coastal\_ts\_variable\_selection

Bryan Milstead

26 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”
* randomForest provides two measures of variables importance (“%IncMSE” and “IncNodePurity”)and the SD of importance (importanceSD); we are using “%IncMSE” but change the name to “per\_inc\_mse”
* a scaled variable importance is caluclated by dividing per\_inc\_mse / importanceSD (or use the function [importance])
  + below are information on the importance values from ?importance

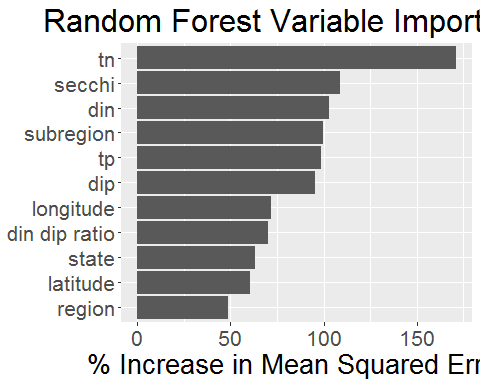
Here are the definitions of the variable importance measures. The first measure is computed from permuting OOB data: For each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification, MSE for regression). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees, and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done (but the average is almost always equal to 0 in that case).

The second measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. For regression, it is measured by residual sum of squares.

* predictor variables, and importance and scaled importance values are shown below (“coast2010\_rf\_imp.csv”)

|  |  |  |
| --- | --- | --- |
| variable | per\_inc\_mse | IncNodePurity |
| tn | 171.25232 | 79.81022 |
| secchi | 108.91501 | 39.50384 |
| din | 102.87908 | 17.62961 |
| subregion | 99.92517 | 50.64981 |
| tp | 98.54266 | 30.64000 |
| dip | 95.42217 | 15.47141 |
| longitude | 71.75630 | 14.93307 |
| din dip ratio | 70.13108 | 12.82000 |
| state | 63.08289 | 35.95041 |
| latitude | 60.37790 | 17.34413 |
| region | 48.89239 | 10.87307 |

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



## Variable Selection Paragraph

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 the percent reduction in to the OOB mean squared error for the (averaged over all trees and normalized by dividing by the standard deviation). 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.

## References

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.