coastal\_ts\_variable\_selection

Bryan Milstead

21 February 2018

## Introduction

We are developing a trophic state index for transitional (estuarine and coastal) waters.

## Data

* Data [CoastalWQ\_20170928.csv] and code [betty\_code\_20171004.r], received from Betty on 20180207
* using Betty code, download and organize the data
  + rename some fields
  + filter for year == 2010, region != "Insular", region != "Great Lakes", col\_loc == "SURFACE", and visnum == 1
  + restrict to complete cases

## Analysis

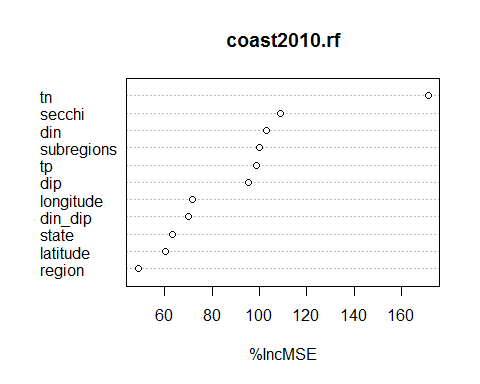
* run random forest
  + save as coast2010rf.rda
* create the variable importance plot & table

## Results

* response variable is "chla"
* predictor variables and their overall imporatance values are shown below in table format (c"oast2010\_rf\_imp.csv")

|  |  |  |
| --- | --- | --- |
| variable | per\_inc\_mse | IncNodePurity |
| tn | 0.2735041 | 79.81022 |
| subregions | 0.1254529 | 50.64981 |
| tp | 0.1062012 | 30.64000 |
| secchi | 0.1028563 | 39.50384 |
| state | 0.0899585 | 35.95041 |
| din | 0.0544333 | 17.62961 |
| dip | 0.0532207 | 15.47141 |
| latitude | 0.0474556 | 17.34413 |
| region | 0.0400098 | 10.87307 |
| longitude | 0.0389727 | 14.93307 |
| din\_dip | 0.0270457 | 12.82000 |

* importance values shown graphically ("coast2010\_variable\_importance.jpeg")



Variable Selection

The goal of variable selection is to identify an optimal reduced subset of predictor variables. Here we used the results from random forest modeling as a means of variable selection. Random forest modeling is a machine learning algorithm that builds numerous statistical decision trees in order to attain a consensus predictor model (Breiman, 2001). Each tree is based on recursively bootstrapped data, and the out-of-bag (OOB) data, cases left out of the sample, provides an unbiased estimation of model error and measure of predictor variable importance. Random forest modeling was conducted in randomForest package in R (Liaw and Wiener, 2002; R Core Team, 2016). We developed random forest models to select predictor variables to model trophic state, nitrogen, and phosphorus. The random forest model for trophic state included in situ water quality data and universally available GIS data, e.g. landscape data (see Hollister et al. (2016) for detailed methods). The models for nitrogen and phosphorus included only universally available GIS data. We used percent increase in mean square error to examine variable importance.