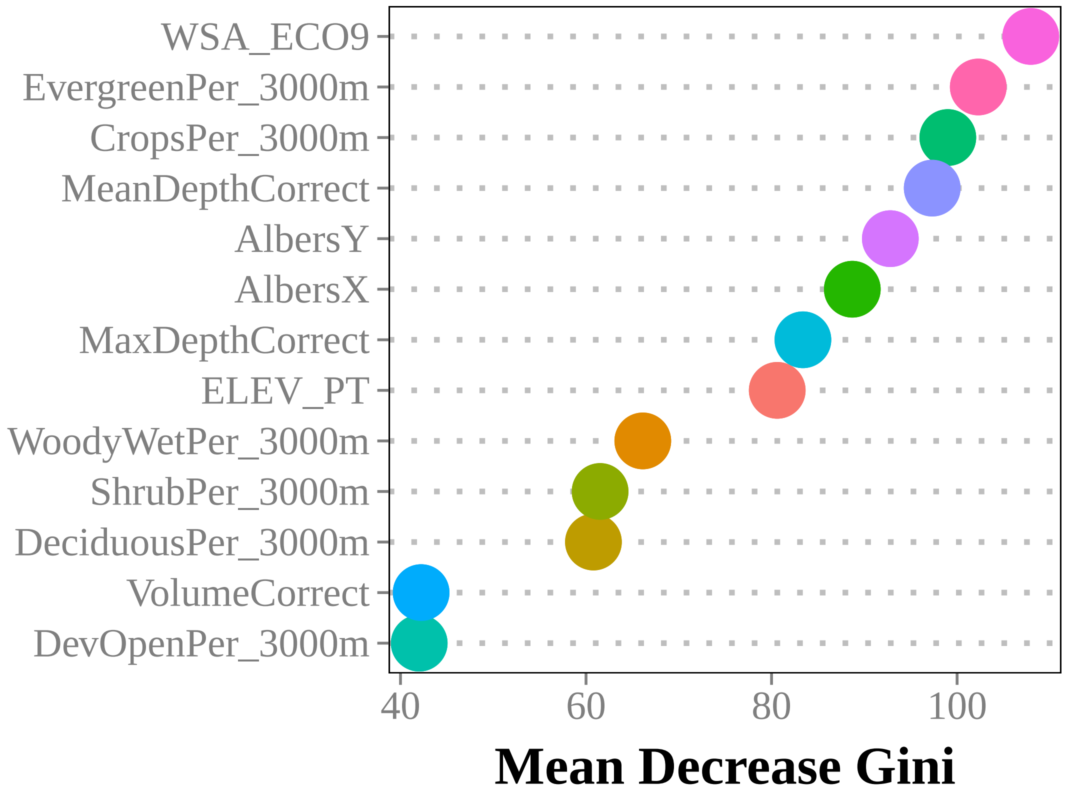

TURB
PTL
NTL
TOC
ORGION
PH_FIELD
NPratio
DOC
DEPTHMAX
AlbersX
K
ELEV_PT
WSA_ECO9
NH4
AlbersY
CropsPer_3000m
CL
50
100
150
200
250
Mean Decrease Gini

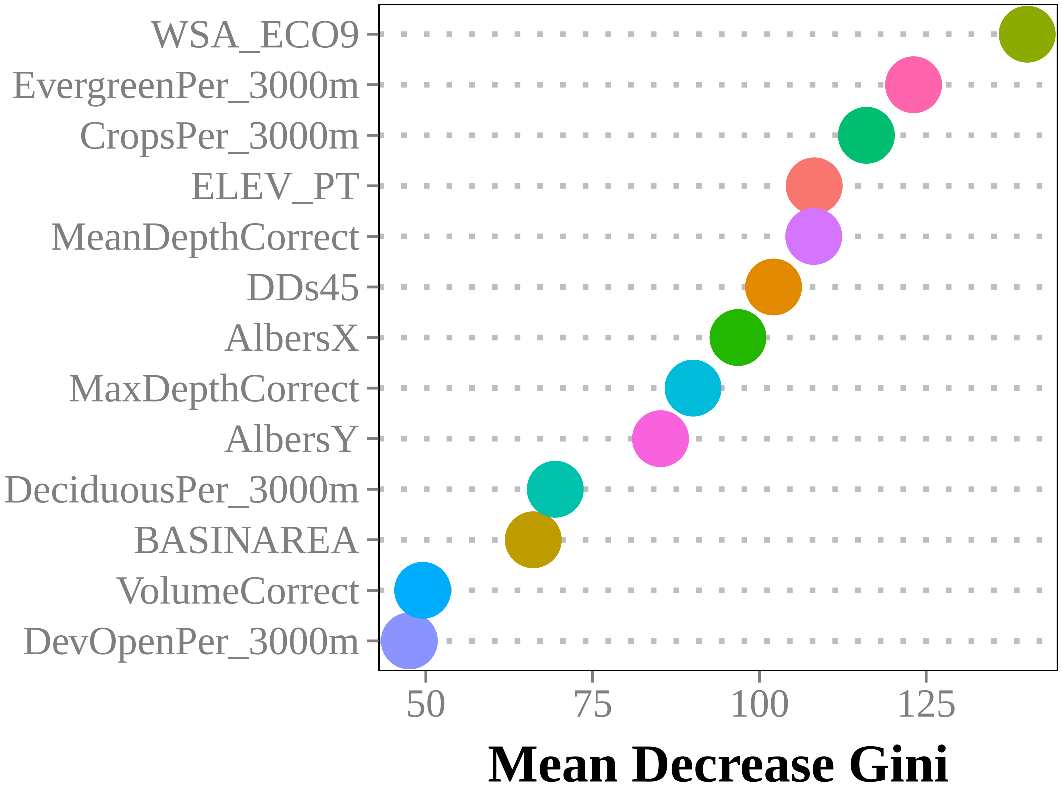
137


TURB
PTL
NTL
ELEV_PT
DDs45
TOC
NPratio
ORGION
DOC
DEPTHMAX
AlbersX
WSA_ECO9
K
Na
AlbersY
50
100
150
200
Mean Decrease Gini

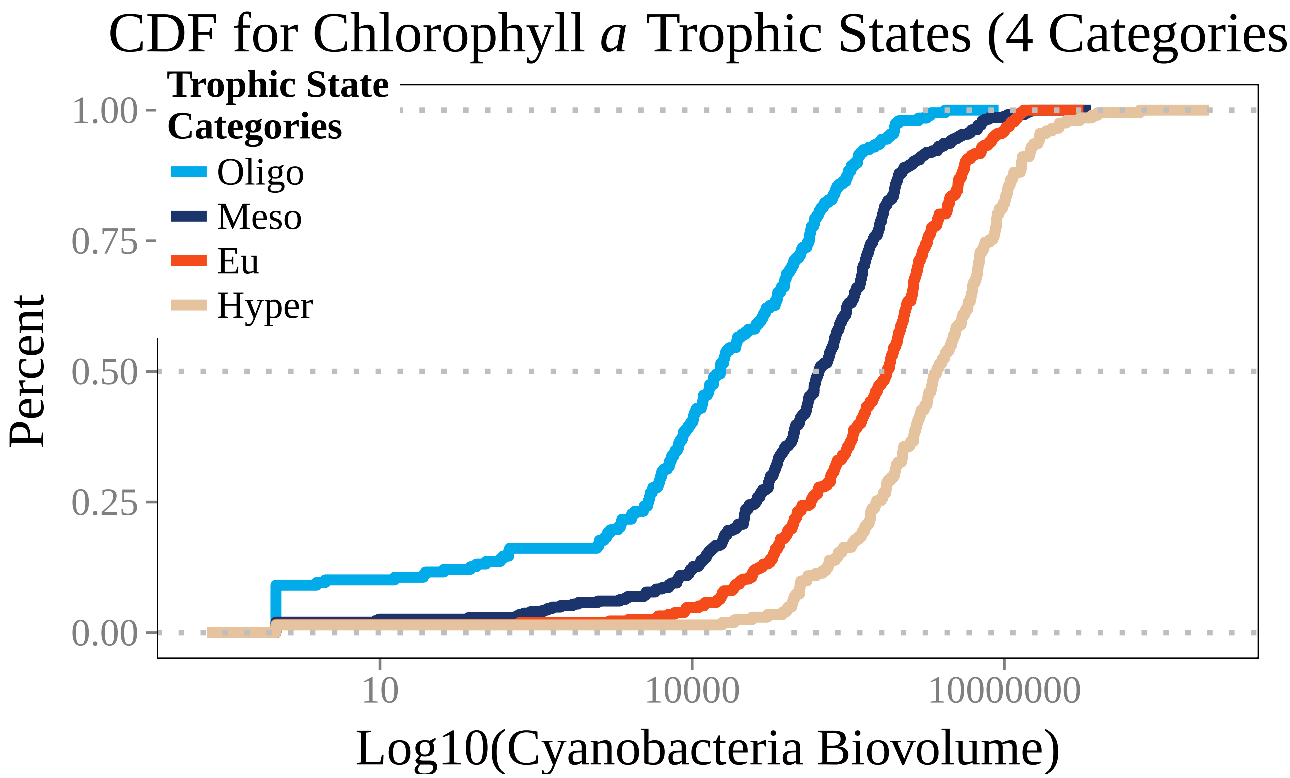


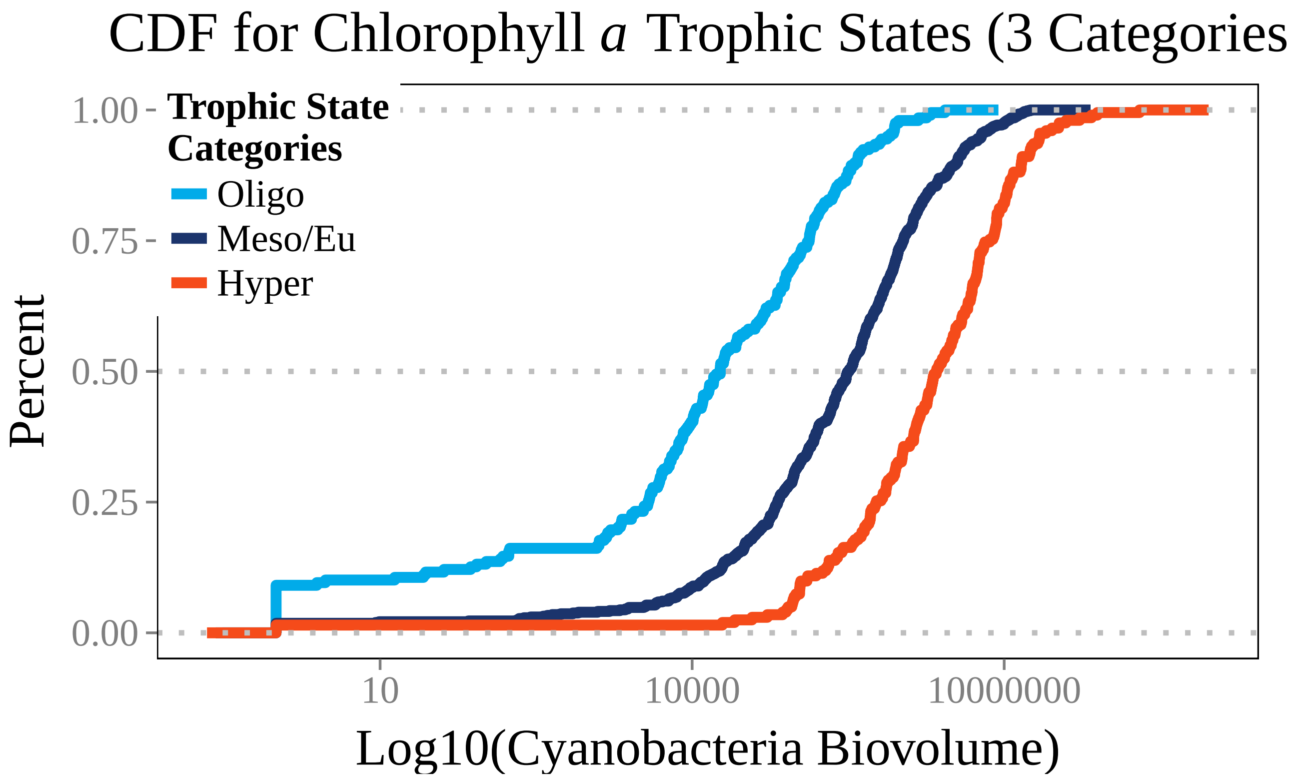
:Importance plot for Model 4


WSA_ECO9
EvergreenPer_3000m
CropsPer_3000m
MeanDepthCorrect
AlbersY
AlbersX
MaxDepthCorrect
ELEV_PT
WoodyWetPer_3000m
ShrubPer_3000m
DeciduousPer_3000m
VolumeCorrect
DevOpenPer_3000m
40
60
80
100
Mean Decrease Gini

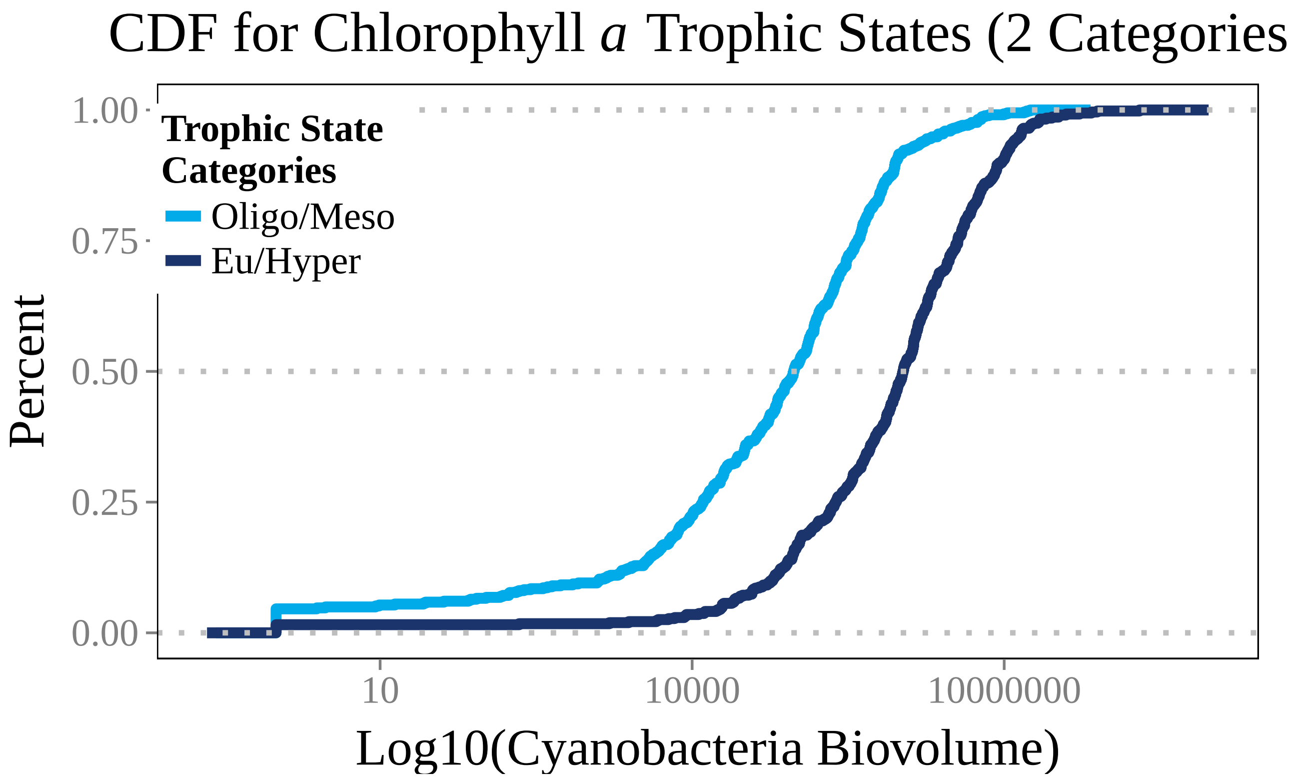

WSA_ECO9
EvergreenPer_3000m
CropsPer_3000m
ELEV_PT
MeanDepthCorrect
DDs45
AlbersX
MaxDepthCorrect
AlbersY
DeciduousPer_3000m
BASINAREA
VolumeCorrect
DevOpenPer_3000m
50
75
100
125
Mean Decrease Gini

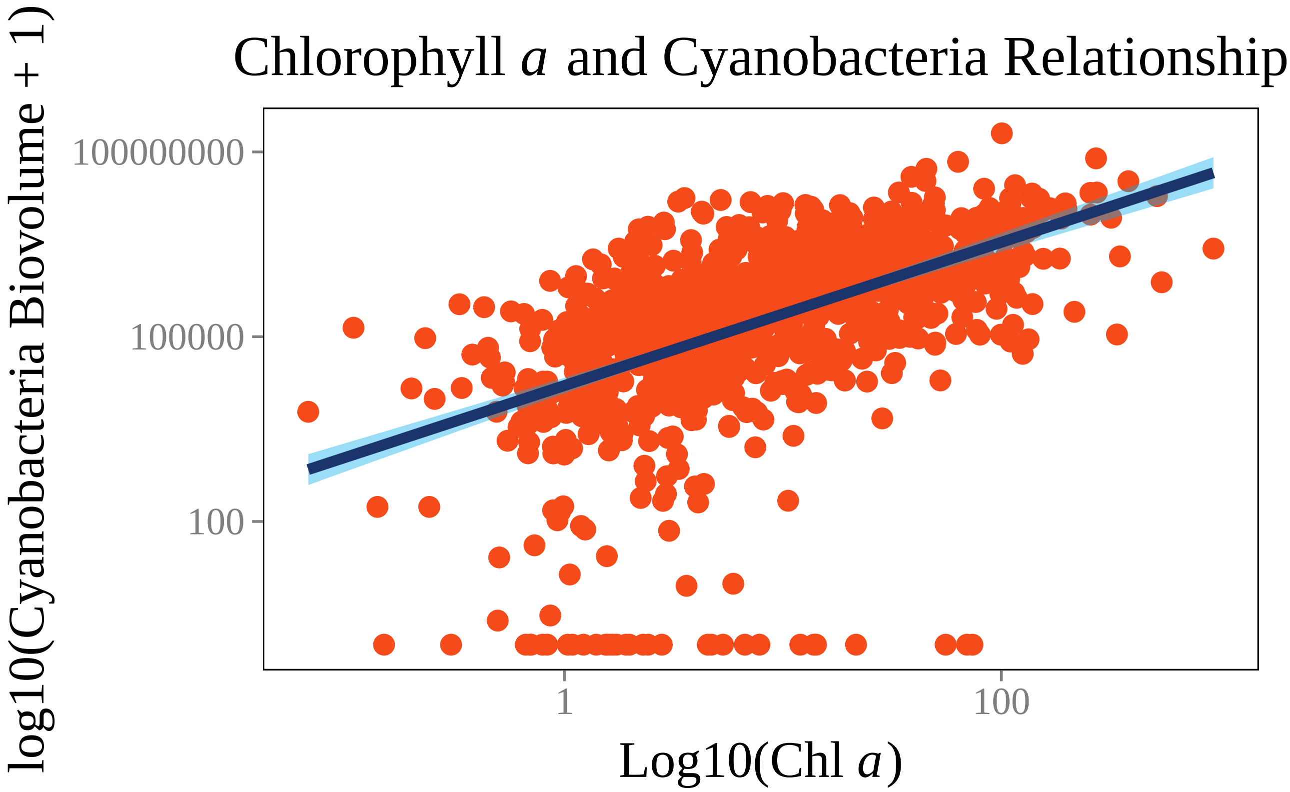
142


CDF for Chlorophyll *a* Trophic States (4 Categories)
Trophic State Categories
Oligo
Meso
Eu
Hyper
Percent
1.00
0.75
0.50
0.25
0.00
10
10000
10000000
Log10(Cyanobacteria Biovolume)


CDF for Chlorophyll *a* Trophic States (3 Categories)
Trophic State Categories
Oligo
Meso/Eu
Hyper
Percent
1.00
0.75
0.50
0.25
0.00
10
10000
10000000
Log10(Cyanobacteria Biovolume)

144


CDF for Chlorophyll a Trophic States (2 Categories)
Trophic State
Categories
Oligo/Meso
Eu/Hyper
Percent
1.00
0.75
0.50
0.25
0.00
10
10000
10000000
Log10(Cyanobacteria Biovolume)


Chlorophyll *a* and Cyanobacteria Relationship
log10(Cyanobacteria Biovolume + 1)
100000000
100000
100
1
100
Log10(Chl *a*)

5 Tables

Trophic State (4)	Trophic State (3)	Trophic State (2)	Cut-off
oligo	oligo	oligo/meso	<= 0.2
meso	meso/eu	oligo/meso	>2-7
eu	meso/eu	eu/hyper	>7-30
hyper	hyper	eu/hyper	>30

Table 1: Chlorophyll a based trophic state cut-offs

Variable	Percent
NPratio	1.00
NTL	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
K	0.99
ORGION	0.33
DOC	0.22
DEPTHMAX	0.11

Table 2: Variable selection results for Model 1

Oligo	Meso	Eu	Hyper	class.error
135	58	4	1	0.32
42	233	77	10	0.36
2	66	222	46	0.34
0	3	69	174	0.29

Table 3: Random Forest confusion matrix for Model 1

Variable	Percent
DEPTHMAX	1.00
DOC	1.00
K	1.00
NTL	1.00
ORGION	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
NPratio	0.88
AlbersX	0.58
CropsPer_3000m	0.36
ELEV_PT	0.23
NH4	0.06
AlbersY	0.04
CL	0.03
PH_FIELD	0.02

Table 4: Variable selection results for Model 2

Oligo	Meso/Eu	Hyper	class.error
122	74	0	0.38
43	604	42	0.12
0	72	173	0.29

Table 5: Random Forest confusion matrix for Model 2

Variable	Percent
K	1.00
NPratio	1.00
NTL	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
ORGION	0.98
DEPTHMAX	0.91
DDs45	0.89
ELEV_PT	0.85
DOC	0.42
AlbersX	0.11
AlbersY	0.03
Na	0.03

Table 6: Variable selection results for Model 3

Oligo/Meso	Eu/Hyper	class.error
485	75	0.13
77	505	0.13

Table 7: Random Forest confusion matrix for Model 3

Variable	Percent
AlbersX	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
AlbersY	0.30
ELEV_PT	0.05
MaxDepthCorrect	0.01

Table 8: Variable selection results for Model 4

Oligo	Meso	Eu	Hyper	class.error
94	72	28	2	0.52
50	201	80	30	0.44
21	110	131	73	0.61
1	34	80	131	0.47

Table 9: Random Forest confusion matrix for Model 4

Variable	Percent
AlbersX	1.00
AlbersY	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MaxDepthCorrect	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
ELEV_PT	0.97
DeciduousPer_3000m	0.94
ShrubPer_3000m	0.32
WoodyWetPer_3000m	0.18
DevOpenPer_3000m	0.13
VolumeCorrect	0.11

Table 10: Variable selection results for Model 5

Oligo	Meso/Eu	Hyper	class.error
80	115	1	0.59
50	585	61	0.16
0	142	104	0.58

Table 11: Random Forest confusion matrix for Model 5

Variable	Percent
AlbersX	1.00
AlbersY	1.00
CropsPer_3000m	1.00
DDs45	1.00
ELEV_PT	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
MaxDepthCorrect	0.98
DeciduousPer_3000m	0.91
DevOpenPer_3000m	0.71
BASINAREA	0.33
VolumeCorrect	0.01

Table 12: Variable selection results for Model 6

Oligo/Meso	Eu/Hyper	class.error
428	129	0.23
147	434	0.25

Table 13: Random forest confusion matrix for Model 6

6 References

Beaulieu, M., F. Pick, and I. Gregory-Eaves. 2013. Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. Limnol. Oceanogr 58:1736–1746.

Breiman, L. 2001. Random forests. Machine learning 45:5–32.

Carlson, R. E. 1977. A trophic state index for lakes. Limnology and oceanography 22:361–369.

Diaz-Uriarte, R. 2010. varSelRF: Variable selection using random forests.

Díaz-Uriarte, R., and S. A. De Andres. 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7:3.

Hollister, J. W. 2014. lakemorpho: Lake morphometry in r.

Hollister, J. W., and W. B. Milstead. In Preparation. National lake morphometry dataset v1.0.

Hollister, J. W., W. B. Milstead, and M. A. Urrutia. 2011. Predicting maximum lake depth from surrounding topography. PLoS ONE 6:e25764.

Hollister, J., and W. B. Milstead. 2010. Using gIS to estimate lake volume from limited data. Lake and Reservoir Management 26:194–199.

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan. 2004. Development of a 2001 national land-cover database for the united states. Photogrammetric Engineering & Remote Sensing 70:829–840.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Smith, V. H. 1998. Cultural eutrophication of inland, estuarine, and coastal waters. Pages 7–49 *in* Successes, limitations, and frontiers in ecosystem science. Springer.

Smith, V. H., S. B. Joye, R. W. Howarth, and others. 2006. Eutrophication of freshwater and marine ecosystems. Limnology and Oceanography 51:351–355.

Smith, V. H., G. D. Tilman, and J. C. Nekola. 1999. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environmental pollution 100:179–196.

USEPA. 2009. National lakes assessment: a collaborative survey of the nation's lakes. ePA 841-r-09-001. Office of Water; Office of Research; Development, US Environmental Protection Agency Washington, DC.

Xian, G., C. Homer, and J. Fry. 2009. Updating the 2001 national land cover database land cover classification to 2006 by using landsat imagery change detection methods. Remote Sensing of Environment 113:1133–1147.