Application of a continuous lake trophic state index on lakes with limited data

Farnaz Nojavan A.[1], Betty J. Kreakie[2], Jeffrey W. Hollister[2], and Song S. Qian[1]

[1]**Farnaz's new affiliation address**
[2]**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**
[3]**Department of Environmental Sciences, The University of Toledo, Toledo, OH, United States**

Corresponding author:
Farnaz Nojavan A.[1]

Email address: f.nojavan@gmail.com

ABSTRACT

Lake trophic state indices have long been used to provide a measure of the trophic state of lakes. Over time it has been determined that these indices perform better when they utilize mutliple metrics and provide a continuous measurement of trophic state. We utilize such a method for trophic state that is based upon a Proportional Odds Logistic Regression (POLR) model and extend this model with a Bayesian multilevel model that predicts nutrient concentrations from universally available GIS data. This Bayesian multilevel model provides relatively accurate measures of trophic state and has an overall accuraccy of 60%. The approach illustrates a method for estimating a continuous, mutli-metric trophic state index for any lake in the United States. Future improvements to the model will focus on improving overall accuracy and use variables that are more sensitive to change over time.

INTRODUCTION

In this brief research note, we extend a model for estimating a continuous mulit-metric trophic state index described by Nojavan et al. (n.d.). This model uses lake elevation and *in situ* measurements of total nitrogen, total phosphorus, and secchi depth to provide a continuous index of trophic state. The drawback of the developed POLR model is the cost of monitoring multiple predictor variables (e.g., nutrients). This is addressed in the extended application by linking nitrogen and phosphorus to universally available GIS variables. The goal of the extended POLR model is to allow prediction of the trophic state of all lakes (i.e. lakes with limited field data) in the United States.

METHODS

We present the extended application of the developed POLR model using a Bayesian multilevel model. Our modeling work flow is as follows:

1. Develop a random forest model, using R's `randomForest` package, with 5000 trees using only GIS variables to identify the best predictor variables for nitrogen and phosphorus.
2. Develop the extended application model (the Bayesian multilevel model) using R's `rjags` package to run Just Another Gibbs Sampler (JAGS) from inside of R. JAGS is a program for simulation and analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC).
3. Assess the performance of the extended application model using a hold-out validation method (90% training set, 10% evaluation set).

We link nitrogen and phosphorus in the POLR model to a separate nutrient model built from universally available GIS data, thereby, avoiding the need for nitrogen and phosphorus data, costly variables to measure for all lakes. The number of variables for each response variable, nitrogen or phosphorus, was decided using random forest model's variable selection plots(Hollister, Milstead, and Kreakie 2016).

RESULTS AND DISCUSSION

Selected GIS variables for nitrogen and phosphporus were initially screened with variable selection plots (Figures~1 &~2).The figures show model mean squared error as a function of the number of variables. The best representation of nitrogen and phosphorus could be achieved using three variables, adding more than three variables had incremental (< 0.1) impact on root mean square error. The three most important variables were ecoregion, % evergreen forest, and latitude. The random forest models provided estimates of variable importance for nitrogen and phosphorus and the results are reported in figures~3 &4.

Figure~5 represents the regression models. The extended POLR model is grouped into two blocks (gray shaded rectangles). The trophic state classification regression, the POLR model in the lower block, includes nitrogen, phosphorus, secchi disk, and elevation as predictors. The nutrient model, in the upper block, estimates the means of nitrogen and phosphorus based on ecoregion, % evergreen forest, and latitude. The two blocks are connected through the estimated means of nitrogen ($\mu_{Nitrogen}$) and phosphorus ($\mu_{Phosphorus}$) to form the combined model which enables trophic state classification for all lakes without the costly sampling requirement. The relationship between nitrogen, phosphorus, and their predictors was examined using multilevel linear regression models. The standard deviation of the normal distribution, as well as each parameter in the regression model, were then assigned non-informative prior distributions (uniform, or nearly so, to allow the information from the likelihood to be interpreted probabilistically).

The three selected variables, latitude, eco-region, and % evergreen forest, appear to be capturing patterns of total nutrient concentration at three different spatial scales. Figures~6 &~7 depict the partial dependency plot for latitude, the marginal effect of latitude on the predicted outcome of nitrogen or phosphorus in the random forest model. For example for predicted total nitrogen, high concentrations in the northern and southern extremes of the continental US and the lowest predicted concentrations correspond to the mid-latitudes. The ecoregion variable represents an intermediate scale among these three variables and represents the variation between the regions. Finally, the % evergreen variable was summarized within a 3 kilometer buffer around each lake and is presumably summarizing more local land use decisions that are adjacent to lakes.

As mentioned, the extension of the developed POLR model uses eco-region, latitude, and watershed level % evergreen forest as predictors for nitrogen and phosphorus. This contrasts with prior trophic state classification models that are applied to all lakes, regardless of the differences across scale. Lake trophic index, and hence lake trophic classes, should be calculated differently in different eco-regions to accommodate variation in landform and climate characteristics and our proposed model and extension bares this out by identifying and including and eco-regional approach to quantifying trophic state. Furthermore, the developed multilevel model structure can be further expanded to lake-specific trophic state index, upon availability of multiple measurements for each lake.

Mathematically, the models were set up as follows:

$$\mathrm{Nitrogen_{ij}} \sim \mathcal{N}(\mu_{Nitrogen_{ij}}, \sigma^2_{Nitrogen}) \tag{1}$$

where $\mu_{Nitrogen_{ij}} = X_{Nitrogen}B$, $X_{Nitrogen}$ is the matrix of predictors, and B is the vector of coefficients. $Nitrogen_{ij}$ is the $\{i\}$*th nitrogen observation in the* $\{j\}$*th ecoregion.*

$$\mathrm{Phosphorus_{ij}} \sim \mathcal{N}(\mu_{Phosphorus_{ij}}, \sigma^2_{Phosphorus}) \tag{2}$$

where $\mu_{Phosphorus_{ij}} = X_{Phosphorus}\Gamma$, $X_{Phosphorus}$ *is the matrix of predictors, and* Γ *is the vector of coefficients.* $Phosphorus_{ij}$ *is the* $\{i\}$*th phosphorus observation in the* $\{j\}$*th ecoregion.*\

The overall accuracy of the extended POLR model was 0.6 and the balanced accuracies were 0.78, 0.77, 0.69, 0.68 for oligotrophic, mesotrophic, eutrophic, and hypereutrophic classes, respectively (Table~1). Table~2 shows the confusion matrix for the extended POLR model.

The extended POLR model calculates lake trophic index and classes differently for different eco-regions. Please refer to Table~1 for varying coefficients in different eco-regions. For example, eco-regions 3, 6, and 5, corresponding to Northern Plains, Temperate Plains, and Southern Plains, have the highest *positive coefficients for nitrogen. Hence, nitrogen plays a significant role in moving the trophic state index and class toward the eutrophic/hypereutrophic side of the trophic continuum. Further Table˜1 shows the coefficients for latitude and % evergreen. We included these predictors as they were selected as important variables by the random forest model. They may not help predictions dramatically but they do not hurt the results.*

TABLES

Table 1. Coefficients for the extended POLR model.

		Mean	Standard Deviation
Cutoff points/Thresholds	$C_{Oligo\|Meso}$	-156.60	44.04
	$C_{Meso\|Eu}$	-6.18	8.29
	$C_{Eu\|Hyper}$	121.32	35.04
POLR model coefficients	$\alpha_{Elevation}$	-40.20	12.86
	$\alpha_{Nitrogen}$	-44.33	29.29
	$\alpha_{Phosphorus}$	165.90	46.96
	$\alpha_{\text{Secchi Disk Depth}}$	0.18	5.23
Multilevel model coefficients for nitrogen	$\beta_{\%Evergreen}$	0.00	0.01
	$\beta_{Ecoregion_1}$	0.34	0.13
	$\beta_{Ecoregion_2}$	-0.78	0.12
	$\beta_{Ecoregion_3}$	0.96	0.15
	$\beta_{Ecoregion_4}$	-0.37	0.10
	$\beta_{Ecoregion_5}$	0.59	0.10
	$\beta_{Ecoregion_6}$	0.68	0.09
	$\beta_{Ecoregion_7}$	-0.01	0.10
	$\beta_{Ecoregion_8}$	-1.00	0.10
	$\beta_{Ecoregion_9}$	0.11	0.12
	$\beta_{Latitude}$	0.11	0.05
Multilevel model coefficients for phosphorus	$\gamma_{\%Evergreen}$	-0.00	0.01
	$\gamma_{Ecoregion_1}$	0.40	0.09
	$\gamma_{Ecoregion_2}$	-0.90	0.09
	$\gamma_{Ecoregion_3}$	0.73	0.11
	$\gamma_{Ecoregion_4}$	-0.38	0.08
	$\gamma_{Ecoregions_5}$	0.53	0.08
	$\gamma_{Ecoregion_6}$	0.71	0.07
	$\gamma_{Ecoregion_7}$	-0.32	0.08
	$\gamma_{Ecoregion_8}$	-0.69	0.08
	$\gamma_{Ecoregion_9}$	0.07	0.09
	$\gamma_{Latitude}$	-0.03	0.03
Logistic distribution's scale parameter	σ	75.64	21.27

Table 2. Confusion matrix for multilevel POLR model. Each element of the matrix is the number of cases for which the actual state is the row and the predicted state is the column.

	Oligo	Meso	Eu	Hyper
Oligo	5	3	0	0
Meso	3	12	7	1
Eu	0	0	16	10
Hyper	0	1	3	9

FIGURES

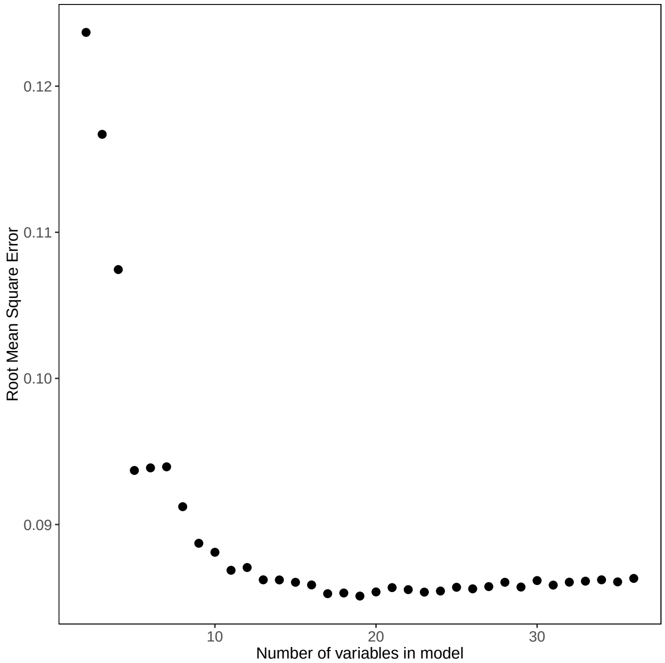

Figure 1. Random Forest model's output for nitrogen with GIS only variables as predictors. Shows model mean squared error as a function of the number of variables.

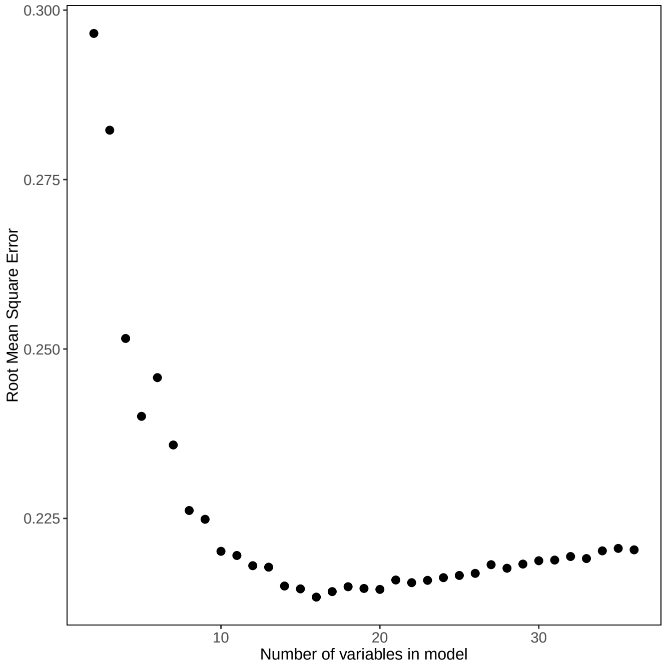

Figure 2. Random Forest model's output for phosphorus with GIS only variables as predictors. Shows model mean squared error as a function of the number of variables.

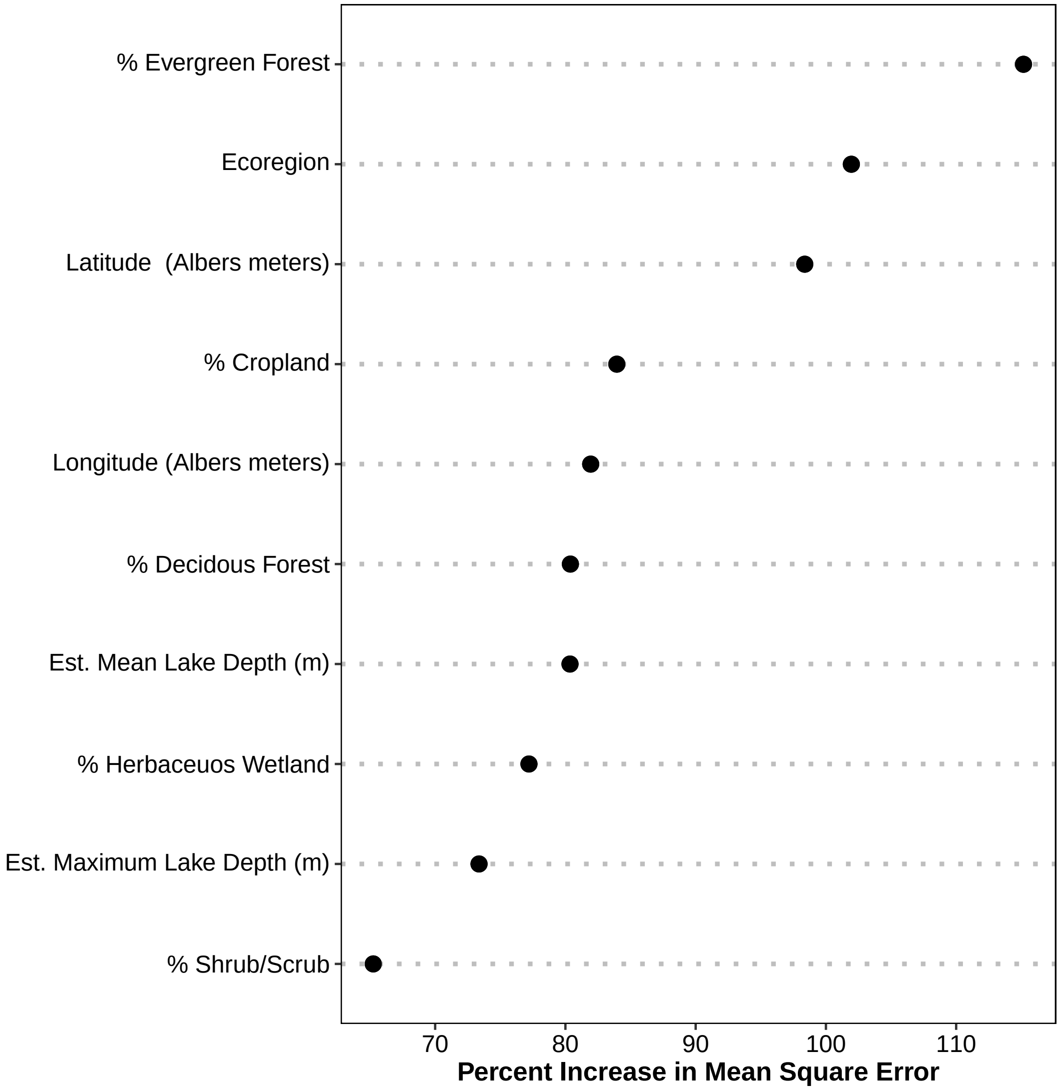

Figure 3. Random Forest model's output for nitrogen predictors. Importance plot for GIS variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

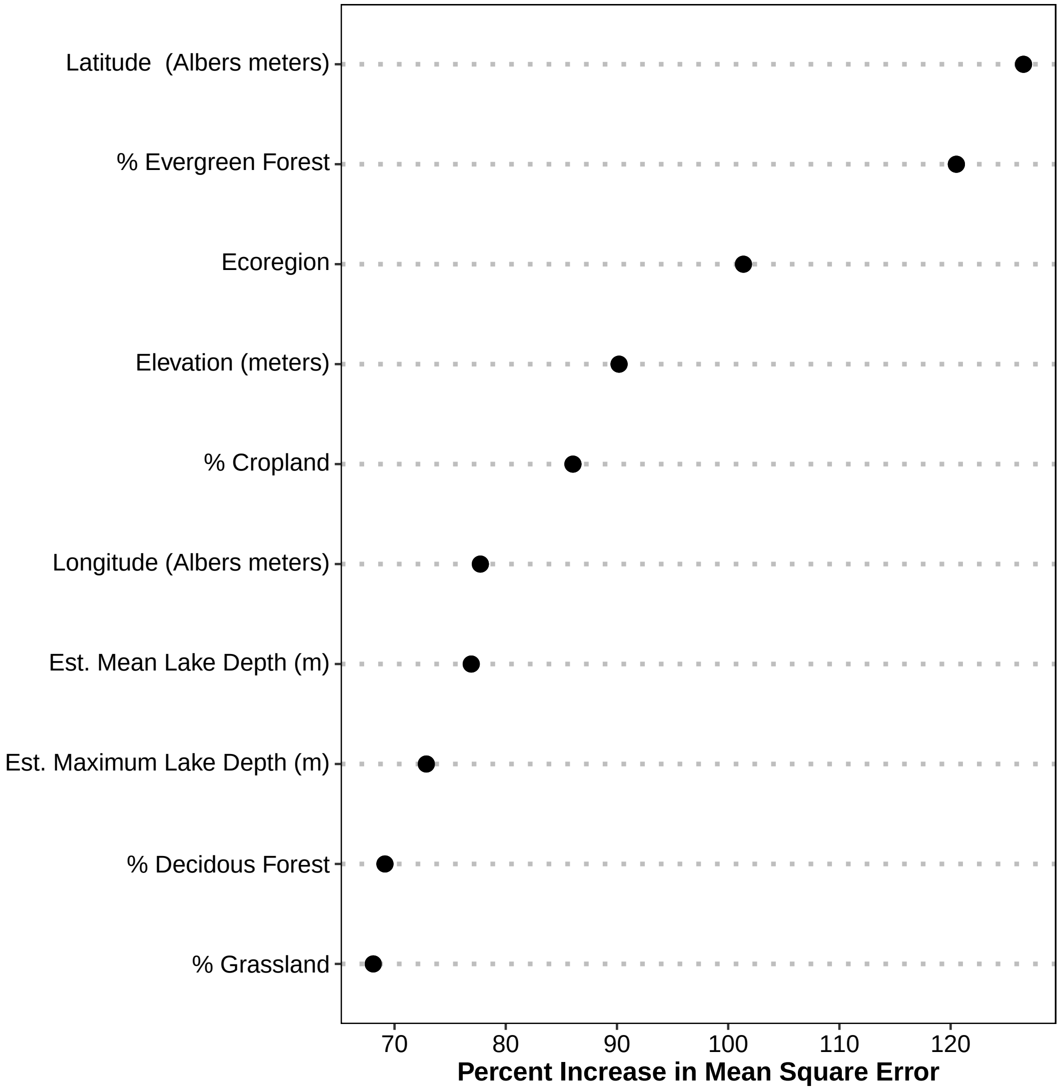

Figure 4. Random Forest model's output for nitrogen predictors. Importance plot for GIS variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

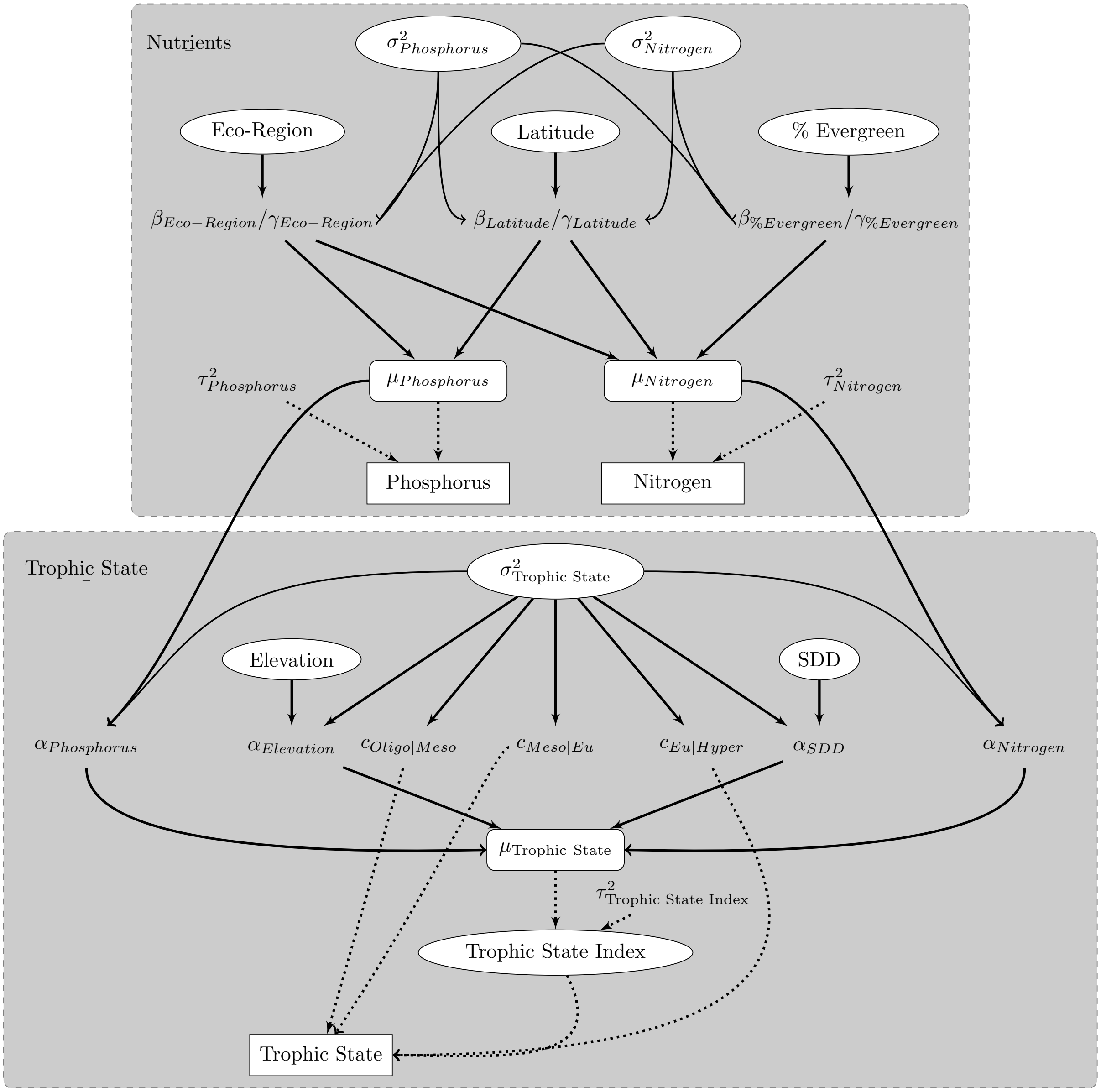

Figure 5. Directed Acyclic Graphical (DAG) model. The lower box depicts the POLR model with its four predictors of secchi disk depth (SDD), elevation, nitrogen, and phosphorus. The upper box is the extension to the POLR model to predict nitrogen and phosphorus using universally available GIS variables.

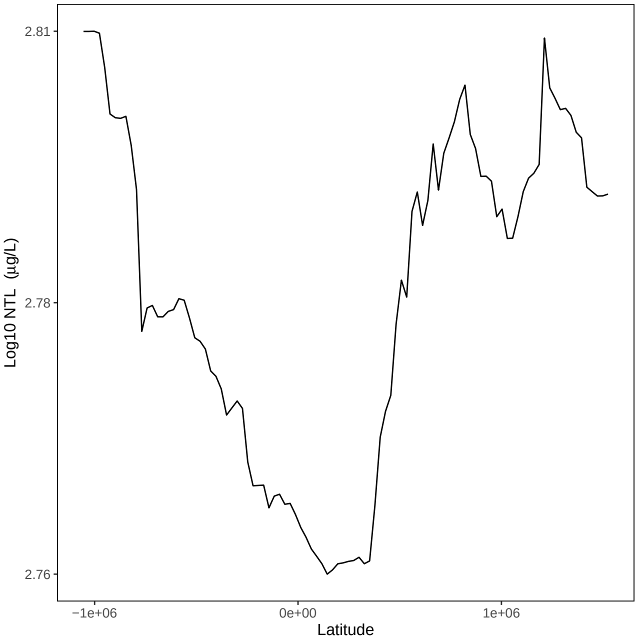

Figure 6. Partial dependency plot for predicted total nitrogen over the range latitude: the effect of latitude on predcited total nitrogen when the rest of the predictors are held constant.

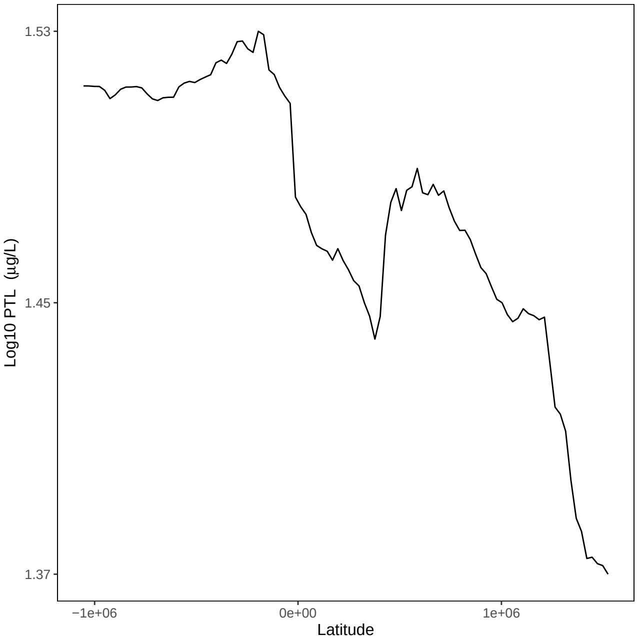

Figure 7. Partial dependency plot for predicted total phosphorus over the range latitude: the effect of latitude on predicted total phosphorus when the rest of the predictors are held constant.

REFERENCES

Hollister, Jeffrey W, W Bryan Milstead, and Betty J Kreakie. 2016. "Modeling Lake Trophic State: A Random Forest Approach." Ecosphere *7 (3).*

Nojavan A., Farnaz, Betty J. Kreakie, Hollister Jeffrey W., and Song Qian. n.d. "Rethinking the Lake Trophic State Index." PeerJ.