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

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

Introduction

- Trophic State related to stuff we care about
- Largely determined by primary productivity and thus can be estimate with Chl a (among others)
- Most studies of trophic state are limited in spatial extent and don't look for broad scale patterns of variables that drive trophic state
- Most studies of trophic state focus on in-lake variables (i.e. nurients), limited ability to predict over large regions
- 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
- 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) [@usepa2009national]. 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 meausers 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. 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.

Independent Variables

- Chl a Trophic status from NLA.

- What are the cut-offs.

Variable Selection

Prior to running the random forest models we need to reduce the total number of variables from the original *gazillion*. 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

- background on random forest modelling
- why we are using it

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