coastal\_ts\_variable\_selection

Bryan Milstead

21 February 2018

## Introduction

We are developing a trophic state index for transitional (estuarine and coastal) waters. Part of this process involves selection of predictor variables. We use random forest to determine the most imporant variables.

## 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

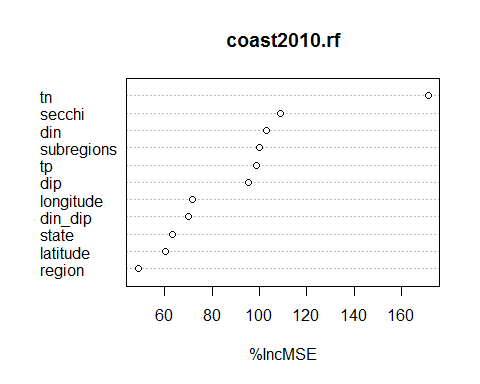
* run random forest
  + save as coast2010rf.rda
* create the variable importance plot & table

## Results

* response variable is "chla"
* predictor variables and their overall imporatance values are shown below in table format ("coast2010\_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 first step in the modelling process is to decide which variables to include. Of the hundreds of measurements taken during the National Coastal Assessment eleven were selected *apriori* as being potentially useful in predicting the trophic state of transitional waters. These variables are shown in Table 1. Random forest modeling was used to select an optimal reduced subset of the original predictor variables. 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 R (R Core Team 2016) with the randomForest package (Liaw and Wiener 2002). The random forest model with 10,000 trees was used to determine which of the original eleven variables had the most impact on predicting Chlorophyll *a* values, a proxy for trophic state. Based on the model, each variable was assigned an importance value based on their contribution to the mean standard error. The importance measures are shown in Table 1 and Figure 1. From Figure 1 it is apparent that the data fallinto at least three groups. Total nitrogen (tn) has the greatest impact followed by five variables of midlevel and five of lower level importances (Table 1 and Figure 1). The six variables with high or mid-level importances were retained for model development and the five lowest importance variables were eliminated.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Liaw, A. andWiener,M. (2002). Classification and regression by randomforest. r news 2 (3): 18–22. URL: <http://CRAN>. R-project. org/doc/Rnews.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.