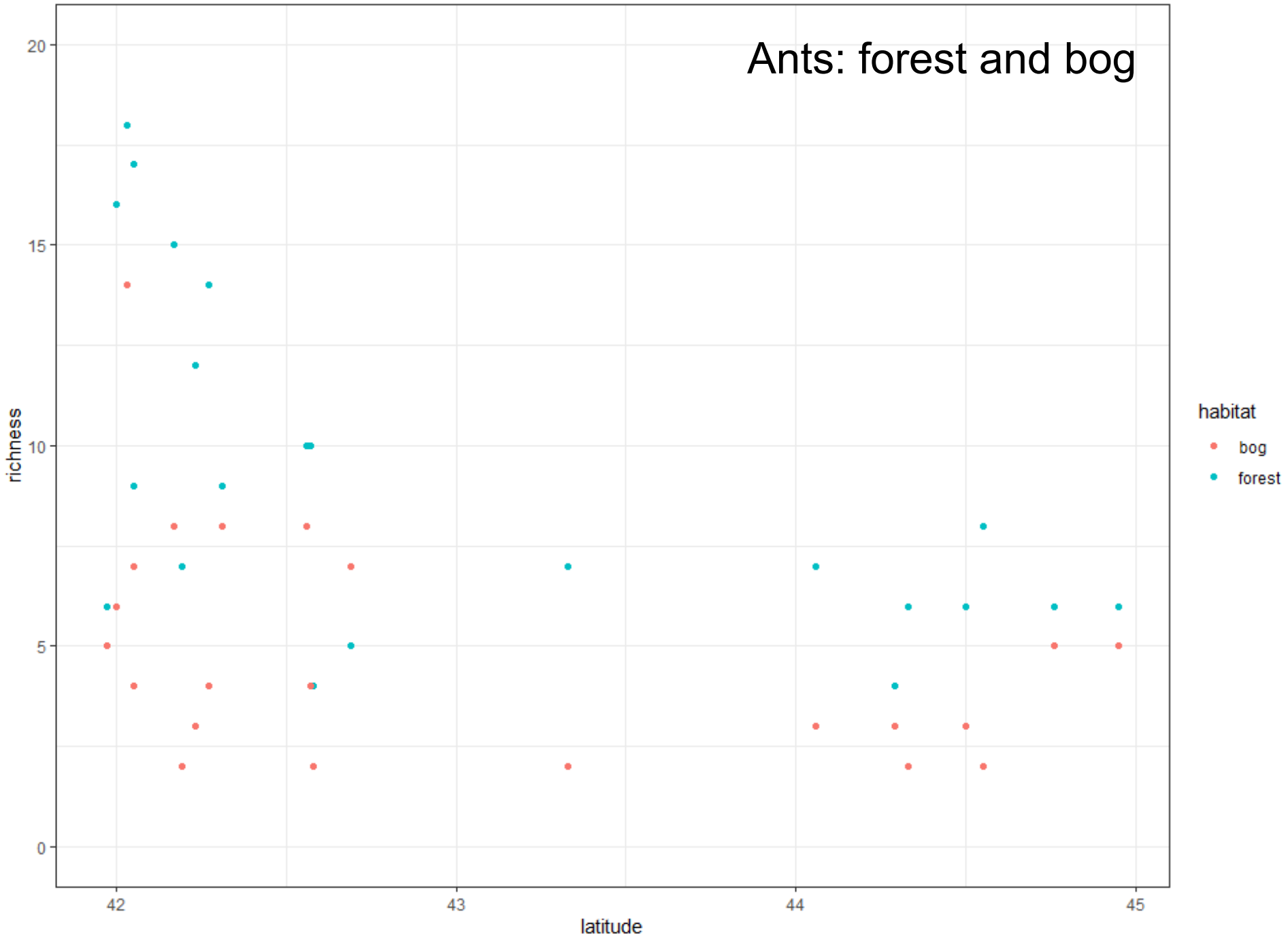
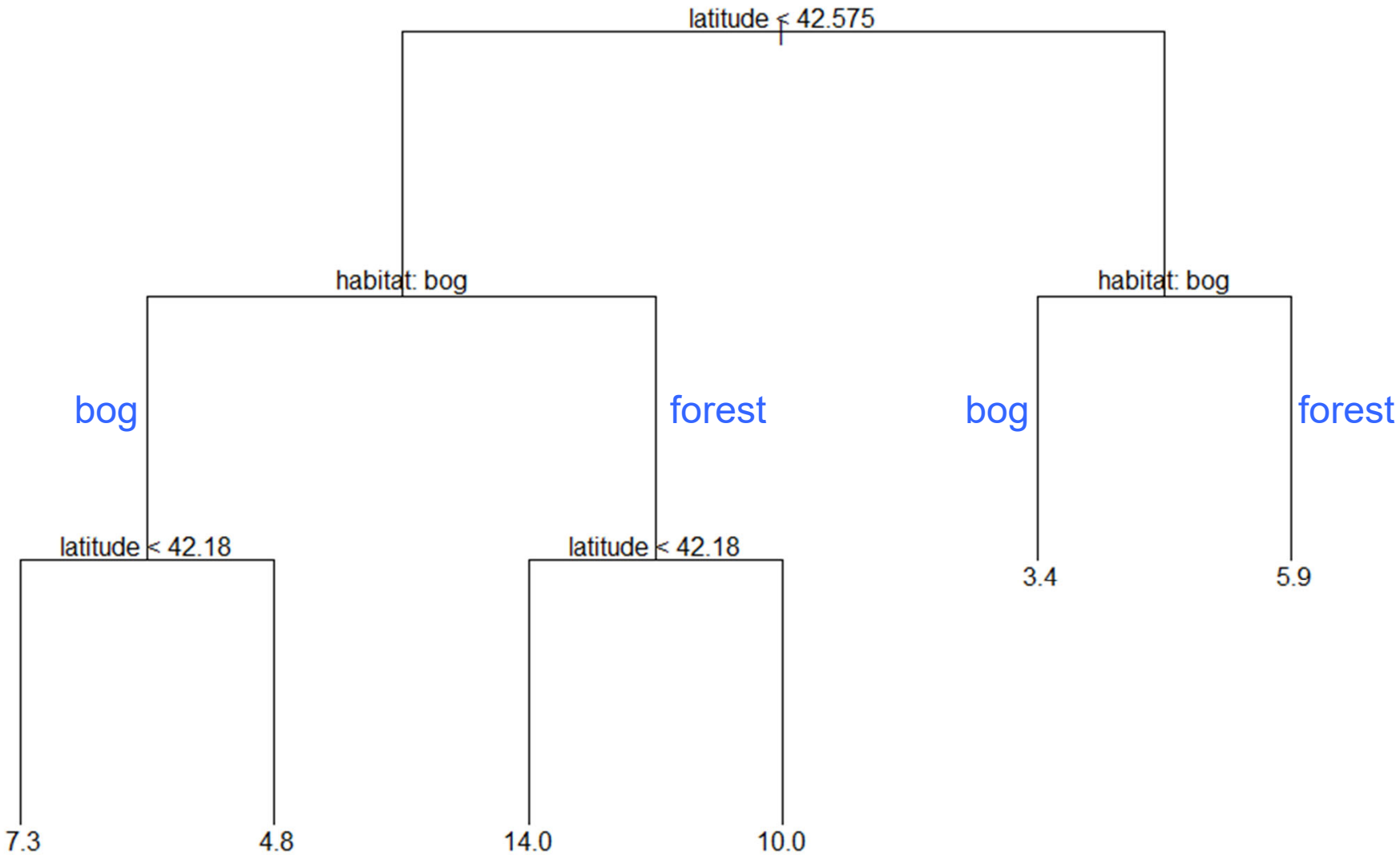
Today

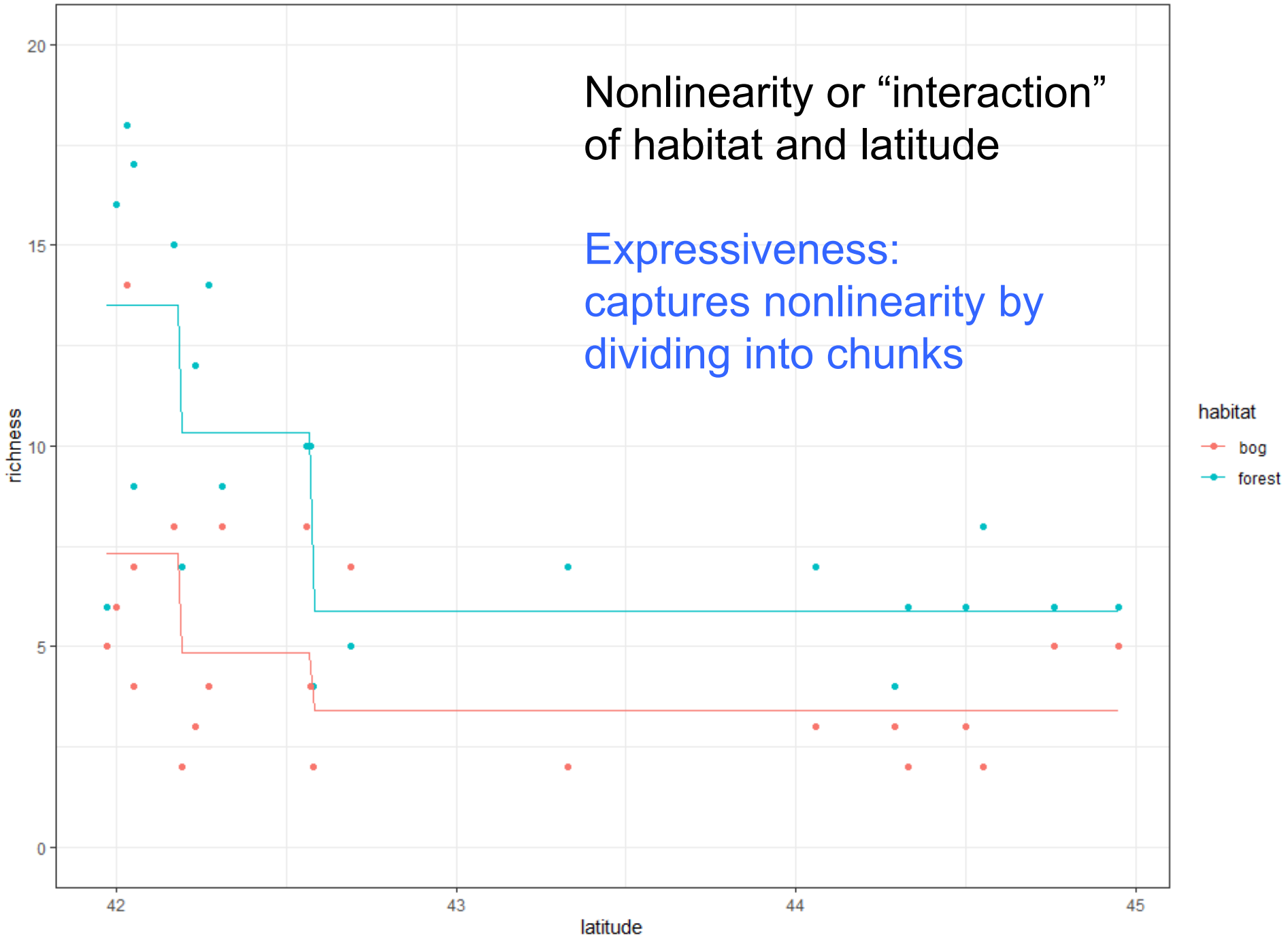
- Finish up basic trees
  - code for model algorithm
  - multiple predictor variables
  - inference algorithm
- Ensemble methods
  - bagging (bootstrap aggregation)

Code

- ants_tree.R
- model algorithm
- translate pseudocode to R

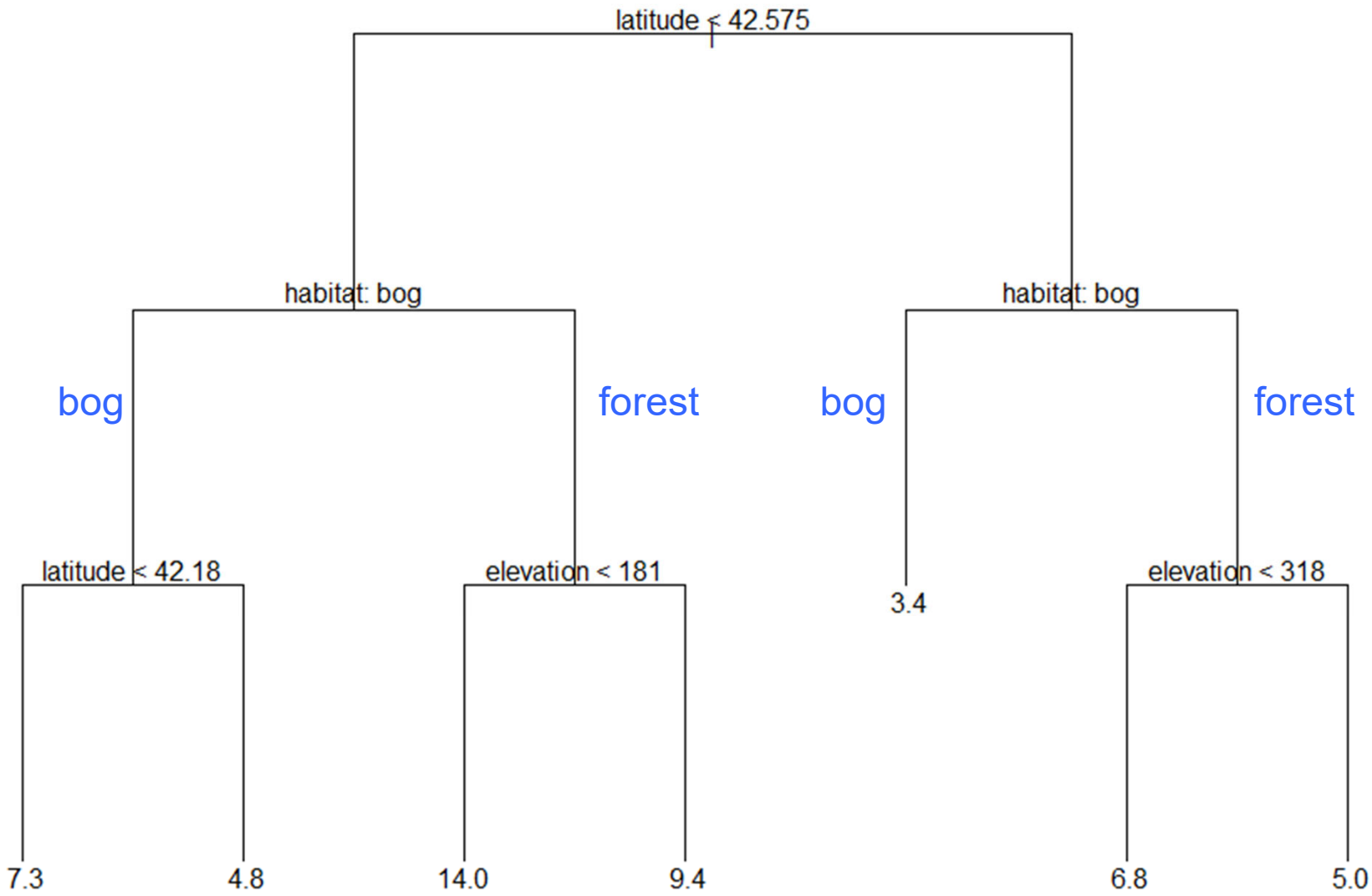

Ants: forest and bog
richness
20
15
10
5
0
42
43
44
45
latitude
habitat
bog
forest

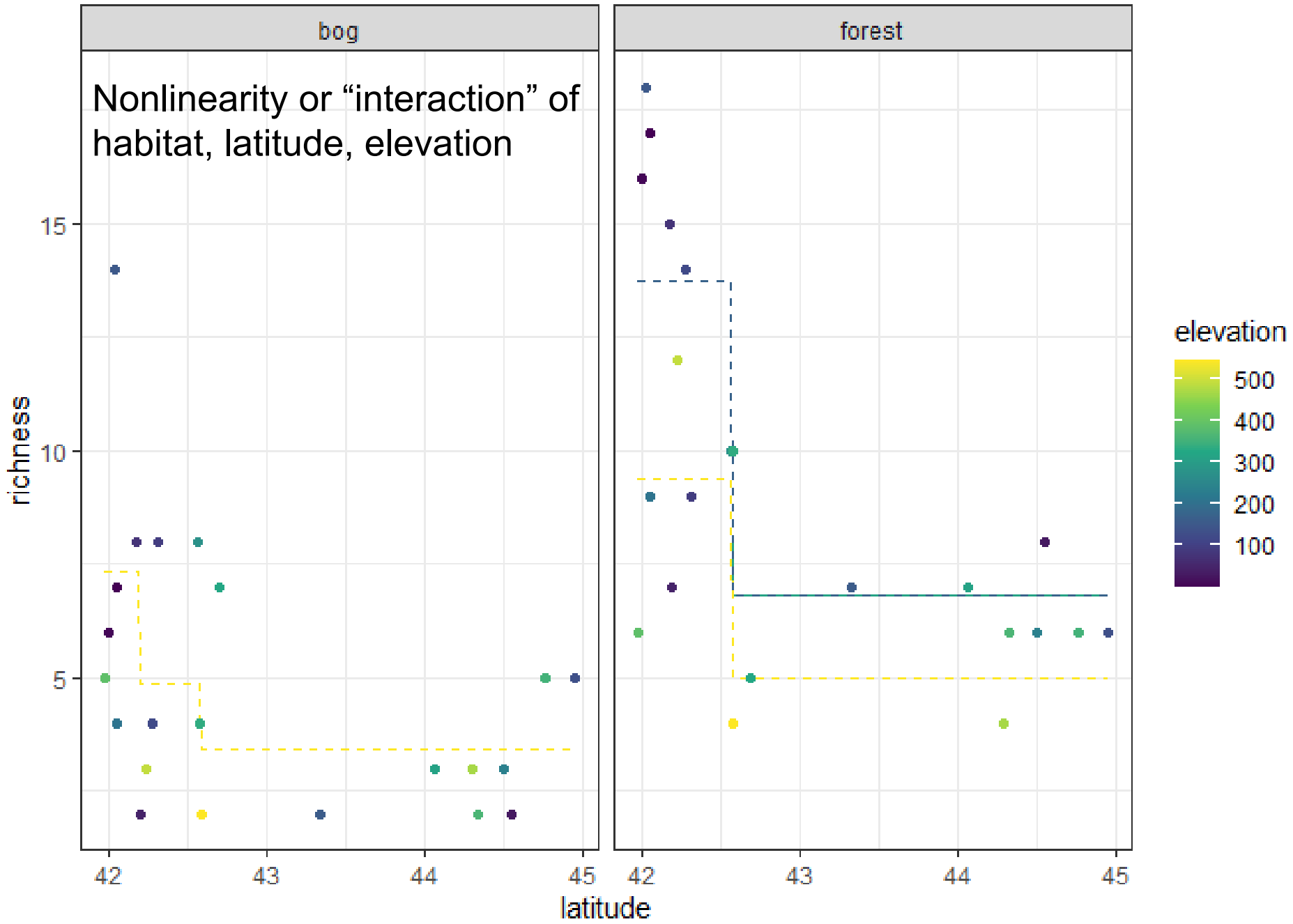

latitude < 42.575
habitat: bog
habitat: bog
bog
forest
bog
forest
latitude < 42.18
latitude < 42.18
3.4
5.9
7.3
4.8
14.0
10.0


Nonlinearity or “interaction” of habitat and latitude
Expressiveness: captures nonlinearity by dividing into chunks
richness
20
15
10
5
0
42
43
44
45
latitude
habitat
bog
forest

```
> head(ants)
  habitat latitude elevation richness
1  forest    41.97       389        6
2  forest    42.00         8       16
3  forest    42.03       152       18
4  forest    42.05         1       17
5  forest    42.05       210        9
6  forest    42.17        78       15
```

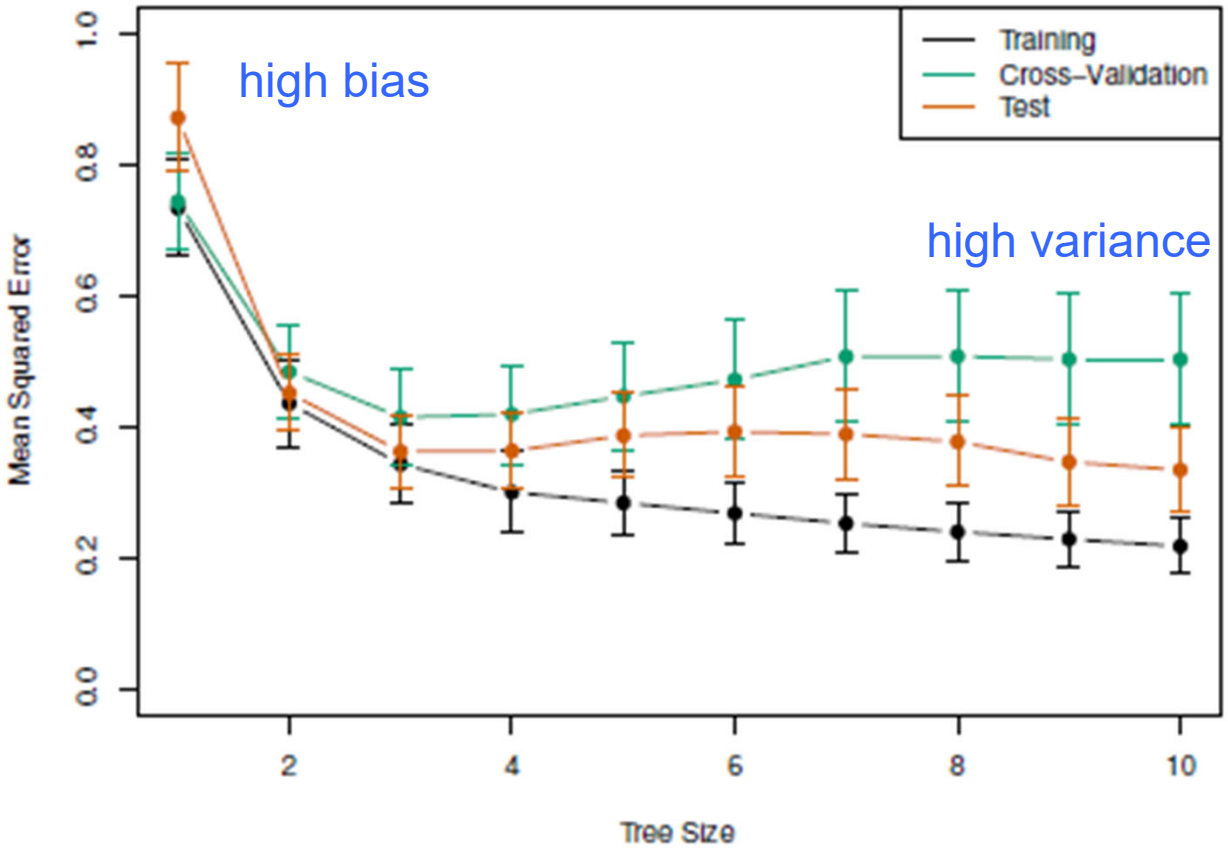
All 3 predictors




bog
forest
Nonlinearity or “interaction” of habitat, latitude, elevation
richness
15
10
5
42
43
44
45
latitude
elevation
500
400
300
200
100

Inference

- k-fold CV
- Can tune tree parameters
  - e.g. tree depth
- or tree complexity: regularization
  - training: complexity penalty
  - e.g. loss = SSQ + αT
  - where α is a tuning parameter, T is number of leaves
  - “pruning” (first fit complex tree, then prune it)


high bias
high variance
Training
Cross-Validation
Test
Mean Squared Error
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Tree Size

Ensemble methods

- Train many models (ensemble)
- Average the models to predict
- Averaging reduces prediction variance

e.g. $\mathrm{Var}(\bar{y}) = \frac{\sigma_y^2}{n}$

variance of the mean of y is
less than the variance of y

Bagging

- Bootstrap
  - form new datasets by resampling from the data
  - sample with replacement
- Aggregate
  - average over bootstrapped model fits

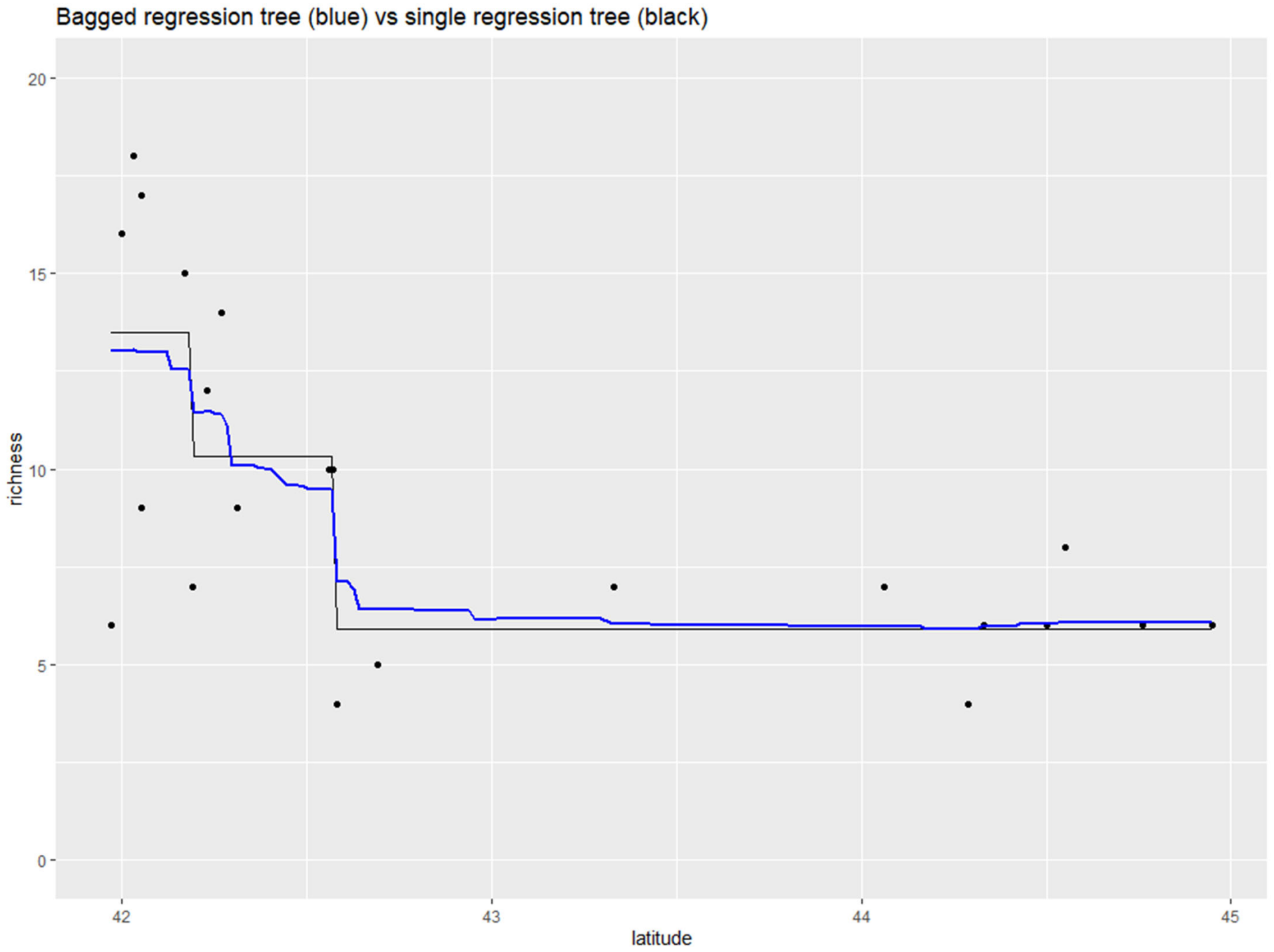
Bagging algorithm

for many repetitions
    resample the data with replacement
    train the base model
    record prediction
final prediction = mean of predictions

Base model: can be any type of model

Bagged regression tree

```
# Bagging algorithm
boot_reps <- 500
n <- nrow(forest_ants)
nx <- nrow(grid_data)
boot_preds <- matrix(rep(NA, nx*boot_reps), nrow=nx, ncol=boot_reps)
# for many repetitions
for ( i in 1:boot_reps ) {
#    resample the data (rows) with replacement
     boot_indices <- sample(1:n, n, replace=TRUE)
     boot_data <- forest_ants[boot_indices,]
#    train the base model
     boot_train <- tree(richness ~ latitude, data=boot_data)
#    record prediction
     boot_preds[,i] <- predict(boot_train, newdata=grid_data)
}
bagged_preds <- rowMeans(boot_preds)
```


Bagged regression tree (blue) vs single regression tree (black)
richness
20
15
10
5
0
42
43
44
45
latitude