Expanding Models of Lake Trophic State to Predict Cyanobacteria in Lakes: A Data Mining Approach

Jeffrey W. Hollister, W. Bryan Milstead, and Betty J. Kreakie

U.S. Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI 02882

Introduction

Cyanobacteria are an important taxonomic group associated with harmful algal blooms in lakes. Understanding the drivers of cyanobacteria presence has important implications for lake management and for the protection of human and ecosystem health. Chlorophyll *a* concentration, a measure of the biological productivity of a lake, is one such driver that is largely determined by nutrient inputs. As nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. hypereutrophic). These broad trophic state classifications are associated with ecosystem health and ecosystem services and disservices. Thus, models of trophic state might be used to predict things like cyanobacteria. In the preliminary work reported here, we:

1. Build and assess models of lake trophic state
2. Assess ability to predict trophic state in lakes without available *in situ* water quality data
3. Explore association between cyanobacteria and trophic state.

Methods

Data

We utilize four primary sources of data for this study.

1. National Lakes Assessment (NLA) 2007: Using consistent methods and metrics, the NLA collected data, across the conterminous United States, on biophysical measures of lake water quality and habitat (Map 1). For this analysis we primarily examined the water quality measurements from the NLA (USEPA 2009).

2. National Land Cover Dataset (NLCD) 2006: The NLCD is a national land use/land cover dataset. We calculated total land use/land cover and total percent impervious surface within a 3 kilometer buffer of each lake to examine larger landscape-level effects (Homer et al. 2004; Xian, Homer & Fry 2009).

3. Modeled lake morphometry: Various measures of lake morphometry (i.e. depth, volume, fetch, etc.) are important in understanding lake productivity, yet are often difficult to obtain for large numbers of lakes. Modeled estimates solved this problem. (Hollister & Milstead 2010; Hollister, Milstead, & Urrutia 2011; Hollister 2013; Hollister & Milstead *in prep*).

4. Estimated Cyanobacteria Biovolumes: Measuring of cyanobacteria dominance is best done with biovolume as there is great size variability within and amon taxa. Beaulieu et al. (2013) used literature values to estimate biovolumes for the taxa in the NLA. They shared these data with our colleagues and we have summed that information on a per-lake basis.

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.

Random forests are able to handle numerous correlated variables without a decrease in prediction accuracy. But large numbers of related variables can reduce accuracy and lead to over-fitting. This problem, faced in gene selection, has been addressed with a variable selection method based on random forest (Diaz-Uriarte & De Andres 2006). Using 100 iterations of the R package varSelRF, we determine how often our variables are included in a final model (Diaz-Uriarte 2010). The most commonly selected variables (i.e. the reduced model) are used to develop a final random forest model (Liaw & Wiener 2002). From these random forests we collect a consensus prediction and calculate a confusion matrix and summary stats.

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 Chlorophyll *a* trophic states (Table 2) with all variables or with the GIS variables (i.e. no *in situ* information). 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 = GIS variables (lake morphometry, and landscape)
5. Chlorophyll *a* trophic state - 3 class = GIS variables (lake morphometry, and landscape)
6. Chlorophyll *a* trophic state - 2 class = GIS variables (lake morphometry, and landscape)

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* trophic state categories used as dependent variable

Model Results - All Variables

Chlorophyll *a* Trophic States: 4 Classes

	Oligo	Meso	Eu	Hyper	class.error
Oligo	153	58	4	1	0.32
Meso	32	253	76	9	0.35
Eu	2	70	217	47	0.35
Hyper	0	3	68	175	0.29

Total Accuracy: 66.7%
Cohen's Kappa: 0.546

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	0.22
DOC	0.18
DEPTHMAX	0.03

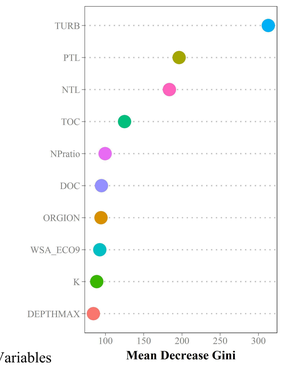

Figure 1. Trophic State (4 Classes) ~ All Variables

Chlorophyll *a* Trophic States: 3 Classes

	Oligo	Meso/Eu	Hyper	class.error
Oligo	121	75	0	0.38
Meso/Eu	10	660	40	0.07
Hyper	0	72	173	0.29

Total Accuracy: 79.9%
Cohen's Kappa: 0.618

VARIABLE	PERCENT
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
DEPTHMAX	0.98
NPratio	0.76
AlbersX	0.48
CropsPer_3000m	0.27
ELEV_PT	0.16
AlbersY	0.05
NH4	0.05
PH_FIELD	0.01
EvergreenPer_3000m	0.01

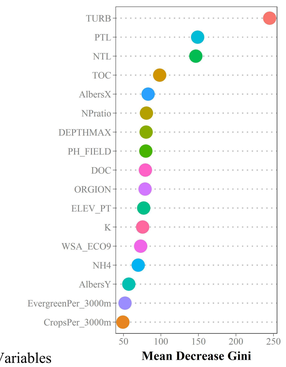

Figure 2. Trophic State (3 Classes) ~ All Variables

Chlorophyll *a* Trophic States: 2 Classes

	Oligo/Meso	Eu/Hyper	class.error
Oligo/Meso	489	71	0.13
Eu/Hyper	77	505	0.13

Total Accuracy: 87.0%
Cohen's Kappa: 0.741

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	0.99
DEPTHMAX	0.96
DDs45	0.90
ELEV_PT	0.85
DOC	0.58
AlbersX	0.06
AlbersY	0.03
Na	0.03

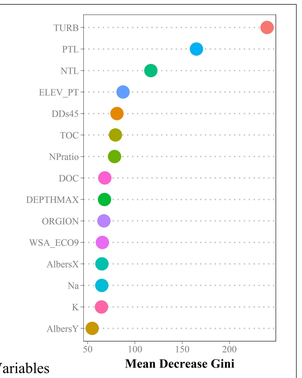

Figure 3. Trophic State (2 Classes) ~ All Variables

Model Results - GIS Only Variables

Chlorophyll *a* Trophic States: 4 Classes

	Oligo	Meso	Eu	Hyper	class.error
Oligo	93	72	27	2	0.52
Meso	48	201	80	32	0.44
Eu	20	114	124	77	0.63
Hyper	2	36	79	129	0.48

Total Accuracy: 48.2%
Cohen's Kappa: 0.292

VARIABLE	PERCENT
AlbersX	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
AlbersY	0.35
ELEV_PT	0.02

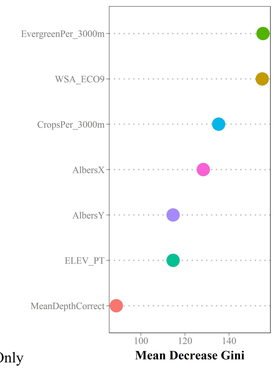

Figure 4. Trophic State (4 Classes) ~ GIS Only

Chlorophyll *a* Trophic States: 3 Classes

	Oligo	Meso/Eu	Hyper	class.error
Oligo	79	116	1	0.6
Meso/Eu	18	552	66	0.16
Hyper	0	141	105	0.57

Total Accuracy: 79.9%
Cohen's Kappa: 0.618

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	0.21
WoodyWetPer_3000m	0.11
DevOpenPer_3000m	0.10
VolumeCorrect	0.04

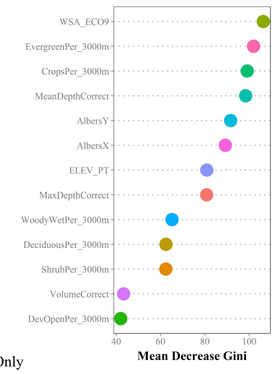

Figure 5. Trophic State (3 Classes) ~ GIS Only

Chlorophyll *a* Trophic States: 2 Classes

	Oligo/Meso	Eu/Hyper	class.error
Oligo/Meso	428	129	0.23
Eu/Hyper	146	435	0.25

Total Accuracy: 79.9%
Cohen's Kappa: 0.618

VARIABLE	PERCENT
AlbersX	1.00
CropsPer_3000m	1.00
DDs45	1.00
ELEV_PT	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
AlbersY	0.98
MaxDepthCorrect	0.98
DeciduousPer_3000m	0.72
DevOpenPer_3000m	0.67
BASINAREA	0.21
PercentImperv_3000m	0.01

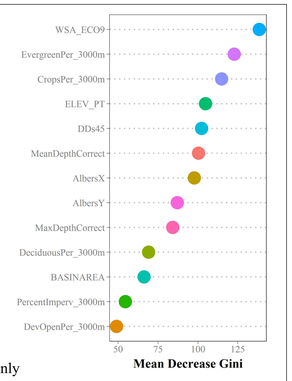

Figure 6. Trophic State (2 Classes) ~ GIS Only

Cumulative Distributions - Chlorophyll *a* and Cyanobacteria Biovolumes

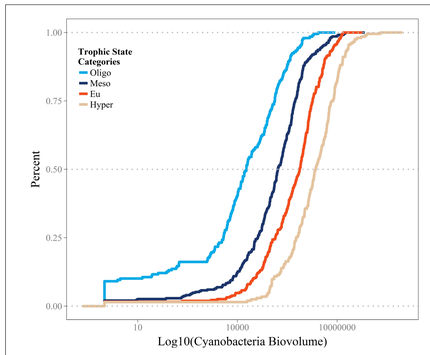

Figure 7. CDF of 4 Class Trophic State and Biovolume

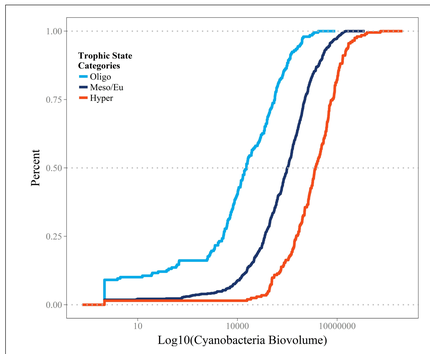

Figure 8. CDF of 3 Class Trophic State and Biovolume

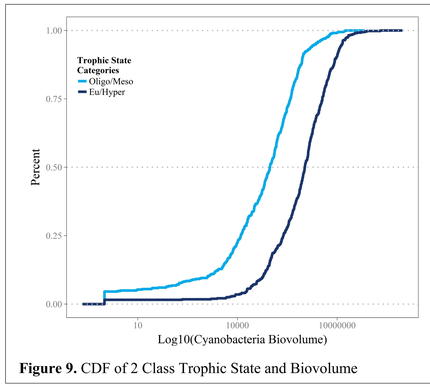

Figure 9. CDF of 2 Class Trophic State and Biovolume

Map 1. Geographic Distribution of Samples in the 2007 National Lakes Assessment.

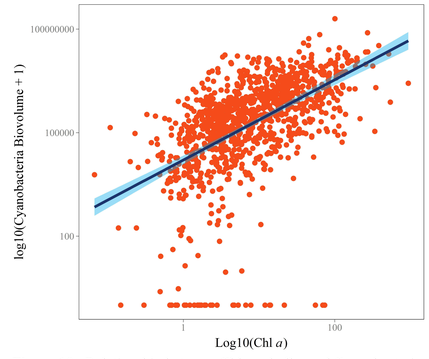

Figure 10. Relationship between Chlorophyll *a* and Cyanobacteria

Poster Source on GitHub

All of the materials that make up this poster are available via GitHub. Included in this repository are an R Markdown document, and R Package with data, and the final poster layout as .svg or .pdf. The repository is available at http://github.com/jhollist/hkm2014ESA, or you may use the QR code below to access this repository.

Acknowledgements

Thanks to John Kiddon, Anne Kuhn-Hines, and Farnaz Nojavan for critical reviews of this work. Thanks to Patricia DeCastro for assistance with poster production. Lastly, we'd be remiss if we left out mention of the fantastic wesanderson R Package developed by Karthik Ram.

References

1.) Beaulieu, Marieke, Frances Pick, and Irene Gregory-Eaves. 2013. "Nutrients and Water Temperature Are Significant Predictors of Cyanobacterial Biomass in a 1147 Lakes Data Set." Limnol. Oceanogr 58 (5): 1736–1746.

2.) Breiman, Leo. 2001. "Random Forests." Machine Learning 45 (1): 5–32.

3.) Diaz-Uriarte, Ramon. 2010. VarSelRF: Variable Selection Using Random Forests. http://CRAN.R-project.org/package=varSelRF.

4.) Diaz-Uriarte, Ramón, and Sara Alvarez De Andres. 2006. "Gene Selection and Classification of Microarray Data Using Random Forest." BMC Bioinformatics 7 (1): 3.

5.) Hollister, Jeffrey. 2013. Lakemorpho: Lake Morphometry in R. http://www.github.com/USEPA/lakemorpho.

6.) Hollister, Jeffrey W, and W Bryan Milstead. *in prep*. "National Lake Morphometry Dataset V1.0."

7.) Hollister, Jeffrey W., W. Bryan Milstead, and M. Andrea Urrutia. 2011. "Predicting Maximum Lake Depth from Surrounding Topography." PLoS ONE 6 (9)

8.) Hollister, Jeffrey, and W Bryan Milstead. 2010. "Using GIS to Estimate Lake Volume from Limited Data." Lake and Reservoir Management 26 (3): 194–199.

9.) Homer, Collin, Chengquan Huang, Limin Yang, Bruce Wylie, and Michael Coan. 2004. "Development of a 2001National Land-cover Database for the 10. United States." Photogrammetric Engineering & Remote Sensing 70 (7):829–840.

10.) Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by RandomForest." R News 2 (3): 18–22.http://CRAN.R-project.org/doc/Rnews/.

11.) 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.

12.) Xian, George, Collin Homer, and Joyce Fry. 2009. "Updating the 2001 National Land Cover Database Land Cover Classification to 2006 by Using Landsat Imagery Change Detection Methods." Remote Sensing of Environment113 (6): 1133–1147.