Random Forest Model of Biogeoclimatic Units for Western North America

William H MacKenzie

22/03/2020

Points from a 8km hex grid of western north america are generated in R and submitted to ClimateNA to extract annual and seasonal variables for the historic normal period (1961-90) and an ensemble future climate scenario (rcp45 2040-2070). These data sets are combined. Several additional climate variables are generated to test in the variable selection step including several monthly climate sums for precipitation and growing degree days. All other monthly variables are removed. A winter rechange moisture deficit was calculated and summed with the climatic moisture deficit to account for regions that begin the growing season in soil moisture deficit.

The preprocessing function from the caret package was used to identify variables with near-zero variance or correlation >0.95 in the combined data set. These variables were removed leaving a final variable set of 23 variables.

A larger training point set was built from 2 km hex grid of points for western north america. Points are overlayed the WNA Biogeoclimatic map to identify BGC membership.

The 1,175,312 training point population is highly imbalanced between the 362 biogeoclimatic subzone/variants. Geographically large BGCs dominate the machine learning model when raw data set are used in the training set. A partial balancing of the raw training point between classes was applied to counteract this effect. Rebalancing is accomplished by log10 scaling the traininpoint population and then rescaling the logscale to a range of 500 - 1200 training points per BGC. This identifies the desired rebalanced training sample size per BGC. Large BGCs with raw training sets larger than the scaled sample are subsampled using conditioned Latin Hypercube Sampling. Geographically restricted BGCs are upsampled using the SMOTE routine.

Large BGCs are subsampling using a cLHS of the 23-variable space for each BGC removing training points sown to that specified in the rescaling.

Under trained BGC units are upsampled using the SMOTE to add synthetic points up to the number specified in the resampling.

The new rebalanced training set is submitted to ranger to generate a final climate model of BGCs for western north america.

## Ranger result  
##   
## Call:  
## ranger(BGC ~ ., data = X2\_final[, -c(1)], num.trees = 151, importance = "impurity", splitrule = "extratrees", seed = 12345, write.forest = TRUE, classification = TRUE)   
##   
## Type: Classification   
## Number of trees: 151   
## Sample size: 245790   
## Number of independent variables: 23   
## Mtry: 4   
## Target node size: 1   
## Variable importance mode: impurity   
## Splitrule: extratrees   
## Number of random splits: 1   
## OOB prediction error: 17.28 %

