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 importantly, 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. Adding in additional variables to the classic models of Chlorophyl *a* based trophic

state improves predictions only by a small percentage. Models based on universally available data alone are able to predict trophic state

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. Most prior models related to trophic state are either limited in spatial extent, have data from a small number of lakes, model nutrients or chlorophyll a directly, use measures of trophic state not widely used or focus on in-lake information (i.e. nurients) and not on the landscape-level data.

For instance, Imboden and Gächter (1978) built a model of primary prodcuction per unit area with a suite of in lake variables and tested this on only three lakes. Another study by Salas and Martino [-salas1991simplified] focused only on warm water lakes and used a dataset of 27 lakes to build their models. They included loading coefficients of Phosphorus from the surrounding landscapes in addition to lake morphometry nutrient and cholorphyll *a* concentration, and disssolved oxygen. They also focus on the scientific and managerial importance of trophic state. However, the study suffers from data were collected by independent labs with variation in the methods and their focus is on warm-water lakes.

Lastly, … xxx found…

Building on these past efforts, we take advanatage of one the first complete national scale efforts monitoring lakes 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

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 …

- Our primary question is, at the national scale, what are the primary determinants of lake trophic status?
- Can those determinants be used to predict trophic state with an acceptable level of accuracy?

Determinants include, chemical and physical parameters of the lake water column and land use/land cover. Lake trophic status defined by Chl a.

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.

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).

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

Variable Selection

A strength of random forest is its ability to handle numerous correlated variables without a decrease in prediction accuracy. Yet the number of redundant correlated predictor variables in our data requires a cursory reduction through the described variable selection method. To do this we examine the correlation between log transformed chlorophyll a concentration and each of the log transformed variables. The rationale behind this selection method is to discard variables with little to no association with chlorphyll a and thus trophic state. Variables that explained less than 5% of the variance (i.e. a pairwise correlation of less than 0.22) were assumed to not be associated with cholorophyll a concentration and were removed from further consideration. Additionally, variables measuring different attributes of the same distribution (e.g. minimum, maximum or mean temperature) were selected based on the variable with the strongest corelation with chlorophyll a. Lastly, the remaining predictor variables that are highly correlated with one another should not be included in the initial set of variables passed to the random forest, unless sepcified by domain knowledge. As such we examine the pairwise correlations of these remaining variables and make a determination, as determined by knowledge of the system, as to which variables to retain.

Random Forest

As stated above, our goal is to explore relative variable importance in determination of lake trophic status. We selected random forest as our statistical analysis approach, becasuse, among other reasons, random forest provides a robust measure of variable importance. Random forest is a machine learning algorithm that aggregates numerous decision trees in order to obtain a consensus prediction of the response categories. 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.

This randomization provides an intrinsic means to calculate out-of-bag (OOB) error and variables importance.

All random forest analysis was conducted using R's randomForest package; for more details see Breiman (2001).

Variable Importance

- How to use for variable selection
- what we used to identify important variables

Predicted Trophic State

- How random forests makes final predictions,
- what we used to assess accuracy, etc.

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.

Imboden, DM, and R Gächter. 1978. "A Dynamic Lake Model for Trophic State Prediction." *Ecological Modelling* 4 (2): 77–98.

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.