Environmental Determinants of Lake Trophic Status in the Conterminous United States: A Data Mining Approach

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

Jeffrey W. Hollister (hollister.jeff@epa.gov), US EPA, Office of Research and Development, National Health and Environmental Effects Research Lab, Atlantic Ecology Division, Narragansett, RI, 02882

Betty J. Kreakie (kreakie.betty@epa.gov), US EPA, Office of Research and Development, National Health and Environmental Effects Research Lab, Atlantic Ecology Division, Narragansett, RI, 02882

W. Bryan Milstead (milstead.bryan@epa.gov), US EPA, Office of Research and Development, National Health and Environmental Effects Research Lab, Atlantic Ecology Division, Narragansett, RI, 02882

Abstract

Productivity of lentic ecosystems has been well studied and predicting the algal community response is known to be largely a function of nitrogen and phosphorus. Most existing predictive models take advantage of this well studied relationship to predict chlorphyl a and lake trophic state. While this 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 alternative 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 status and we compare the classification accuracy of a variety of existing and novel models. Models based on nutrients alone predict trophic state with and average of XX% accuracy. Models build with universally available data alone are able to correctluy predict trophic state, on average, xx% of the time. Adding in additional variables to the classic models of Chlorophyl *a* based trophic state improves predictions only by a small percentage. These results suggest that when *in situ* data are availble, additional variables do not appreciable improve predictions of trophic state. Additionally, reliable predictions of trophic state are possible without *in situ* data allowing for a much broader application of trophic state models than has been previously applied.

Keywords: National Lakes Assessment, Cyanobacteria, Chlorophyl a, National Land Cover Dataset, Random Forest, Data Mining

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). Lakes naturally occur across the range of trophic state and higher primary productivity is not necessarily a predictor of poor ecological condition. Lakes that are naturally oligotrophic 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. Lakes will naturally shift 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, Tilman, and Nekola 1999; Smith et al. 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.

Classic models for estimating chlorophyl *a*, and thus trophic state, are linear (or log-linear), and rely solely on nitrogen and phosphorus concentrations. These well established models were initially developed in …

Building on these past efforts, we take advanatage of one the first complete national scale efforts monitoring lakes and widely availble spatial datasets (e.g. land use/land cover, lake morphometry, etc.) to try and discern broad patterns in both in-lake parameters that drive trophic state and landscape level parameters that might also drive trophic state Our primary questions are: XXXXXXX

Methods

Data and Study Area

The two primary sources of data for this study are the National Lakes Assessment (NLA) data and the National Land Cover Dataset (NLCD) (USEPA 2009). Both datasets are national in scale and provide a unique snapshot view of the condition of United States' lakes and the patterns of the lakes surrounding landscape.

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, the NLA provides a unique opportunity to examine continental scale patterns in lake productivity. The NLA collected data on biophysical meausres of lake water quality and habitat. For this analysis we primarily examined the water quality measurements from the NLA [TABLE REF].

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 the surroundin landscape of the lake [TABLE REF]. We defined the surrounding landscape of a lake with three different buffer distances: 300 meters, 1500 meters, and 2500 meters. The various distances were used to tease out differences in local landscape effects versus larger landscape-level effects. Lastly, lake morphometry is often linked to the productivity of a lake (NEED REF). To account for this, we included a recently released dataset on lake morphometry (Hollister PUBLISH THIS YOU MOFO).

Defining Trophic State

The dependent variable for this effort is lake trophic state. Trophic state is usually defined over four levels: oligotrphic, mesotrophic, eutorphic, and hypereutrophic. Commonly, cut-off values for each of these four levels may be specified with nitrogen concnetration, phosphorus concentration, secchi depth, or chlorphyll a concentration (Carlson 1977; USEPA 2009). As this study is based largely from the NLA we use the NLA definition of trophic state based on the chlorophyll a concentrations (Table). Additionally, a common need for managment is to identify lakes that are in the greatest need of managment. Most of the mangement needs (e.g. cyanobacteria, low disolved oxygen, general HABs, etc.) are associated with the most eutrophic conditions. As such, we also build models for two classes, hypereutrophic and non-hyperutrophic.

Trophic State	Hypereutrophic Classes	Cut-off
oligotrophic	Non-Hypereutrophic	<= 0.2
mesotrophic	Non-Hypereutrophic	>2-7
eutrophic	Non-Hypereutrophic	>7-30
hypereutrophic	Hypereutrophic	>30

Predicting Trophic State from Classic Linear Models

Classic Linear Models of Chl *a*: 1. Chl *a* ~ TN 2. Chl *a* ~ TP 3. Chl *a* ~ TN + TP

We use these to predict Chl *a*, then convert observed and predicted Chl *a* to trophic state (Table 1). We calcualte a confusion matrix and summary stats of the matrix for each of the classic models.

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 genetics. In that field, a variable selection method based on random forest has been succesfully applied (NEED REF). We use varselRF in R to initially examine the importance of the variables and select a subset, the reduced model, to then pass to random forest. Details on how we conducted the analyses are availble in the R package associated with the manuscript `hkm` availble via Github (citation for Github Repo)

Using R's randomForest package, we pass the reduced models selected with varSelRF and calculate confusion matrices, overall accuracy and kappa coeffecient (Liaw and Wiener 2002).

Results

Summary Statistics

- Narrative summary.
- Table

Variable Selection

- Which variables were selected to include, and why, in the Random Forest.

- Table.

- Pairs plot of selected variables showing little/weak association between selected variables.

Random Forest

- Summary of Random Forest model (number of Params, total oob, etc.)

Variable Importance

- Narrative description of variables.

- Table of Variables with gini or percent explained.

Predicted Trophic State

- Summary stats of percent of lakes in each class
- Confusion matrix of predicted with actual.

Discussion

- What worked
- What didnt
- What are the determinants and why improtant

- How can this be expanded to other non-monitored lakes?
- What else can Trophic State tell us?
- Cyanobacteria association with?
- CDF Plots

Acknowledgements

References

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Carlson, Robert E. 1977. “A Trophic State Index for Lakes.” *Limnology and Oceanography* 22 (2): 361–369.

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

Smith, Val H. 1998. “Cultural Eutrophication of Inland, Estuarine, and Coastal Waters.” In *Successes, Limitations, and Frontiers in Ecosystem Science*, 7–49. Springer.

Smith, Val H, Samantha B Joye, Robert W Howarth, and others. 2006. “Eutrophication of Freshwater and Marine Ecosystems.” *Limnology and Oceanography* 51 (1): 351–355.

Smith, Val H, G David Tilman, and Jeffery C Nekola. 1999. “Eutrophication: Impacts of Excess Nutrient Inputs on Freshwater, Marine, and Terrestrial Ecosystems.” *Environmental Pollution* 100 (1): 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.