Modeling Lake Trophic State: A Data Mining Approach

Jeffrey W. Hollister*[,a], W. Bryan Milstead[a], Betty J. Kreakie[a]

[a] *US Environmental Protection Agency Office of Research and Development National Health and Environmental Effects Research Laboratory Atlantic Ecology Division 27 Tarzwell Drive Narragansett, RI, 02879, USA*

Abstract

Productivity of lentic ecosystems has been well studied and it is widely accepted that as nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. eutrophic). These broad trophic state classifications are good predictors of ecosystem health and ecosystem services/disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). While the relationship between nutrients and trophic state 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 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 state and we compare the performance of two models of trophic state (as determined by chlorophyll a concentration). The first model estimates trophic state with *in situ* as well as universally available data and the second model uses universally available data only. For each of these models we used three separate trophic state categories, for a total of six models. Overall accuracy for the *in situ* and universal data models ranged from xx% to xx% and xx, xx, and xx described the most variation in trophic state. For the universal data only models, Overall accuraccy ranged from xx% to xx% and xx, xx, and xx described the most variation in trophic state. Lastly, it is believed that the presence and abundance of cyanobacteria is strongly associated with trophic state. To test this we examine the association between estimates of cyanobacteria biovolume and the measured and predicted trophic state. Expanding these preliminary results to include cyanobacteria taxa indicates that cyanobacteria are significantly more likely to be found in highly eutrophic lakes. These results suggest that predictive models of lake trophic state may be improved with additional information on the landscape surrounding lakes and that those models provide additional information on the presence of potentially harmful cyanobacteria taxa.

*Corresponding author
Email address: hollister.jeff(at)epa.gov (Jeffrey W. Hollister)

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 and is largely, although not exclusively, 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/disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). Thus, models of trophic state might be used to predict things like cyanobacteria.

We have three goals for this preliminary research:

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 in order to expand models.

Data and Modeling Methods

Data

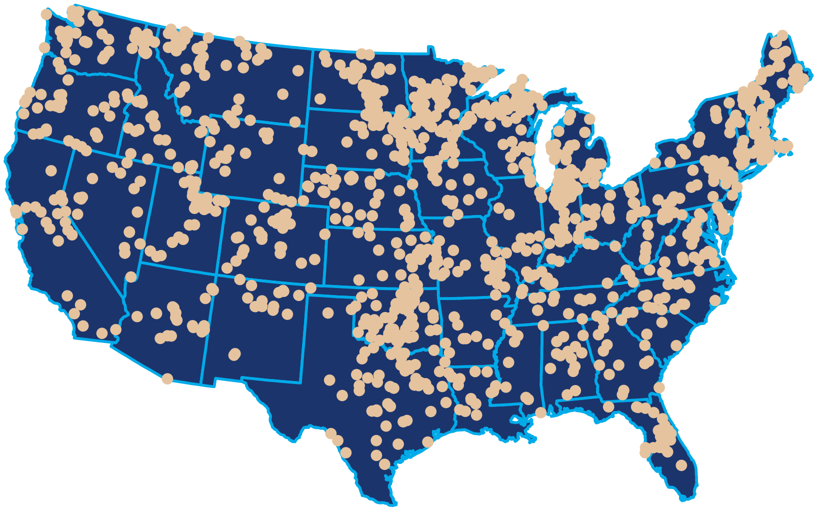
We utilize four primary sources of data for this study. These are outlined below and in Table 1.

1. National Lakes Assessment (NLA) 2007: 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 location s across the conterminous United States (Map 1), the NLA provides a unique opportunity to examine broad scale patterns in lake productivity. The NLA collected data on biophysical measures of lake wat er quality and habitat. For this analysis we primarily examined the water quality measurements from the NLA [1].
2. National Land Cover Dataset (NLCD) 2006: The NLCD is a nationally collected land use land cover dataset. We collected total land use land cover and total percent impervious surface within a 3 kilo meter buffer surrounding the lake to examine larger landscape-level effects [2,3].
3. Modeled lake morphometry: Various measures of lake morphometry (i.e. depth, volume, fetch, etc.) are important in understanding lake productivity, yet many of these data are difficult to obtain for large numbers of lakes over broad regions. To add this information we modeled lake morphometry [4–7].
4. Estimated Cyanobacteria Biovolumes: Cyanobacteria biovolumes are a truer measure of Cyanobacteria dominance than abundance as there is great variability in the size within and between species. To a ccount for this, Beaulieu *et al.* [8] used literature values to estimate biovolumes for the taxa in the NLA. They shared this data and we have summed that information on a per- lake basis.

Variables	Description	Type
PercentImperv_3000m	Percent Impervious	GIS
WaterPer_3000m	Percent Water	GIS
IceSnowPer_3000m	Percent Ice/Snow	GIS
DevOpenPer_3000m	Percent Developed Open Space	GIS
DevLowPer_3000m	Percent Low Intensity Development	GIS
DevMedPer_3000m	Percent Medium Intensity Development	GIS
DevHighPer_3000m	Percent High Intensity Development	GIS
BarrenPer_3000m	Percent Barren	GIS
DeciduousPer_3000m	Percent Decidous Forest	GIS
EvergreenPer_3000m	Percent Evergreen Forest	GIS
MixedForPer_3000m	Percent Mixed Forest	GIS
ShrubPer_3000m	Precent Shrub/Scrub	GIS
GrassPer_3000m	Percent Grassland	GIS
PasturePer_3000m	Percent Pasture	GIS
CropsPer_3000m	Percent Cropland	GIS
WoodyWetPer_3000m	Percent Woody Wetland	GIS
HerbWetPer_3000m	Percent Herbaceuos Wetland	GIS
AlbersX	Longitude	GIS
AlbersY	Latitude	GIS
LakeArea	Lake Surface Area	GIS
LakePerim	Lake Perimeter	GIS
ShoreDevel	Shoreline Development Index	GIS
DATE_COL	Date Samples Collected	Water Quality
WSA_ECO9	Ecoregion	GIS
BASINAREA	Watershed Area	GIS
DEPTHMAX	Maximum Depth	Water Quality
ELEV_PT	Elevation	GIS

Variables	Description	Type
DO2_2M	Dissolved Oxygen	Water Quality
PH_FIELD	pH	Water Quality
COND	Conductivity	Water Quality
ANC	Acid Neutralizing Capacity	Water Quality
TURB	Turbidity	Water Quality
TOC	Total Organic Carbon	Water Quality
DOC	Dissolved Organic Carbon	Water Quality
NH4	Ammonium	Water Quality
NO3_NO2	Nitrate/Nitrite	Water Quality
NTL	Total Nitrogen	Water Quality
PTL	Total Phosphorus	Water Quality
CL	Chloride	Water Quality
NO3	Nitrate	Water Quality
SO4	Sulfate	Water Quality
CA	Calcium	Water Quality
MG	Magnesium	Water Quality
Na	Sodium	Water Quality
K	Potassium	Water Quality
COLOR	Color	Water Quality
SIO2	Silica	Water Quality
H	Hydrogen Ions	Water Quality
OH	Hydroxide	Water Quality
NH4ION	Calculate Ammonium	Water Quality
CATSUM	Cation Sum	Water Quality
ANSUM2	Anion Sum	Water Quality
ANDEF2	Anion Deficit	Water Quality
SOBC	Base Cation Sum	Water Quality

Variables	Description	Type
BALANCE2	Ion Balance	Water Quality
ORGION	Estimate Organic Anions	Water Quality
CONCAL2	Calculated Conductivity	Water Quality
CONDHO2	D-H-O Calculated Conductivity	Water Quality
TmeanW	Mean Profile Water Temperature	Water Quality
DDs45	Growing Degree Days	GIS
MaxLength	Maximum Lake Length	GIS
MaxWidth	Maximum Lake Width	GIS
MeanWidth	Mean Lake Width	GIS
FetchN	Fetch from North	GIS
FetchNE	Fetch form Northeast	GIS
FetchE	Fetch from East	GIS
FetchSE	Fetch from Southeast	GIS
MaxDepthCorrect	Estimated Maximum Lake Depth	GIS
VolumeCorrect	Estimated Lake Volume	GIS
MeanDepthCorrect	Estimated Mean Lake Depth	GIS
NPratio	Nitrogen:Phophorus Ratio	Water Quality



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 [9]. 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 gene selection and in that field, a variable selection method based on random forest has been succesfully applied [10]. We use varselRF in R to initially examine the importance of the water quality and GIS derived variables and select a subset, the reduced model, to then pass to random forest[11].

Using R's randomForest package, we pass the reduced models selected with varSelRF and calculate confusion matrices, overall accuracy and kappa coeffecient [12]. From the reduced model 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 Chlorphyll *a* trophic states (Table 2) along with all variables and the GIS only variables (i.e. no *in situ* infromation). 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 = All variables (lake morphometry, and landscape)
5. Chlorophyll *a* trophic state - 3 class = All variables (lake morphometry, and landscape)

6. Chlorophyll *a* trophic state - 2 class = All variables (lake morphometry, and landscape)

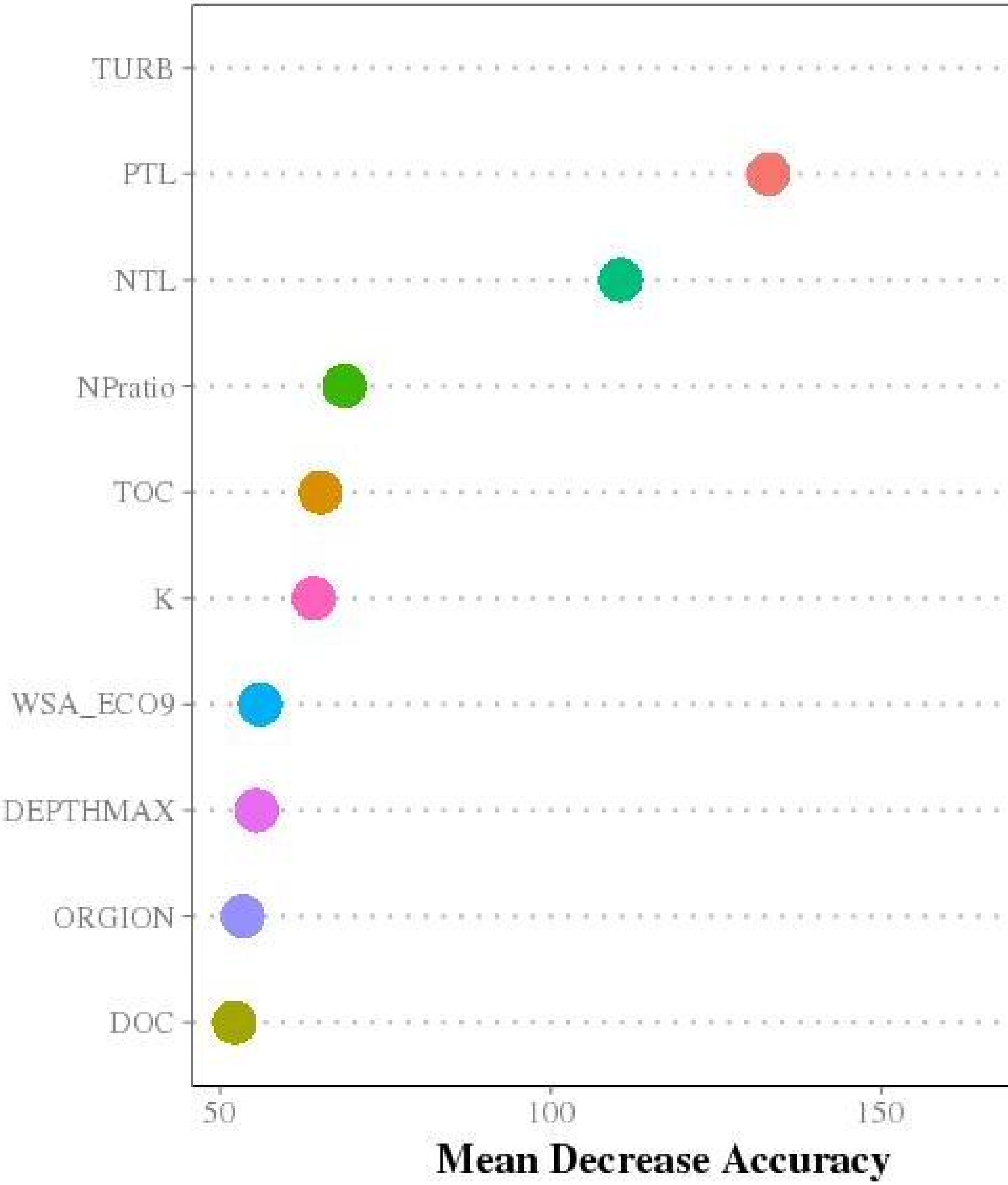
Trophic State (4)	Trophic State (3)	Trophic State (2)	Cut-off
oligo	oligo	oligo/meso	<= 0.2
meso	meso/eu	oligo/meso	>2-7
eu	meso/eu	eu/hyper	>7-30
hyper	hyper	eu/hyper	>30

Results

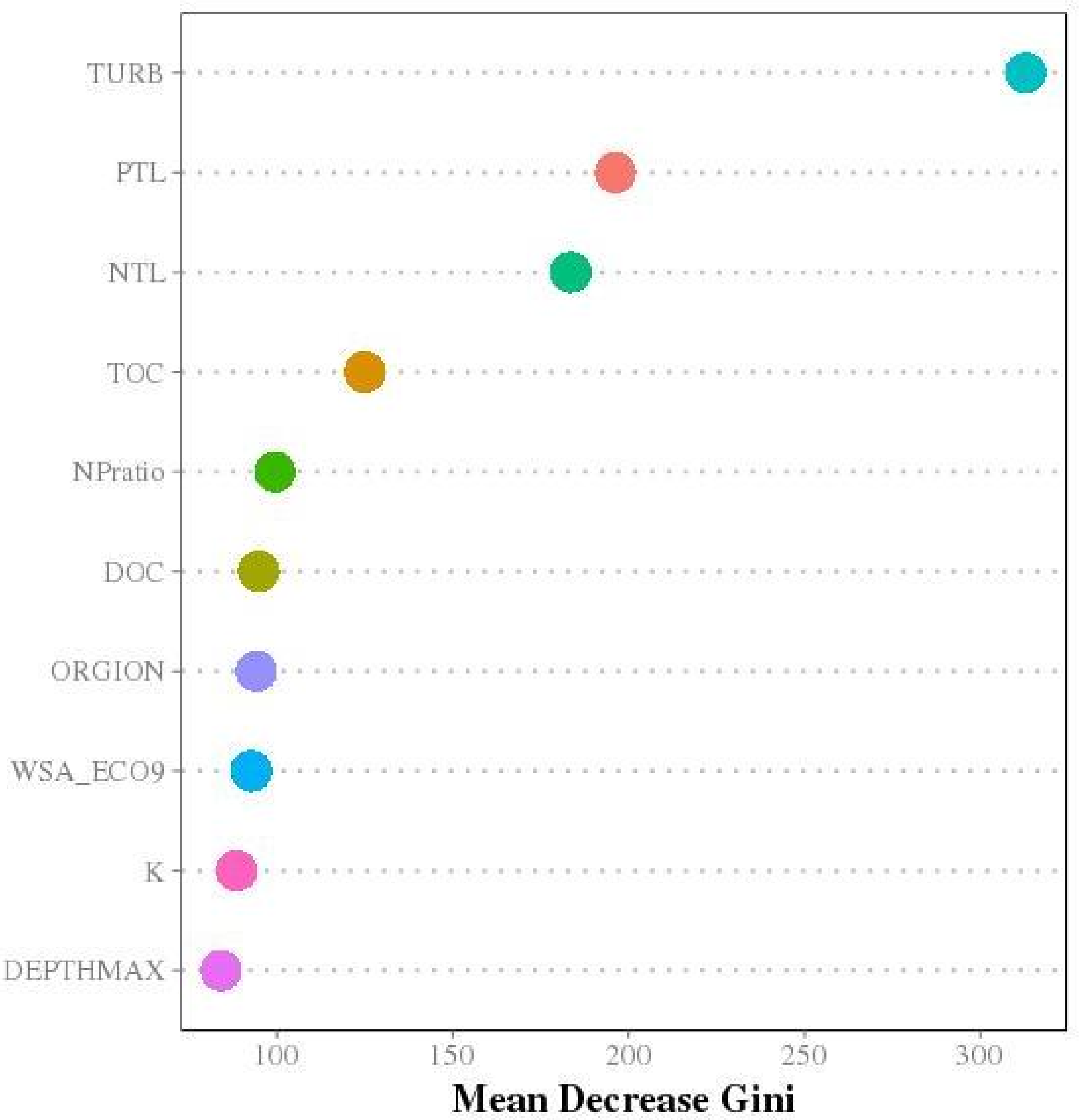
Model 1: 4 Trophic States ~ All Variables

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	0.29
DOC	0.18
DEPTHMAX	0.03

|Oligo |Meso |Eu |Hyper |class.error | |:—–|:—-|:—|:—–|:———–| |135 |58 |4 |1 |0.32 | |42 |235 |76 |9 |0.35 | |2 |70 |217



|47 |0.35 | |0 |3 |68 |175 |0.29 |



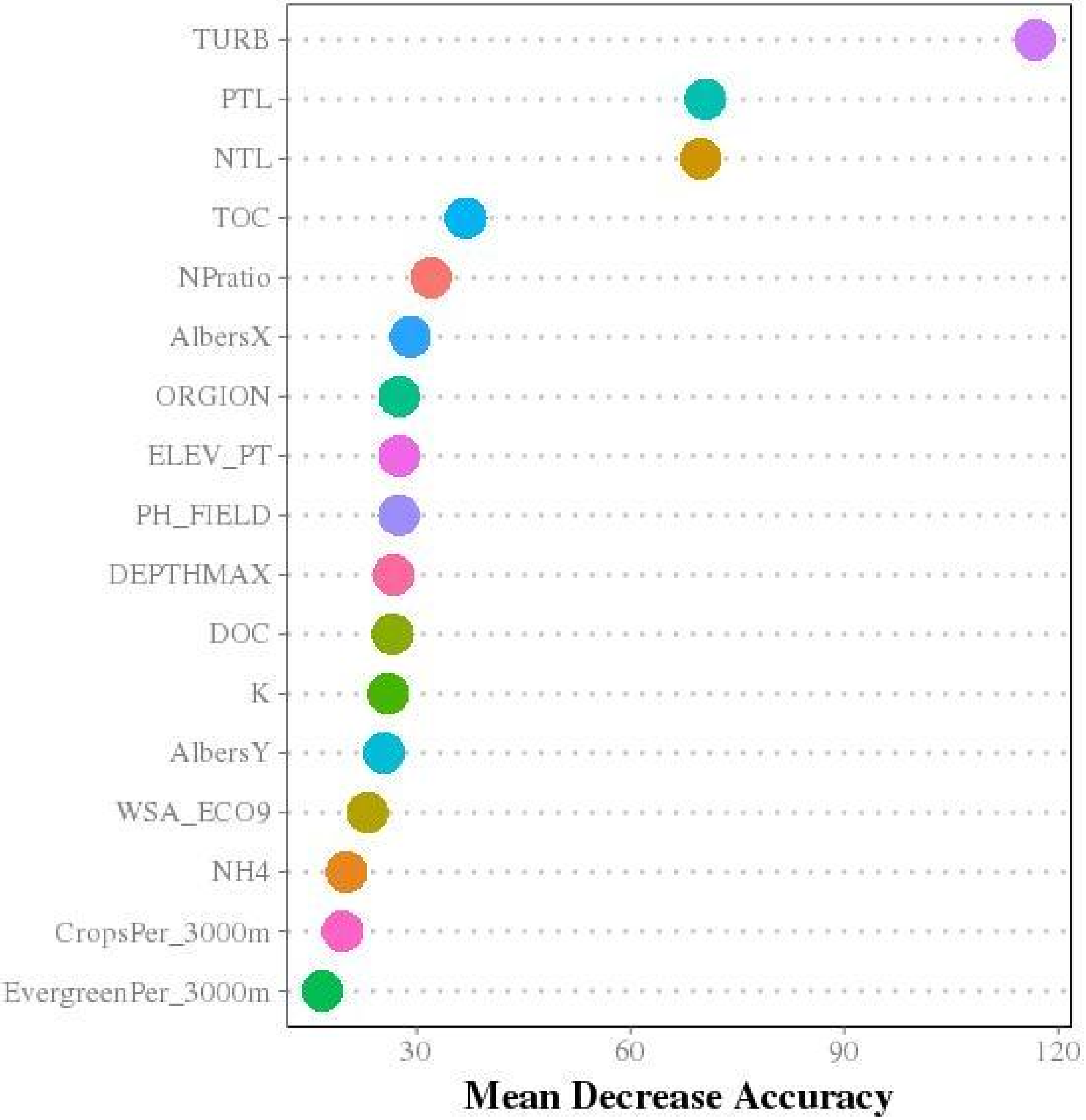
Total accuracy for Model 1 is 0.667% and the Cohen's Kappa is 0.546.

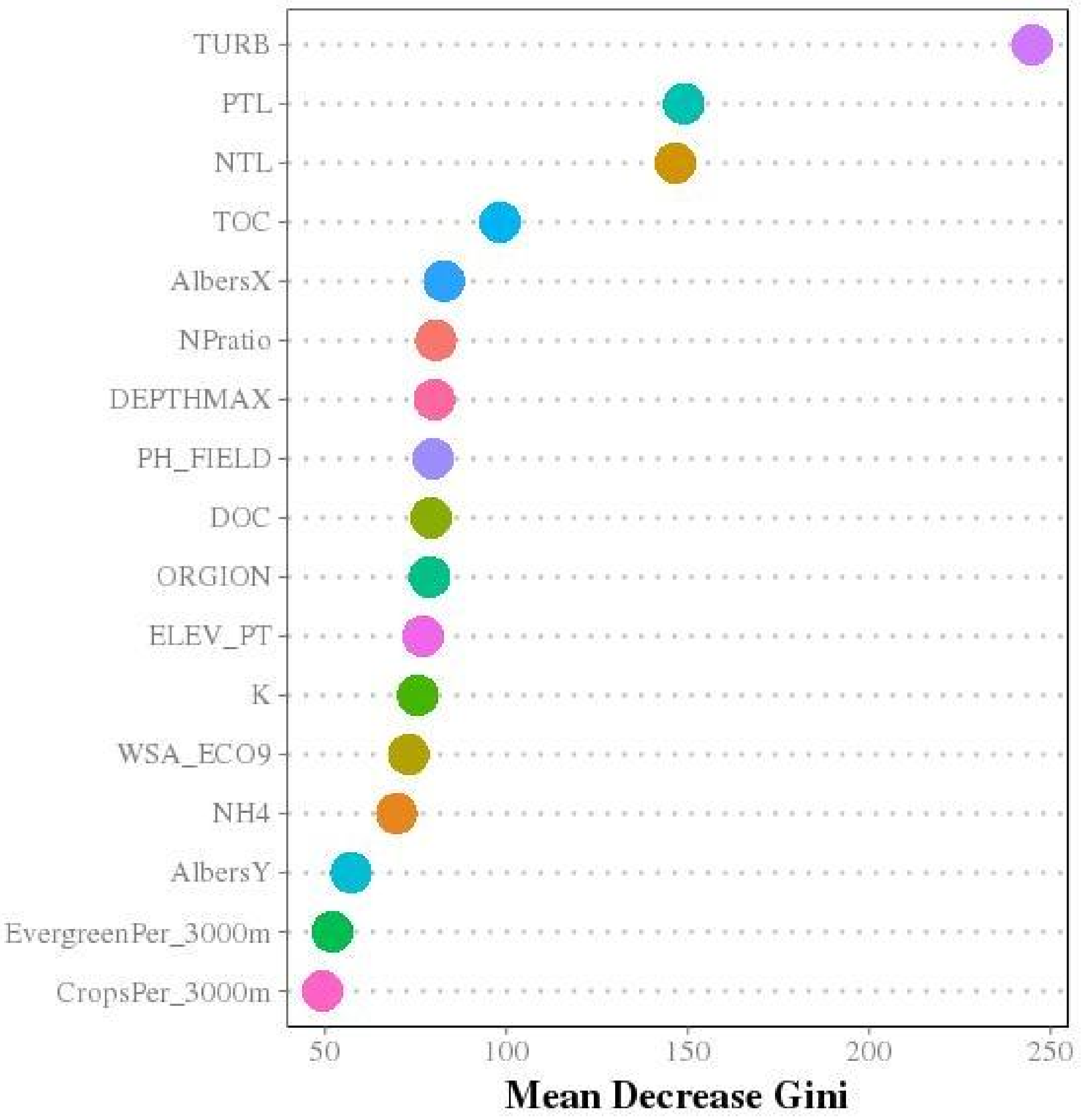
Model 2: 3 Trophic States ~ All Variables

Variable	Percent
DOC	1.00
K	1.00

Variable	Percent
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

|Oligo |Meso/Eu |Hyper |class.error | |:—–|:——-|:—–|:———–| |121 |75 |0 |0.38 | |40 |609 |40 |0.12 | |0 |72 |173 |0.29 |



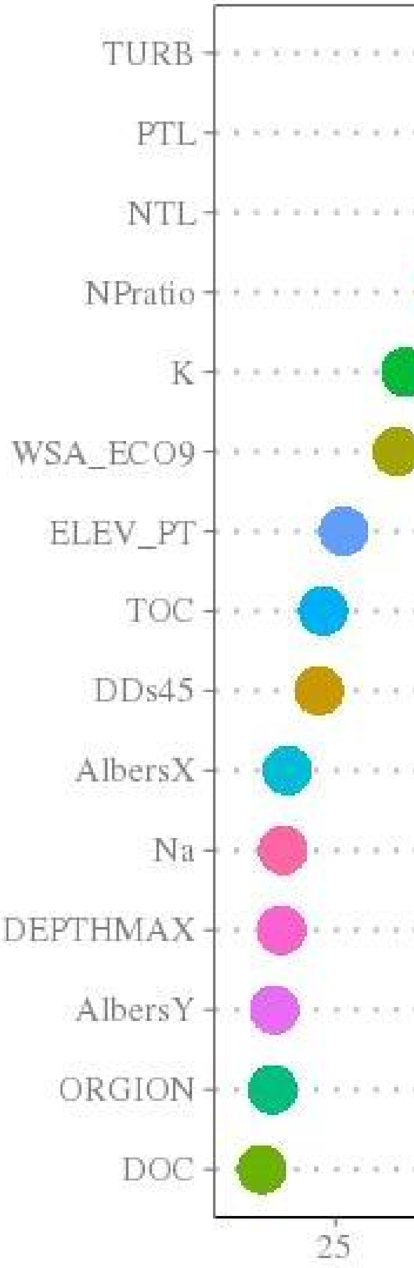


Total accuracy for Model 2 is 0.799% and the Cohen's Kappa is 0.618.

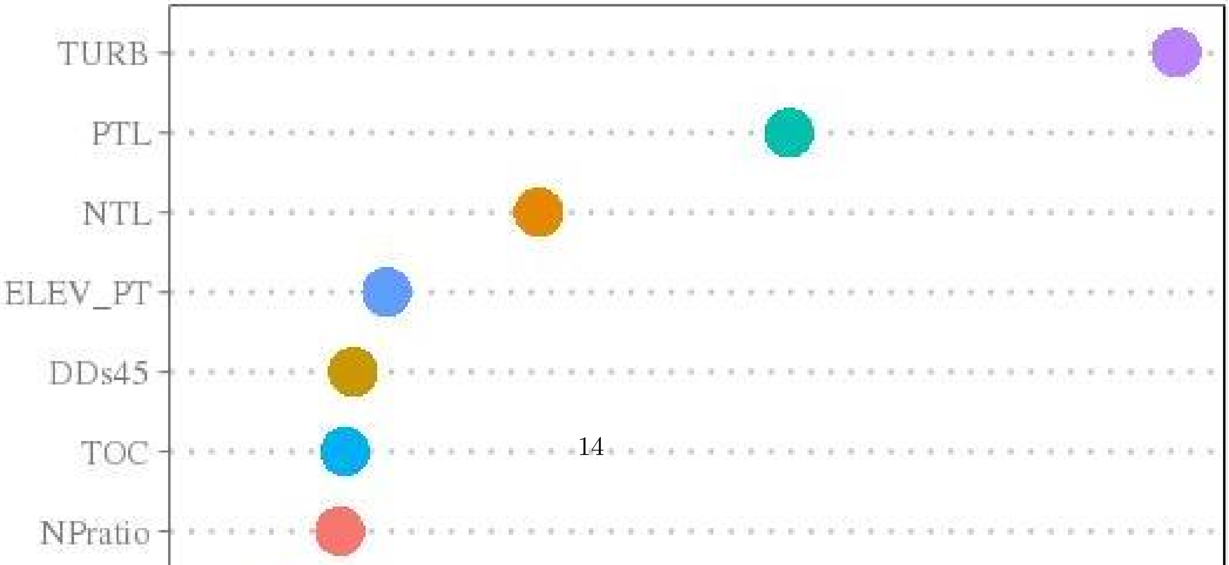
Model 3: 2 Trophic States ~ All Variables

Variable	Percent
K	1.00
NPratio	1.00

Variable	Percent
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



|Oligo/Meso |Eu/Hyper |class.error | |:———-|:——–|:———–| |489 |71 |0.13 | |77 |505 |0.13 |

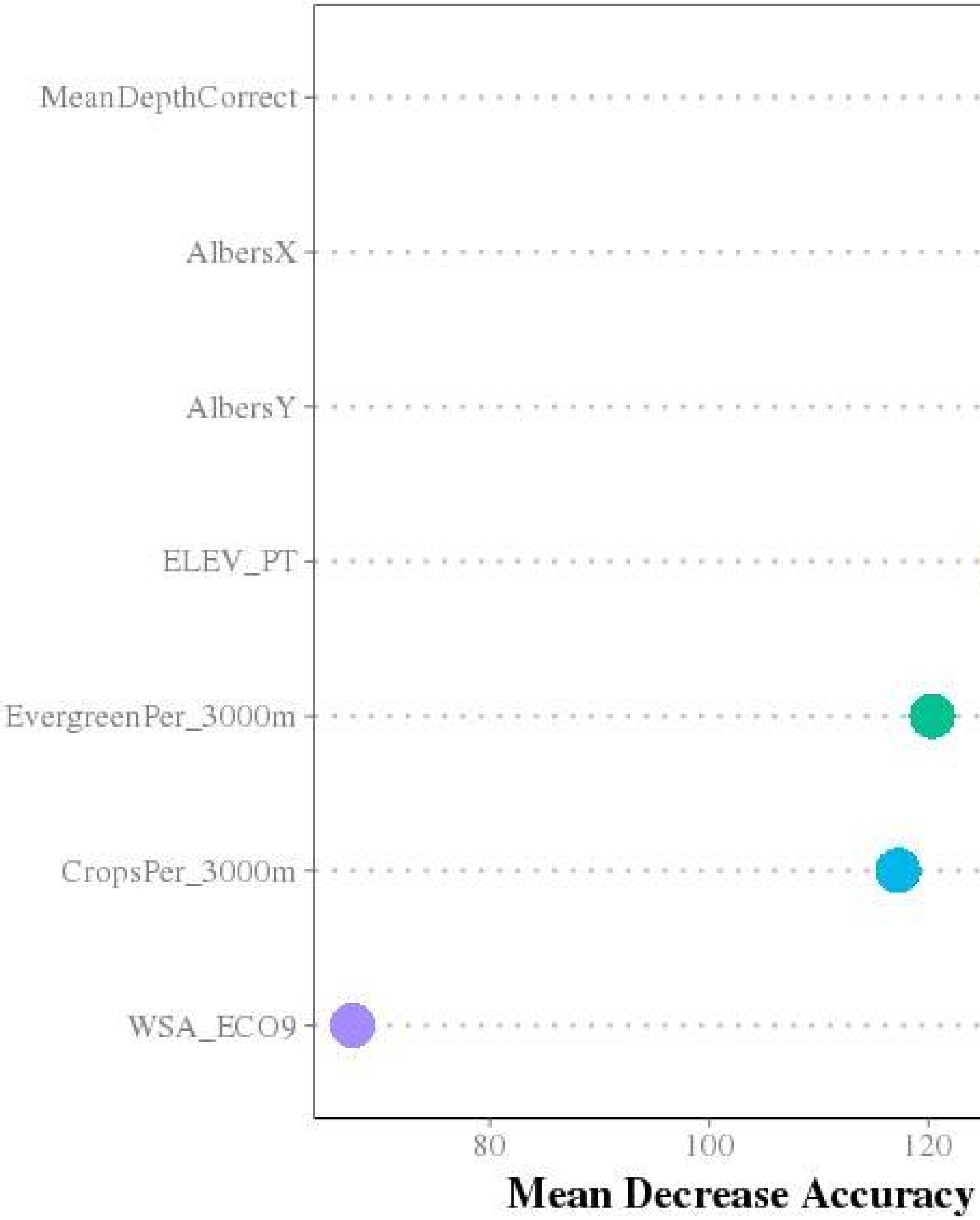


Total accuracy for Model 3 is 0.87% and the Cohen's Kappa is 0.741.

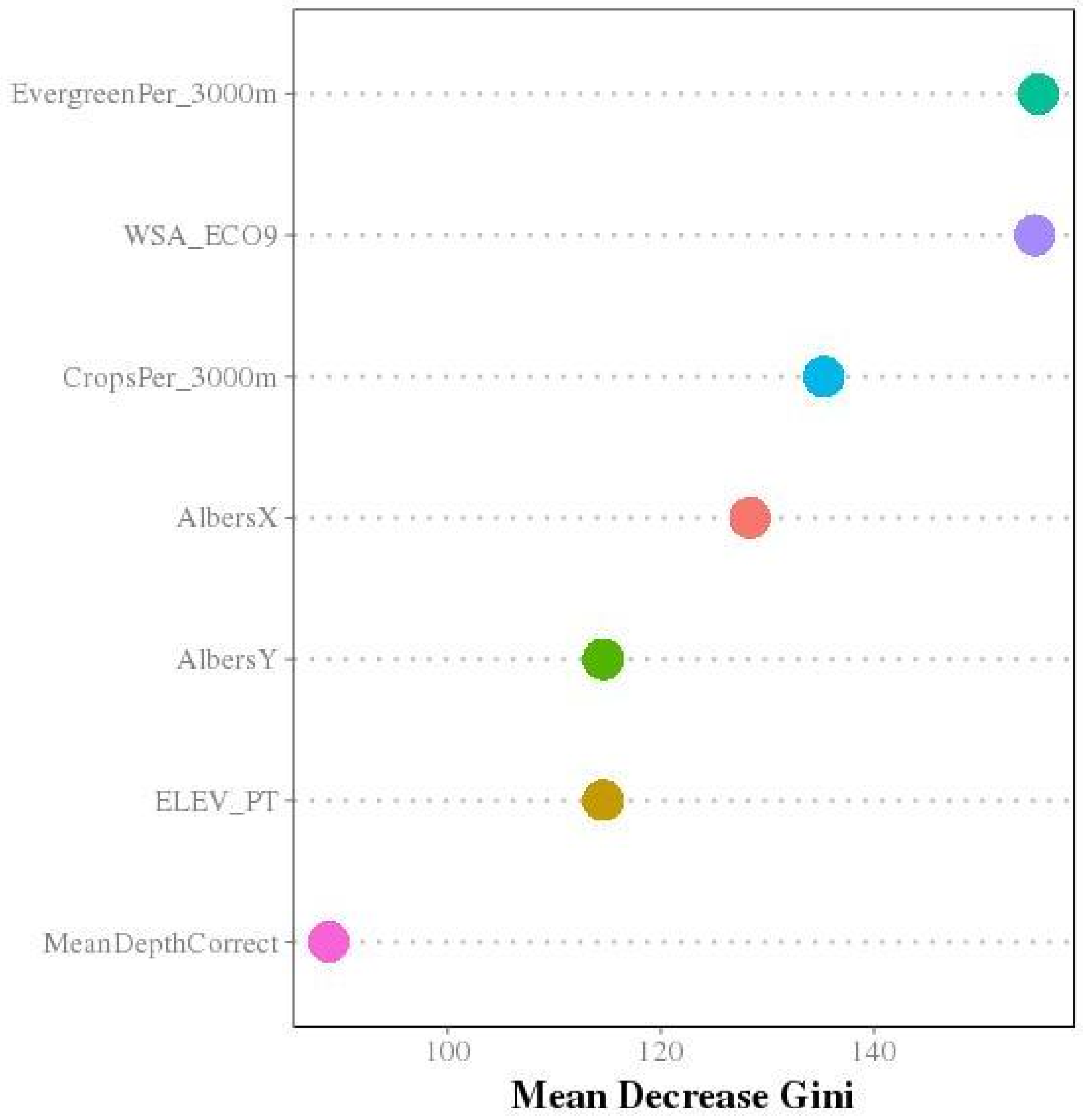
Model 4: 4 Trophic States ~ GIS Only Variables

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

|Oligo |Meso |Eu |Hyper |class.error | |:—–|:—-|:—|:—–|:————| |95 |73 |27 |2 |0.52 | |48 |201 |80 |32 |0.44 | |20 |114



|124 |77 |0.63 | |2 |36 |79 |129 |0.48 |



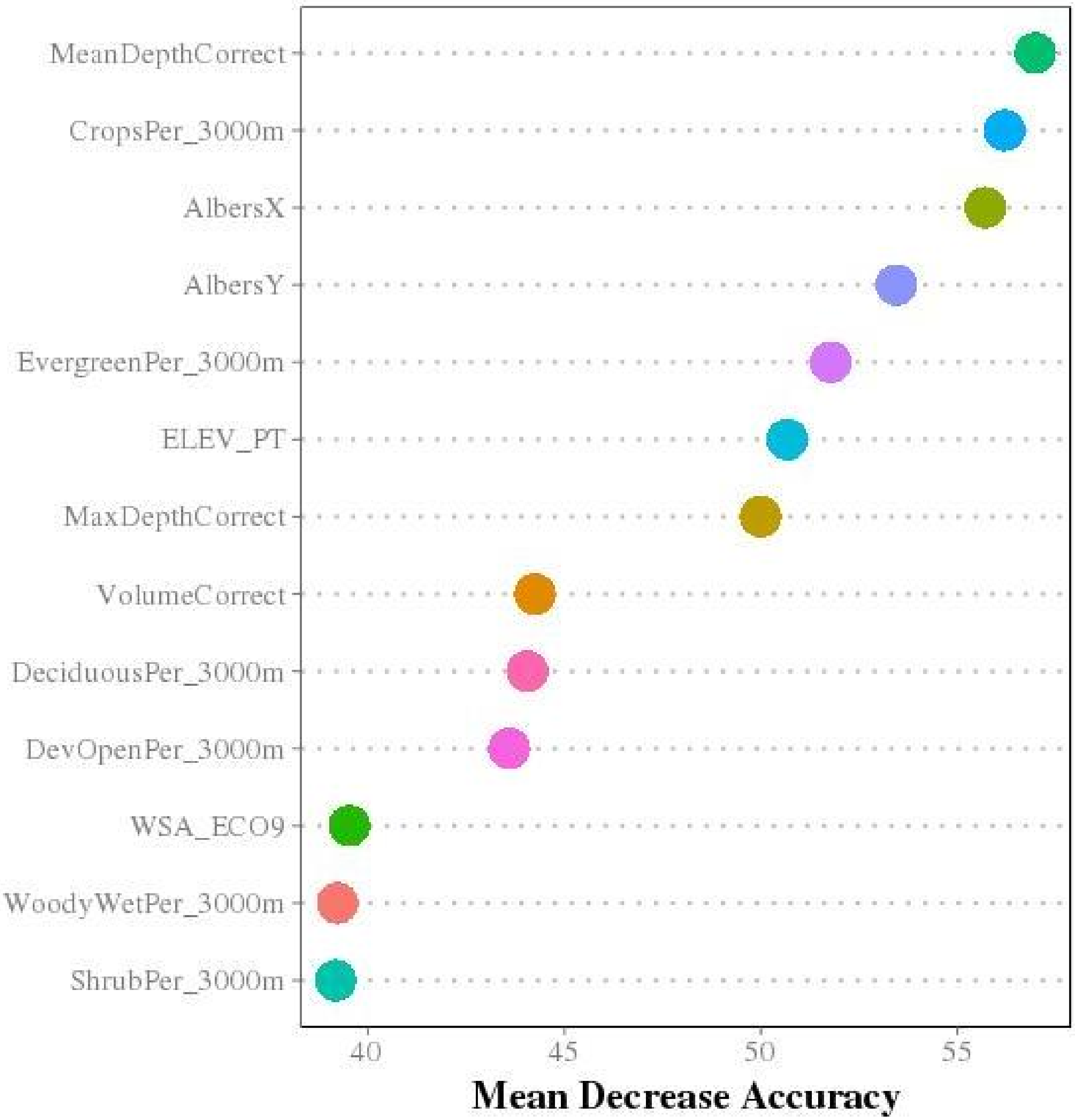
Total accuracy for Model 4 is 0.482% and the Cohen's Kappa is 0.292.

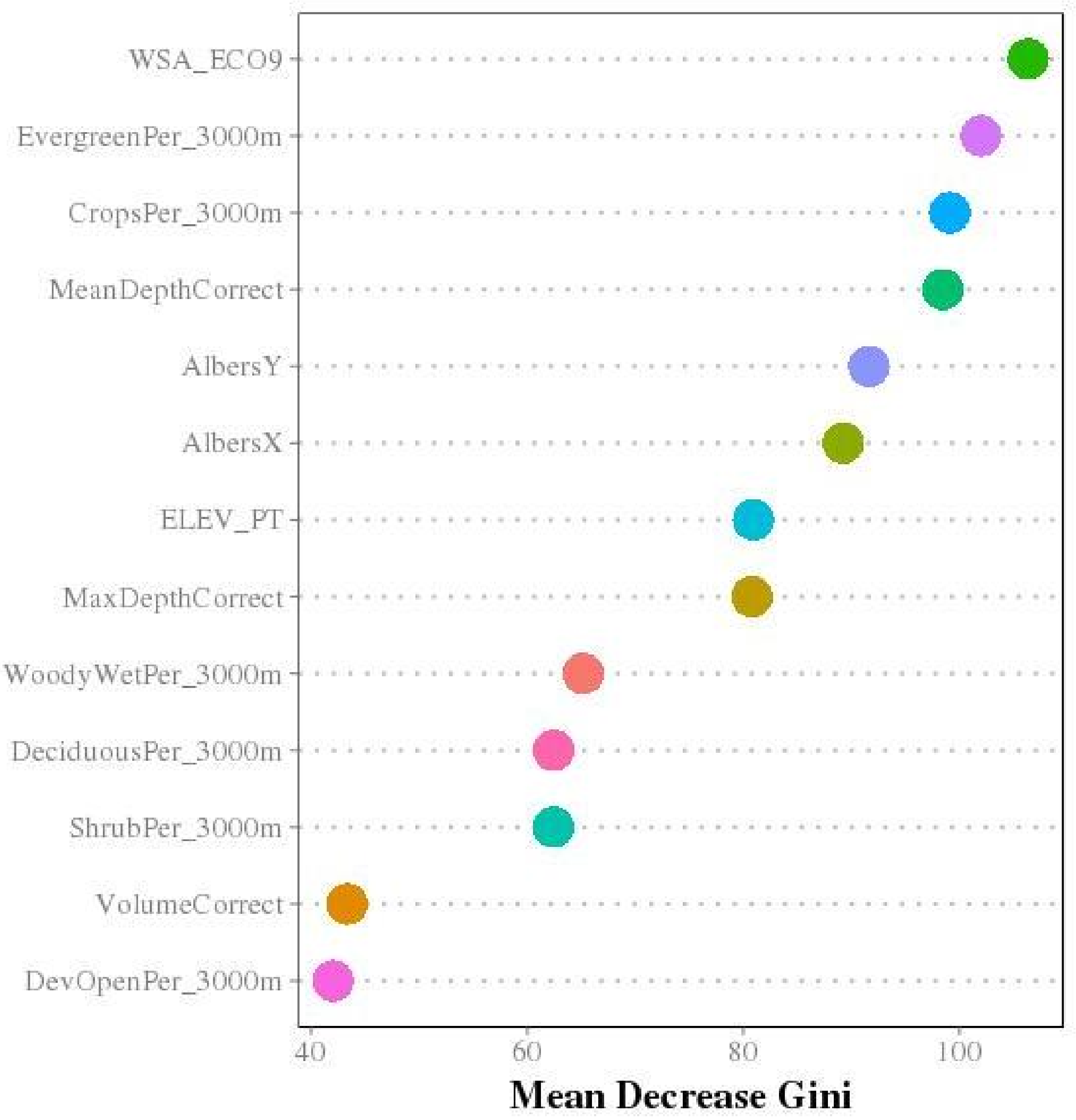
Model 5: 3 Trophic States ~ GIS Only Variables

Variable	Percent
AlbersX	1.00
AlbersY	1.00

Variable	Percent
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

|Oligo |Meso/Eu |Hyper |class.error | |:—|:——-|:—|:———| |79 |116 |1 |0.6 | |48 |582 |66 |0.16 | |0 |141 |105 |0.57 |





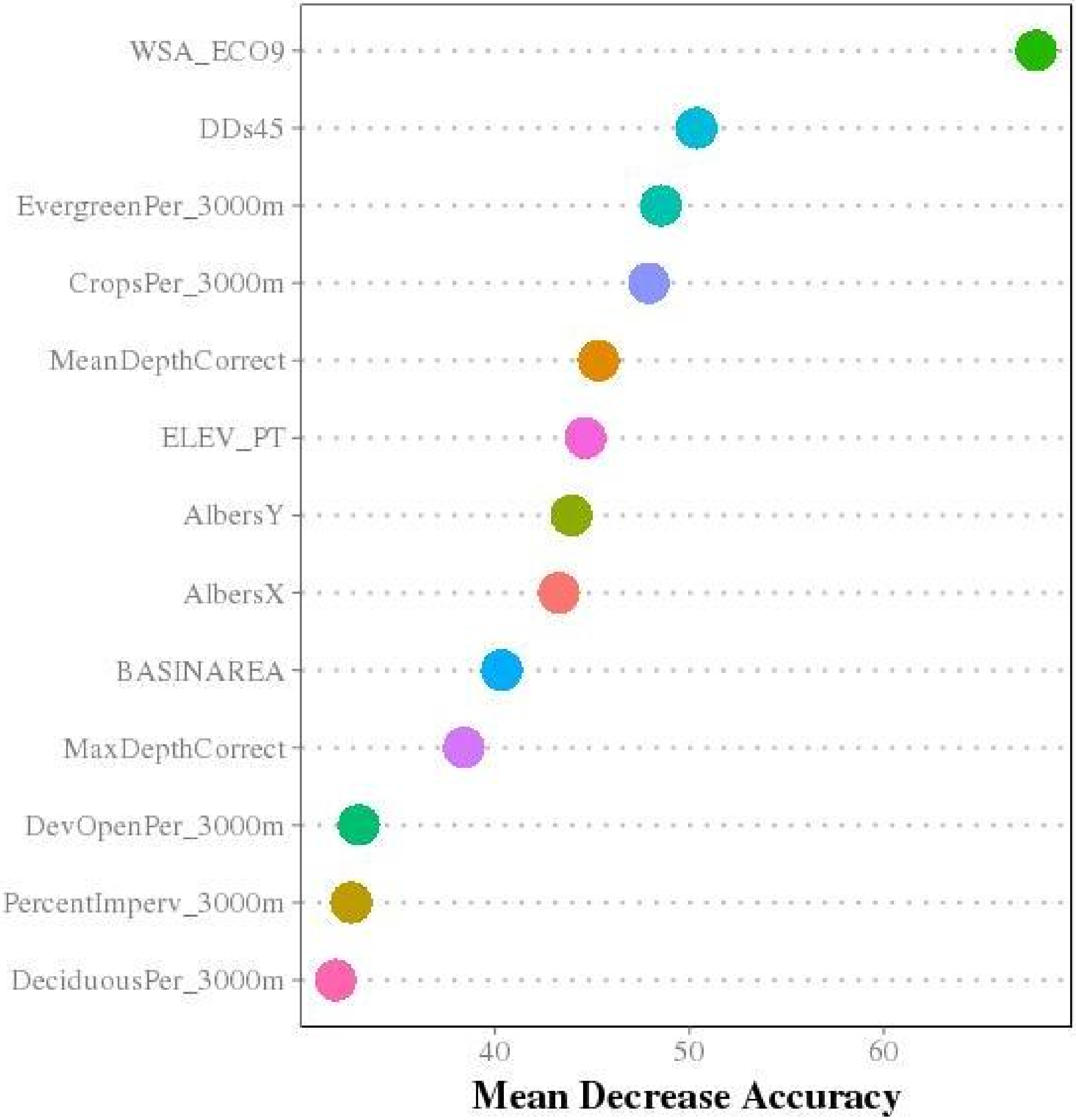
Total accuracy for Model 5 is 0.673% and the Cohen's Kappa is 0.343.

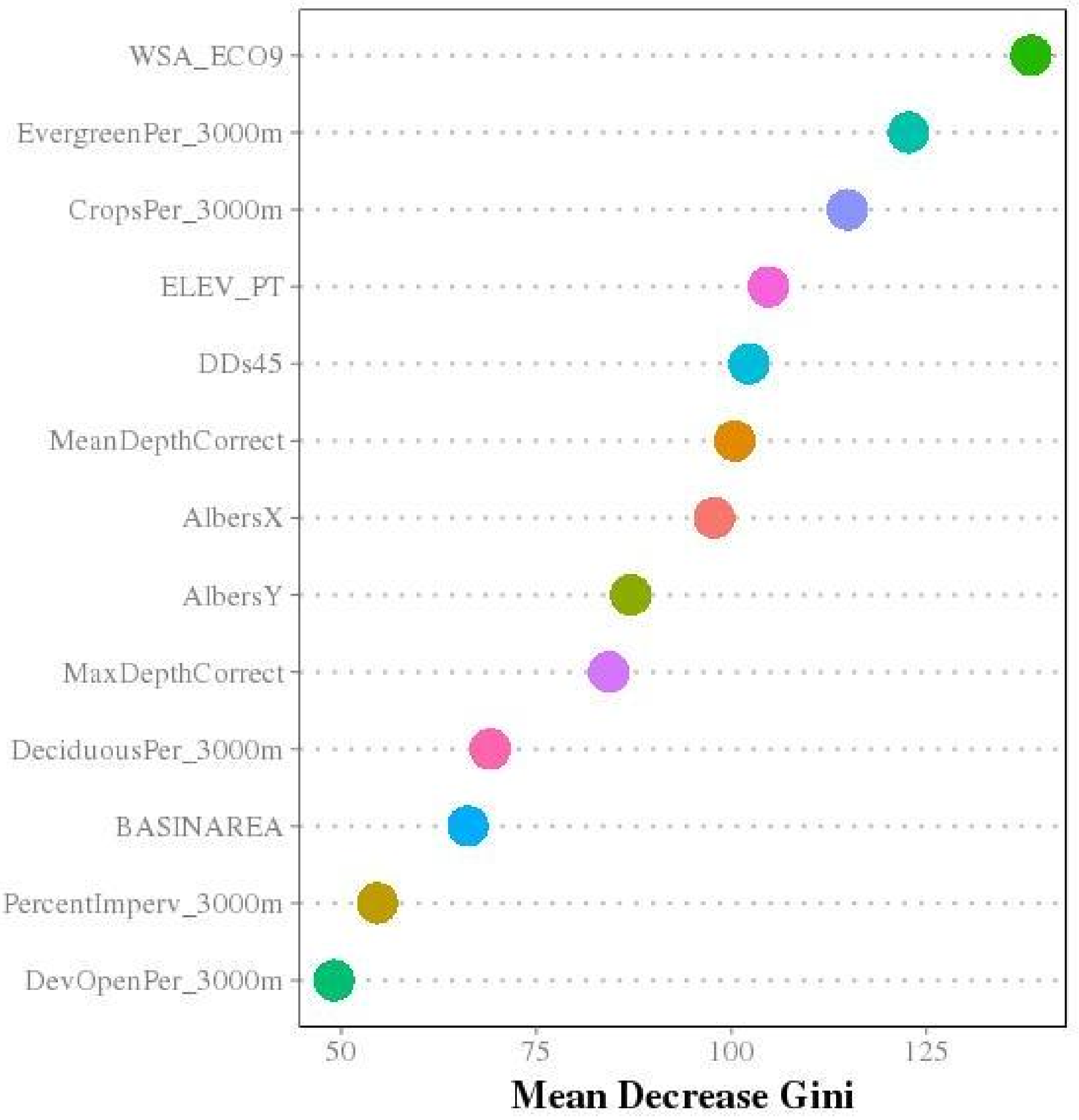
Model 6: 2 Trophic States ~ GIS Only Variables

Variable	Percent
AlbersX	1.00
CropsPer_3000m	1.00

Variable	Percent
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	0.92
DevOpenPer_3000m	0.67
BASINAREA	0.31
PercentImperv_3000m	0.01

|Oligo/Meso |Eu/Hyper |class.error | |:———-|:———|:———-| |428 |129 |0.23 | |146 |435 |0.25 |





Total accuracy for Model 6 0.758% and the Cohen's Kappa is 0.517.

Associating Trophic State and Cyanobacteria

References

1. USEPA (2009) National lakes assessment: a collaborative survey of the nation's lakes. ePA 841-r-09-001.

2. Homer C, Huang C, Yang L, Wylie B, Coan M (2004) Development of a 2001 national land-cover database for

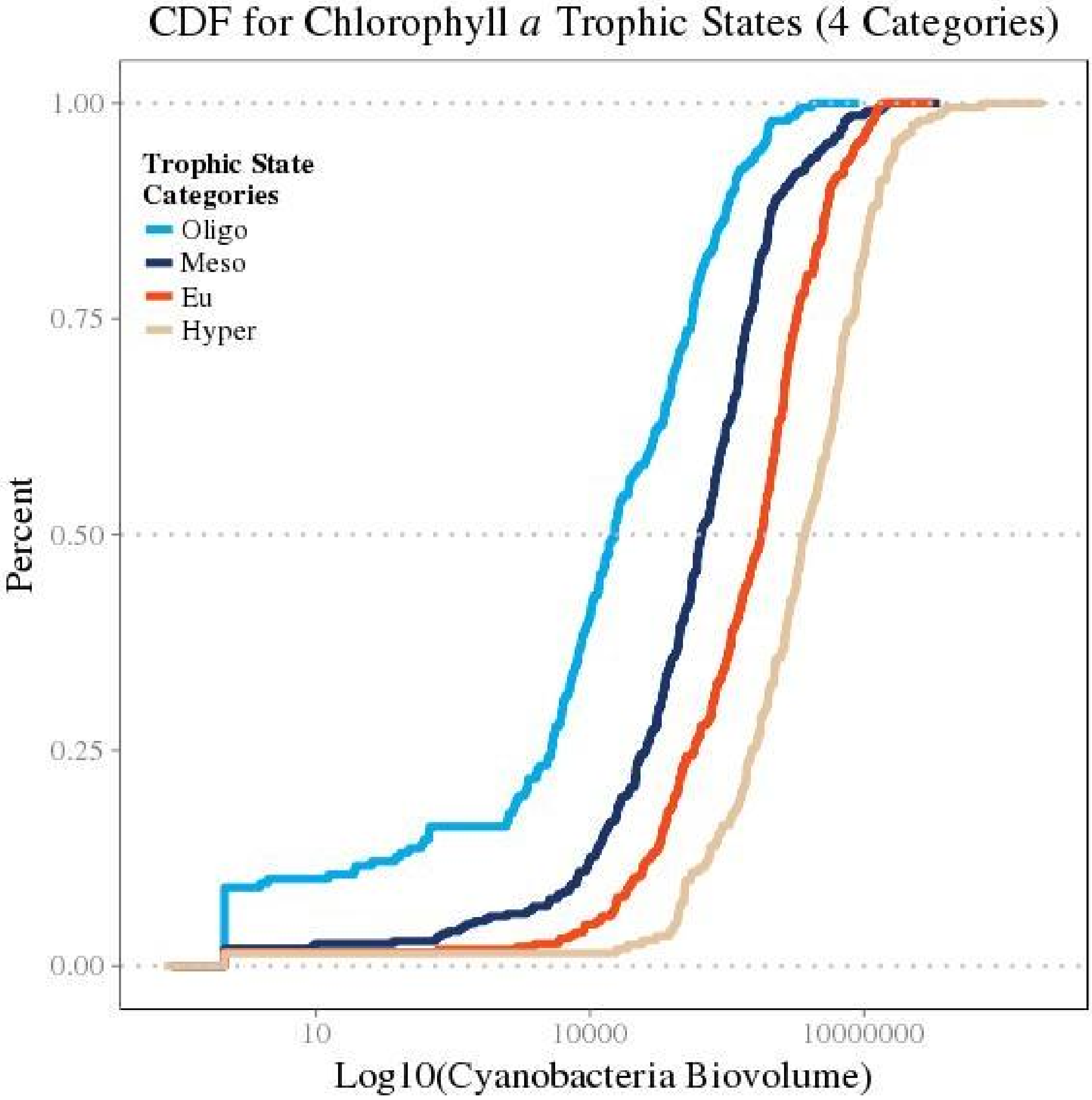


Figure 1: plot of chunk ts_4_biov

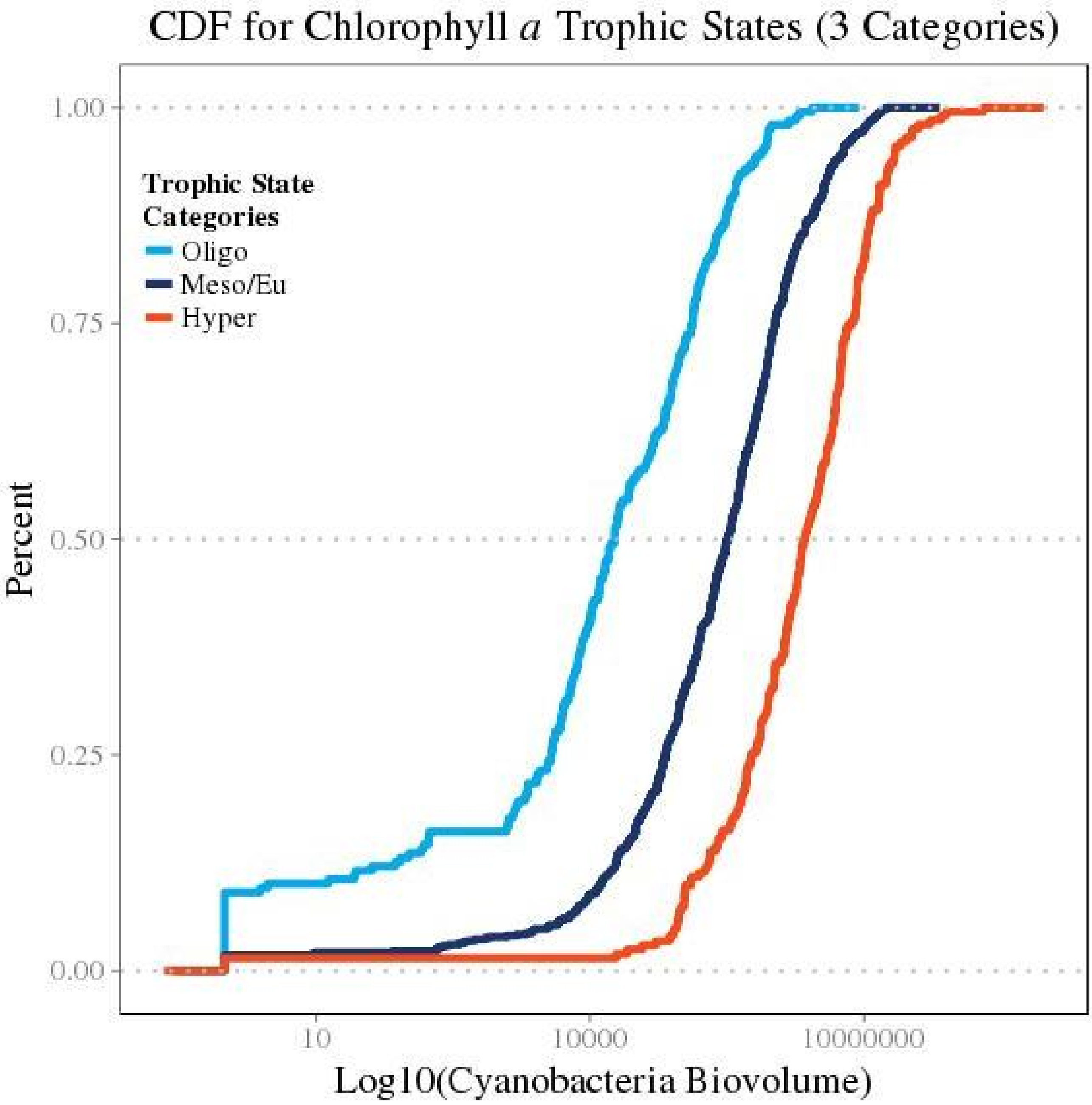


Figure 2: plot of chunk ts_3_biov

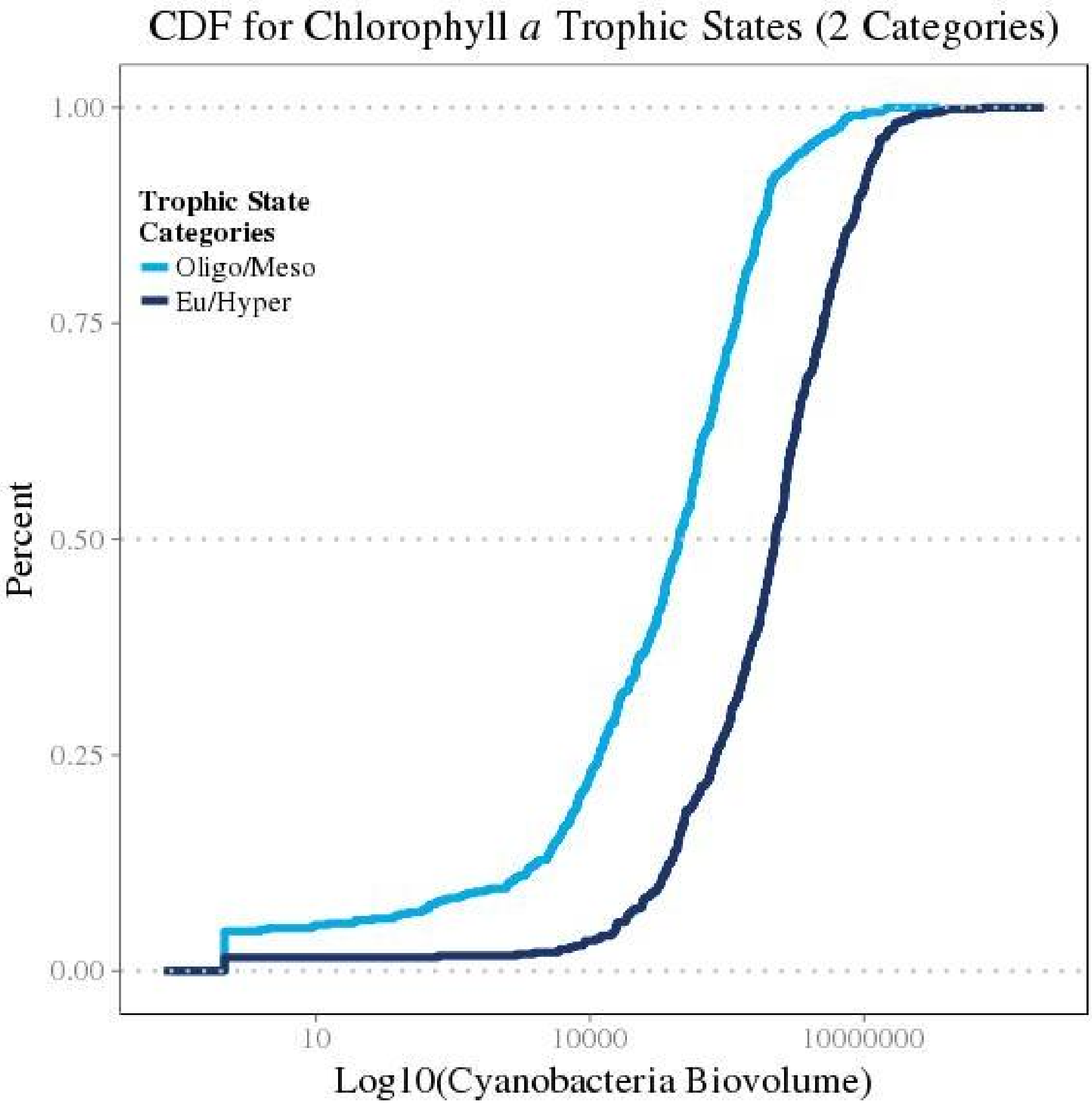


Figure 3: plot of chunk ts_2_biov

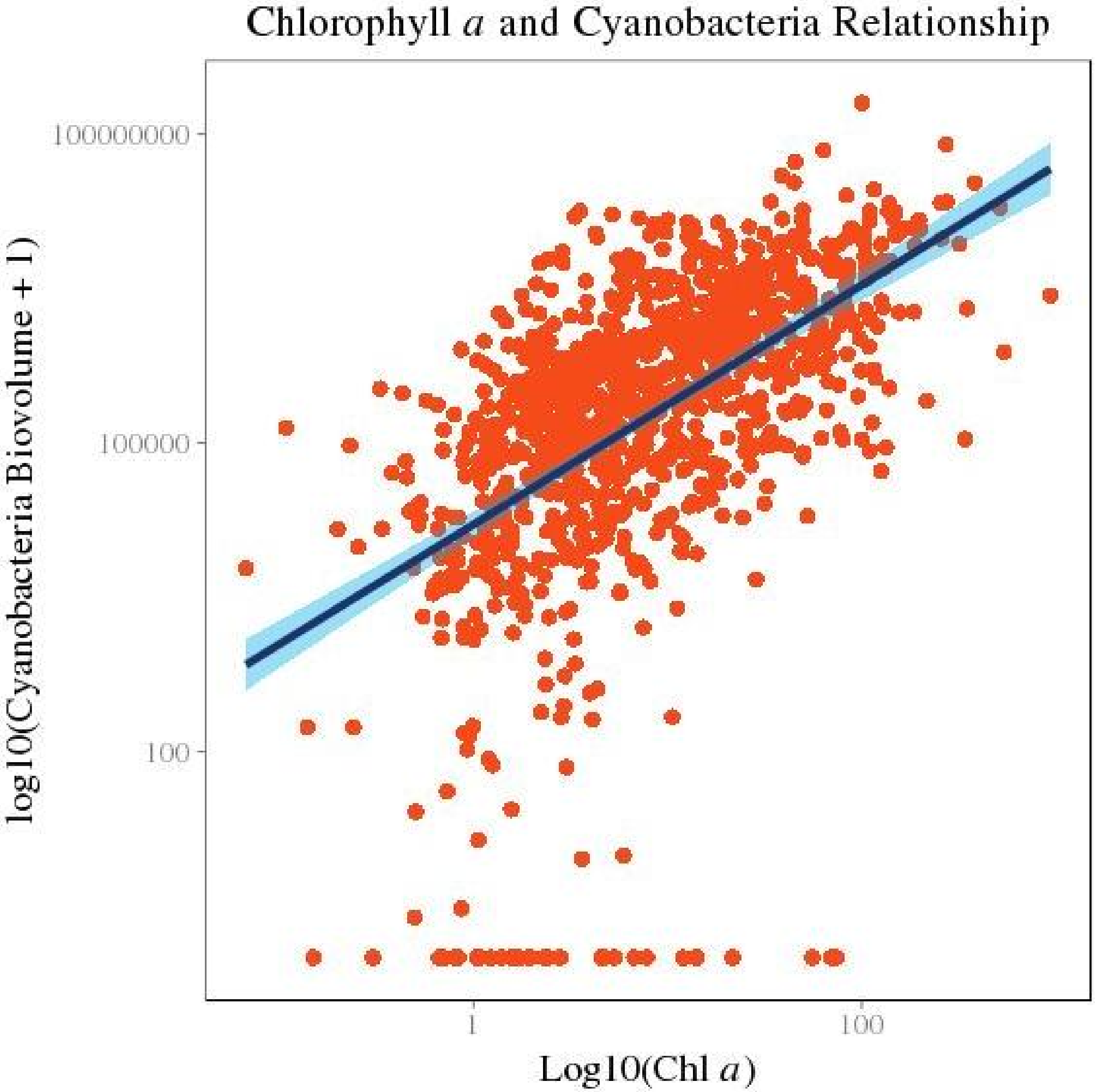


Figure 4: plot of chunk scatterplot

the united states. Photogrammetric Engineering & Remote Sensing 70: 829–840.

3. Xian G, Homer C, Fry J (2009) Updating the 2001 national land cover database land cover classification to 2006 by using landsat imagery change detection methods. Remote Sensing of Environment 113: 1133–1147.

4. Hollister J, Milstead WB (2010) Using gIS to estimate lake volume from limited data. Lake and Reservoir Management 26: 194–199.

5. Hollister JW, Milstead WB, Urrutia MA (2011) Predicting maximum lake depth from surrounding topography. PLoS ONE 6: e25764. Available: http://dx.doi.org/10.1371/journal.pone.0025764. Accessed 28 Jun 2013.

6. Hollister J (2013) lakemorpho: Lake morphometry in r. Available: http://www.github.com/USEPA/lakemorpho.

7. Hollister JW, Milstead WB (In Preparation) National lake morphometry dataset v1.0.

8. Beaulieu M, Pick F, Gregory-Eaves I (2013) Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. Limnol. Oceanogr 58: 1736–1746.

9. Breiman L (2001) Random forests. Machine learning 45: 5–32.

10. Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. BMC bioinformatics 7: 3.

11. Diaz-Uriarte R (2010) varSelRF: Variable selection using random forests. Available: http://CRAN.R-project.org/package=varSelRF.

12. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2: 18–22. Available: http://CRAN.R-project.org/doc/Rnews/.