Error analysis and variable significance with random forests

Ned Horning

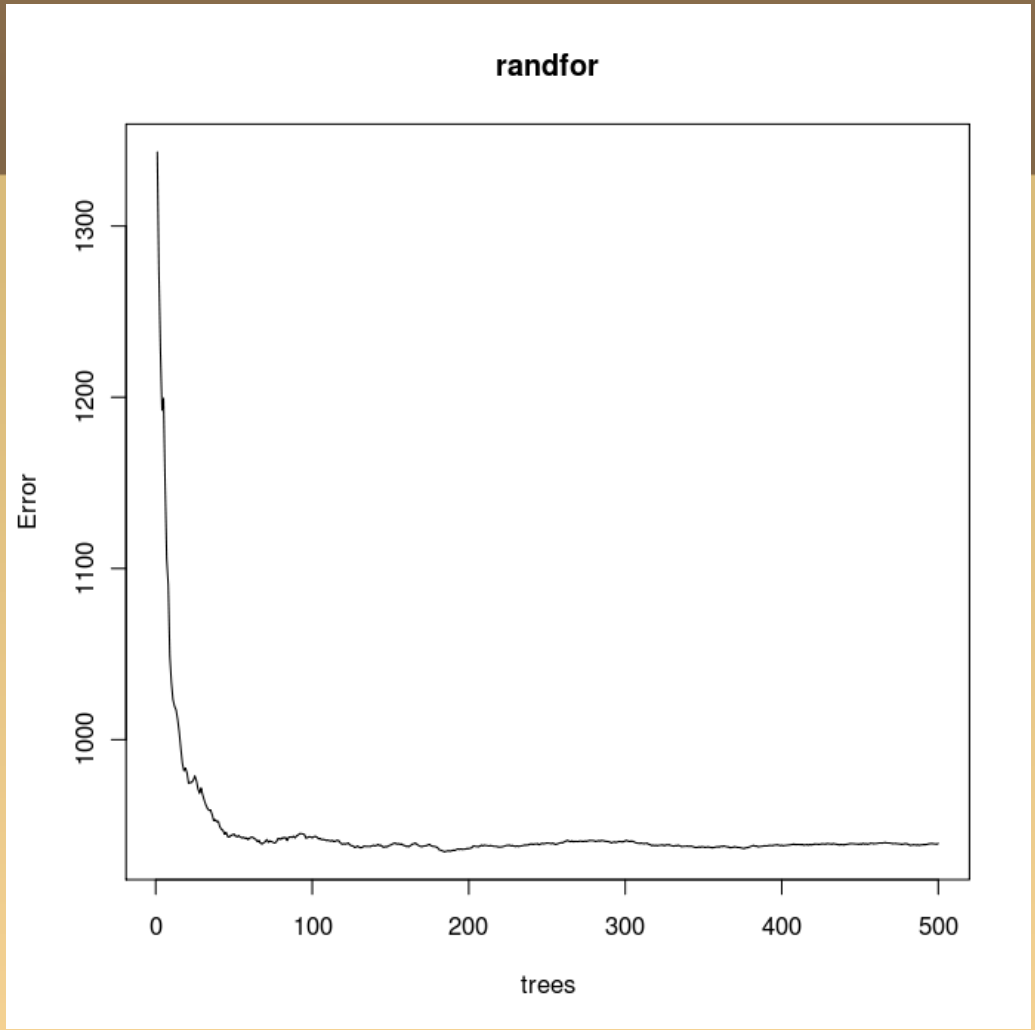
American Museum of Natural History's
Center for Biodiversity and Conservation

horning@amnh.org

Center for Biodiversity and Conservation
American Museum of Natural History

Error estimate

- Provides and unbiased estimate of the error
- Each tree uses a different bootstrap sample (~1/3 of samples) for testing
- Use function print() for OOB error estimate
- Use plot() to view plot of error estimate vs. number of trees



Error rate vs. number of trees

Calculate OOB error estimate

- Put OOB samples down tree after it is constructed and keep track of results
- Proportion of times the result is not accurate averaged over all samples is the OOB error estimate
- For regression “percent variance explained” is also called pseudo R-squared

OOB estimate of error rate: 0.1%
Confusion matrix:

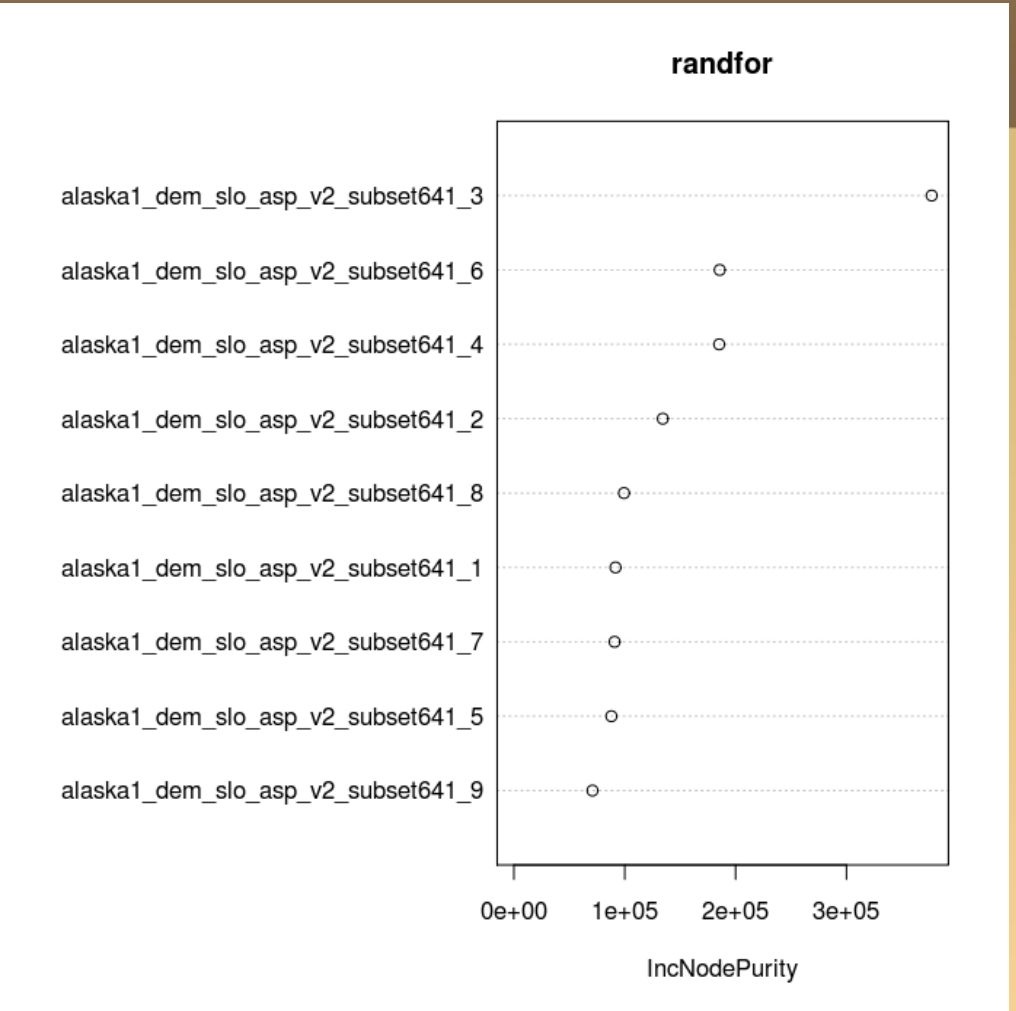
	1	2	3	4	class.error
1	999	1	0	0	0.001
2	3	997	0	0	0.003
3	0	0	1000	0	0.000
4	0	0	0	1000	0.000

Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 3

Mean of squared residuals: 677.6654
% Var explained: 60.66

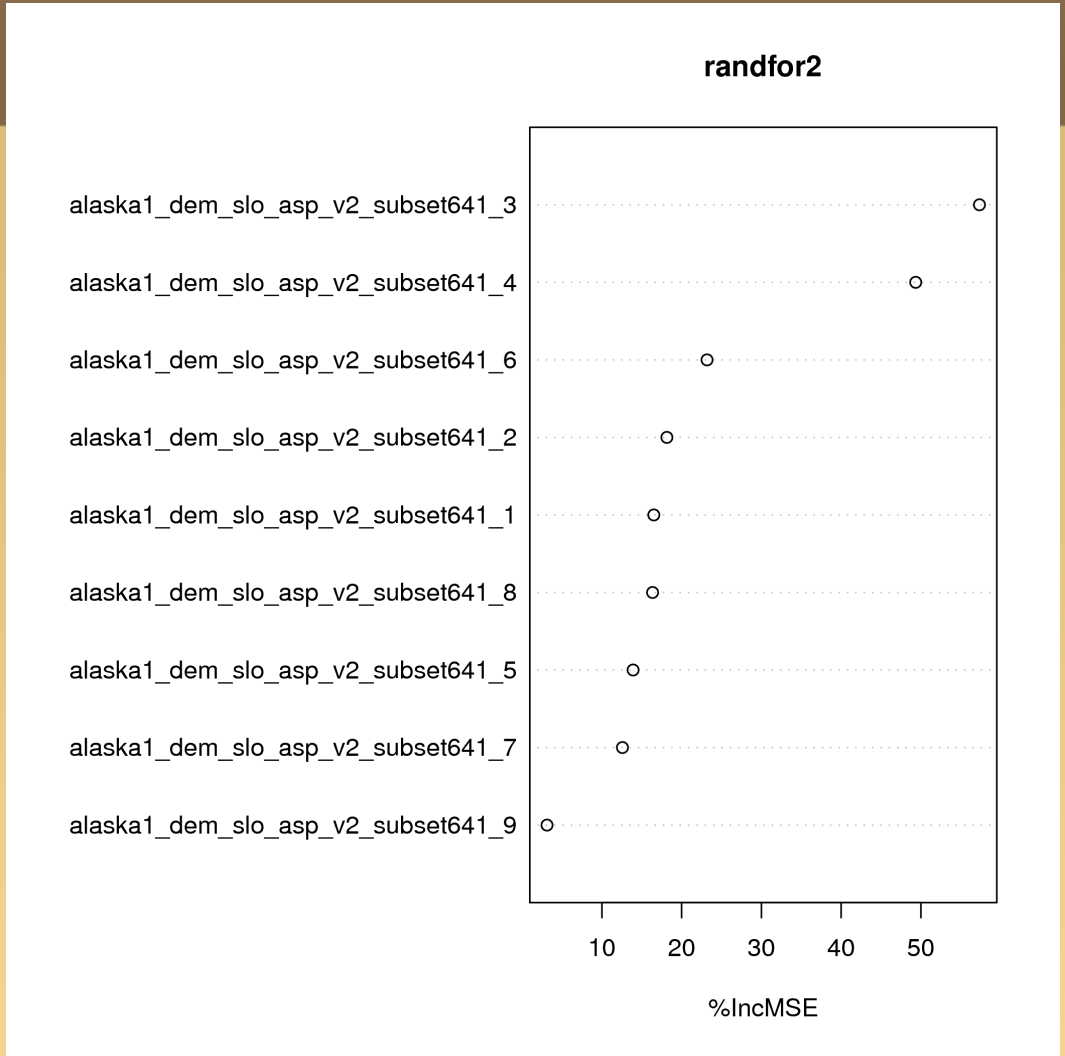
Variable importance

- Put oob samples down a tree then for each variable randomly reorder that variable in each of the oob samples and put these down the trees
- Two types of error can be calculated: mean decrease in accuracy and mean decrease in node impurity
- Actual measures depend if it is classification or regression



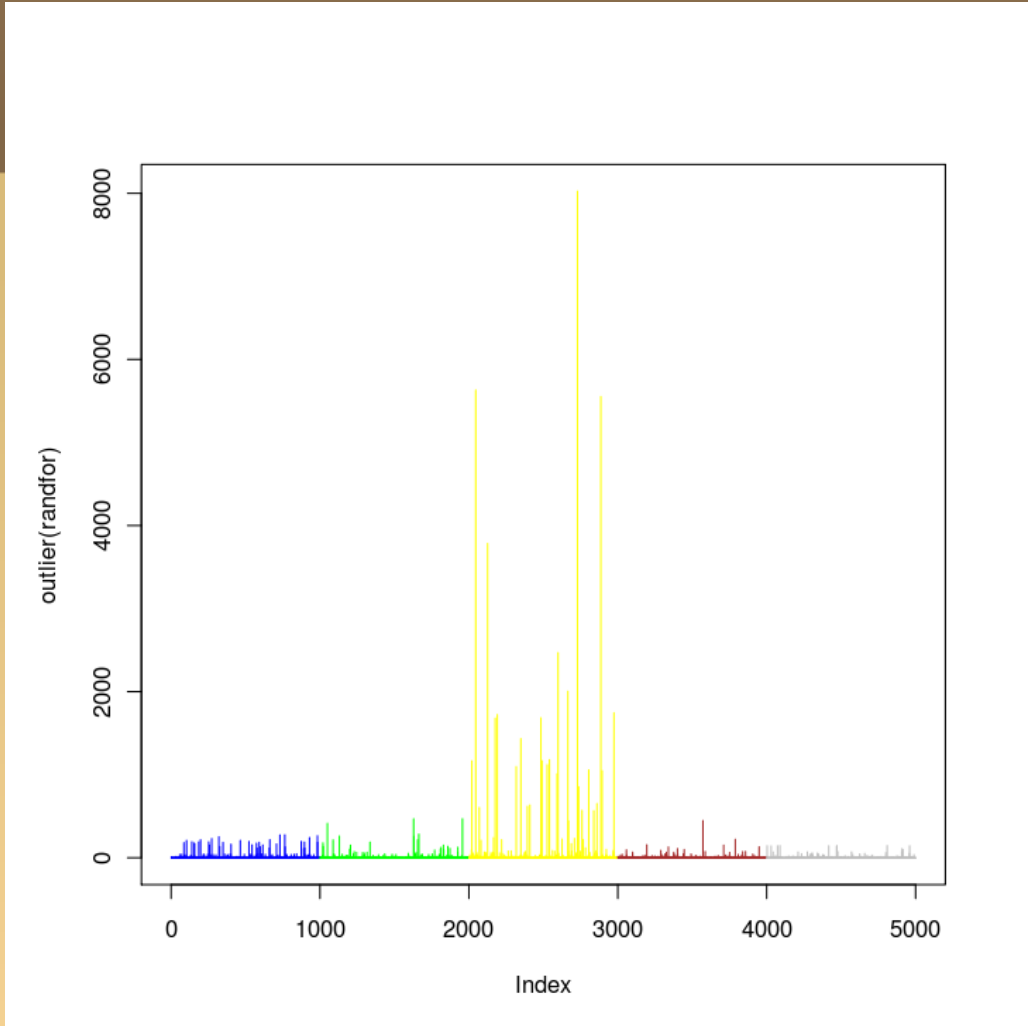
Variable importance steps

- In the randomForest() function specify “importance=TRUE”
- importance() function creates an importance object
- varImpPlot() function plots variable importance
- Specify type = 1 for mean decrease in accuracy and 2 for mean decrease in node impurity



Proximity measure

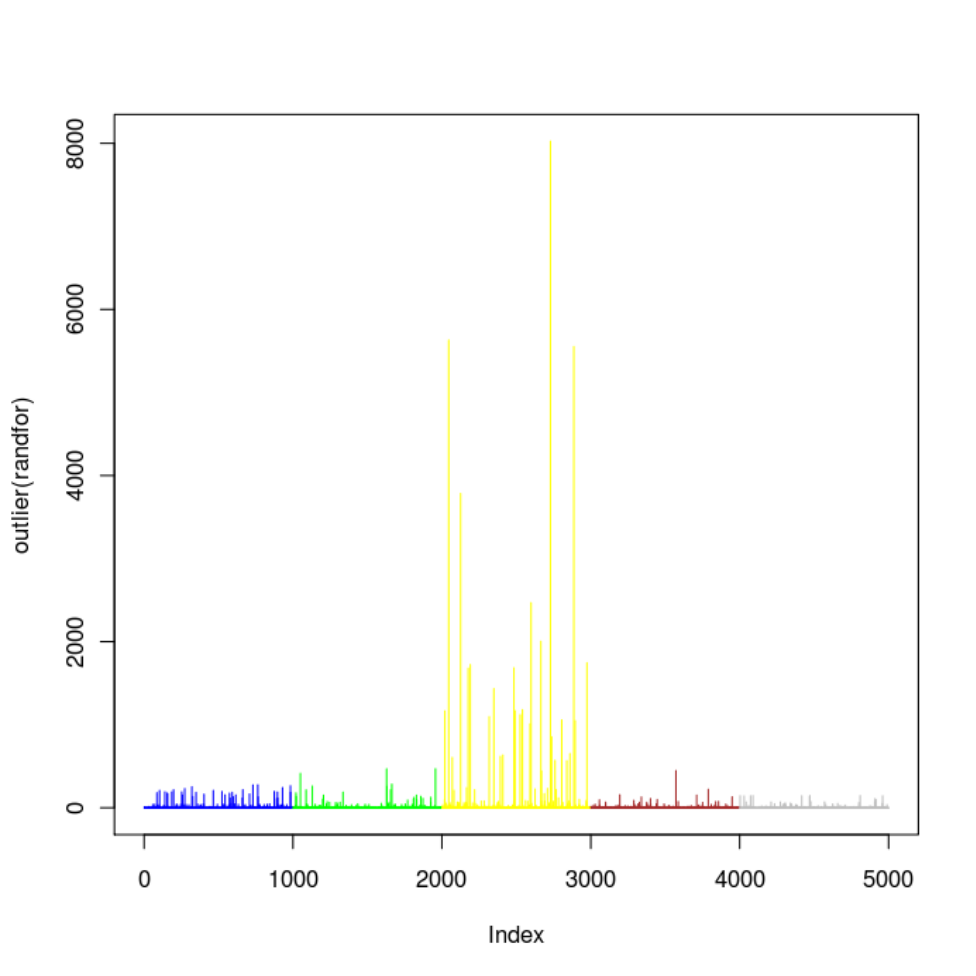
- Measures how frequent unique pairs of training samples (in and out of bag) end up in the same terminal node
- Used to fill in missing data and calculating outliers
- In the randomForest() function specify proximity=TRUE



Outliers for classification

Outlier plots

- Use outlier() function to calculate outlier measures
- Can plot using the R plot() function
- Plot shows which samples contain variables that are outliers



Outliers for classification