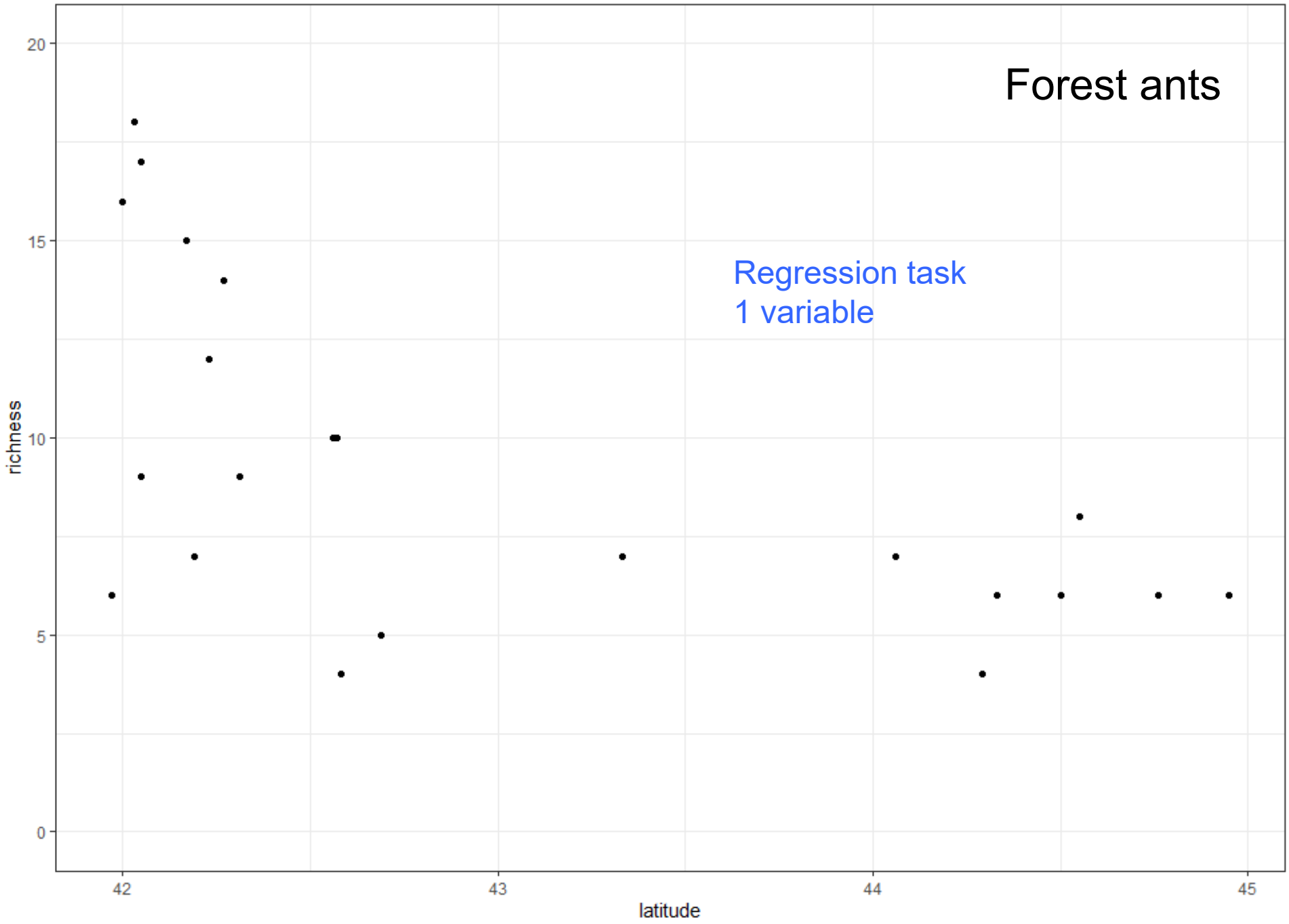
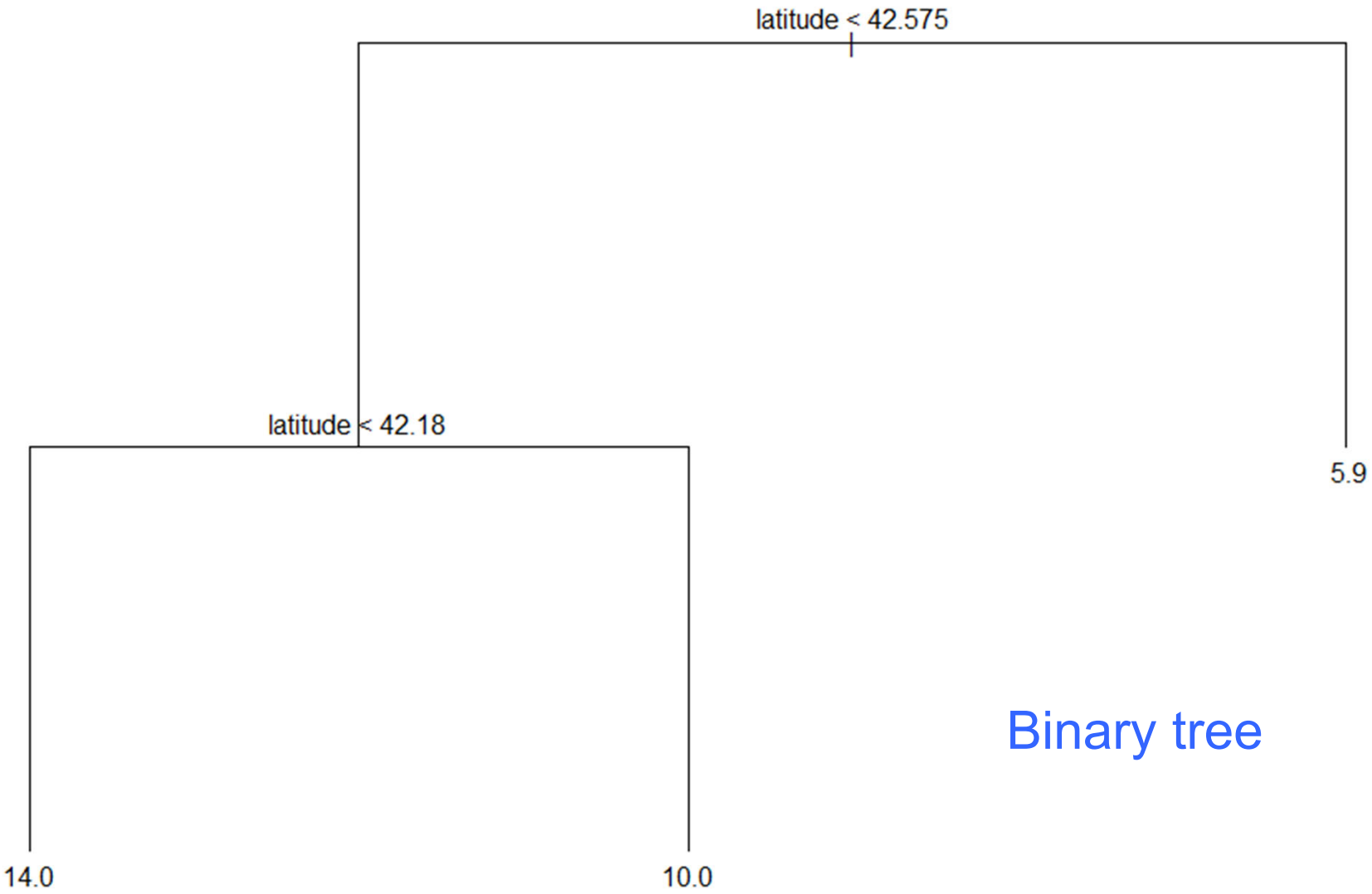
Today

- Decision tree models
- + training and inference
- Coming up:
  - building to xgboost
  - collection of algorithms
  - idea is to understand the building block algorithms

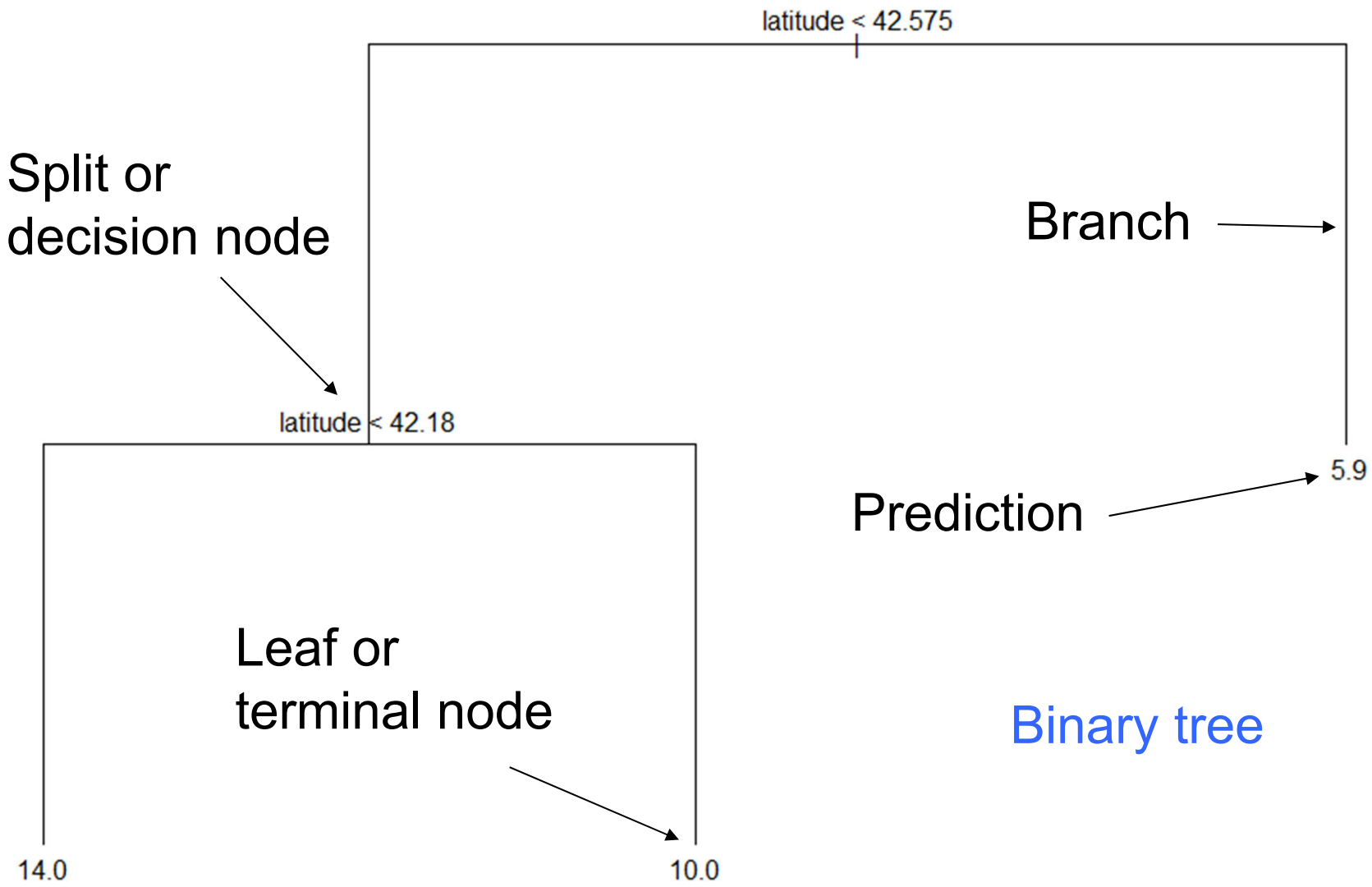

Forest ants
Regression task
1 variable
richness
latitude
20
15
10
5
0
42
43
44
45

A regression tree model

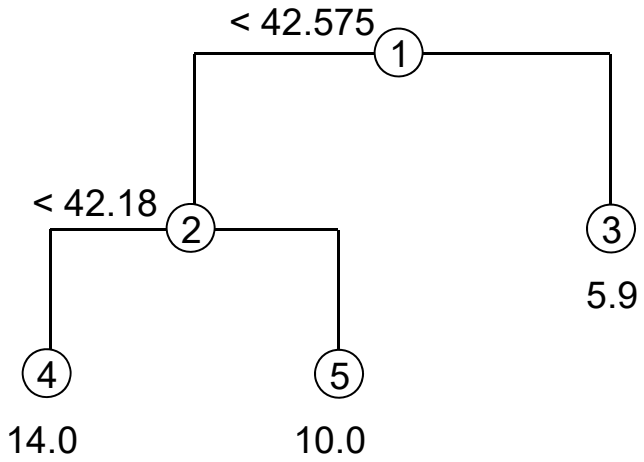


Binary tree

A regression tree model



Model algorithm

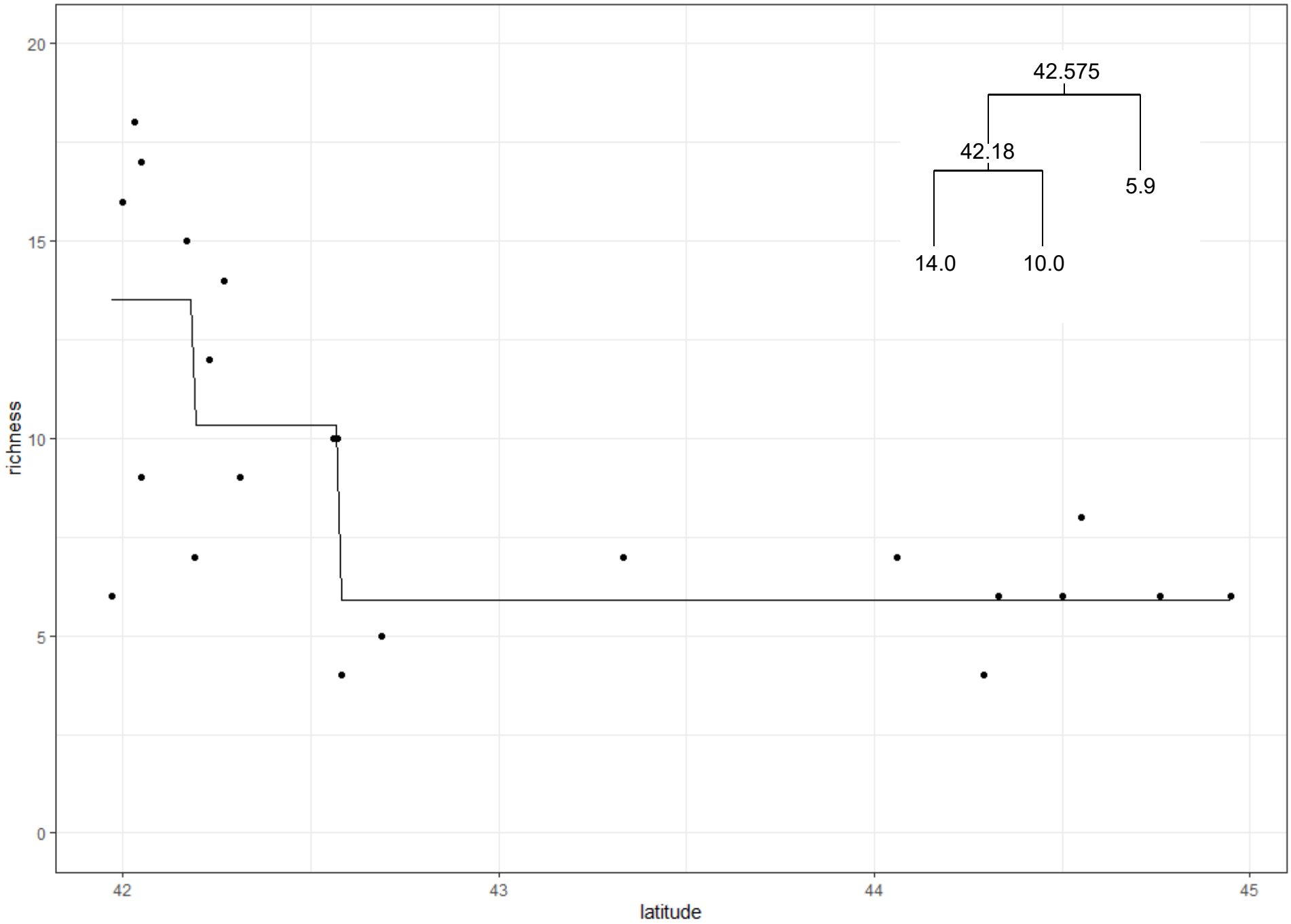


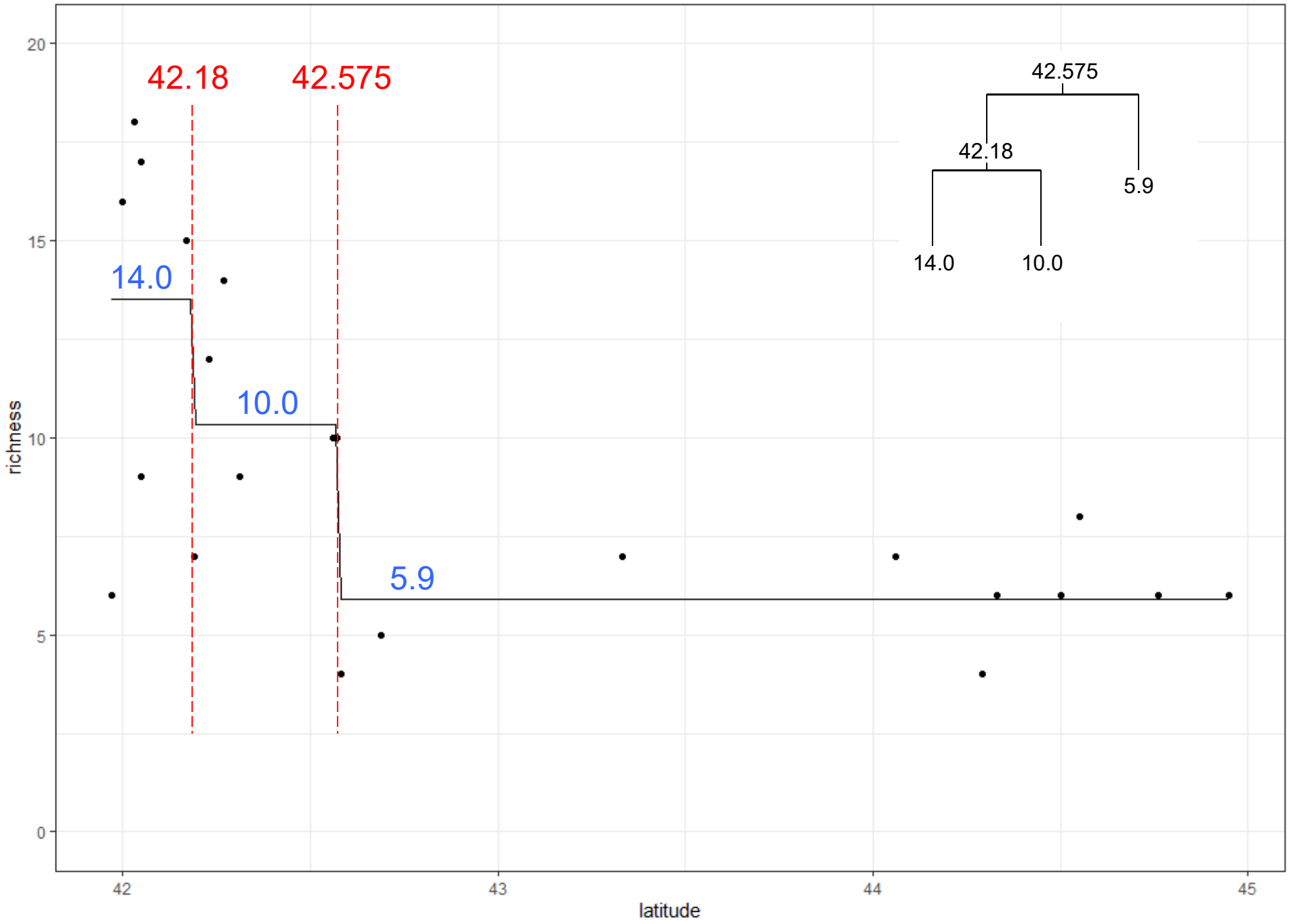
Tree structure

node	type	split	y
1	split	42.575	
2	split	42.180	
3	leaf		5.9
4	leaf		10.0
5	leaf		14.0

Algorithm

```
start at the root node
while node type is split
    if x < split value at the node
        take left branch to next node
    else
        take right branch to next node
return predicted y at node
```


20
15
10
5
0
richness
42
43
44
45
latitude
42.575
42.18
5.9
14.0
10.0


42.18
42.575
14.0
10.0
5.9
richness
latitude
0
5
10
15
20
42
43
44
45
42.575
42.18
5.9
14.0
10.0

Training algorithm

Binary recursive partitioning

```
define build_tree(y, x)
    if stop = TRUE
        calculate prediction (mean of y)
    else
        find x_split    #best x to split the data
        build_tree( (y, x)[x   < x_split] )   #L branch
        build_tree( (y, x)[x >= x_split] )   #R branch
```

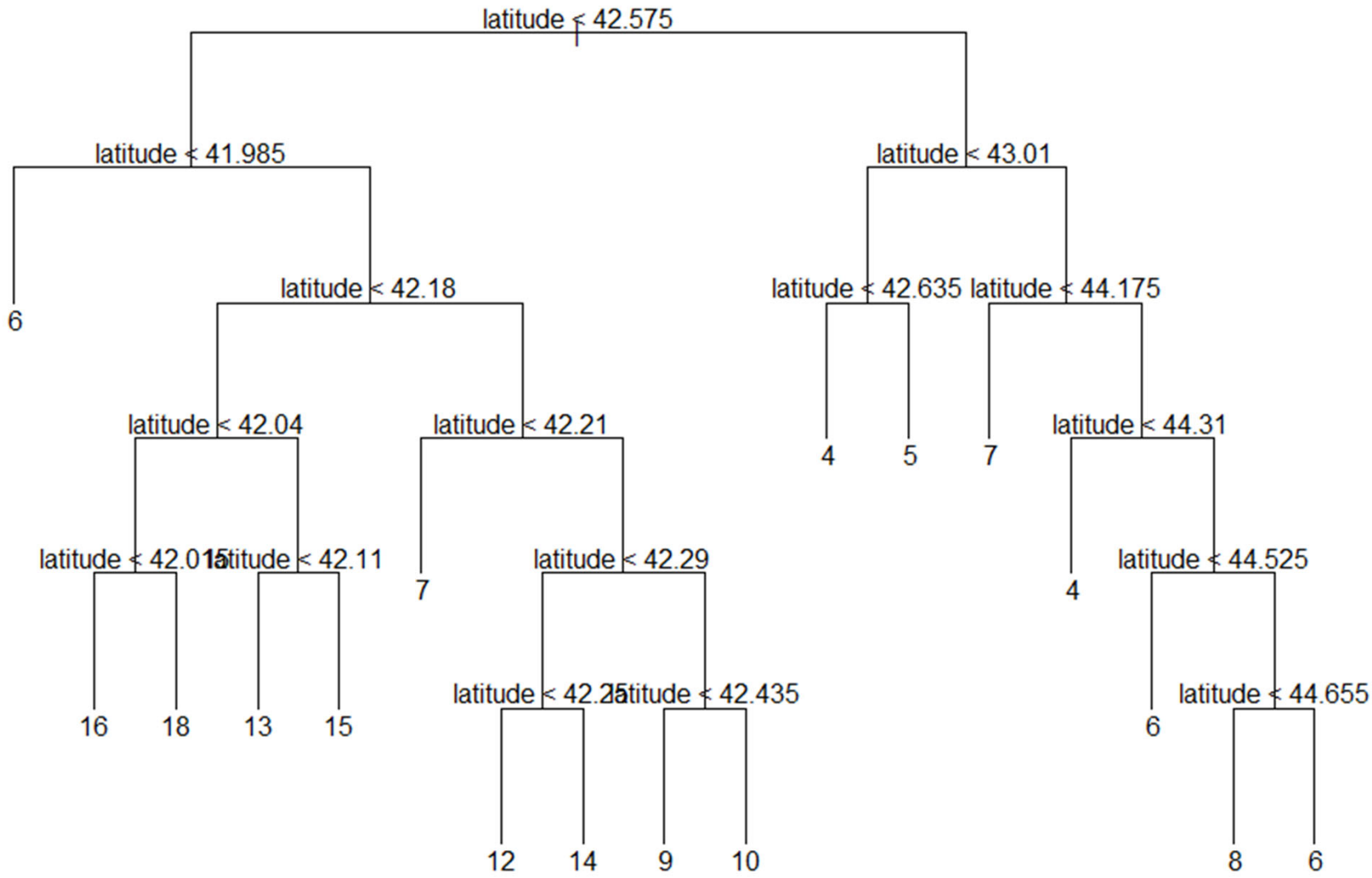
Stopping rules e.g.

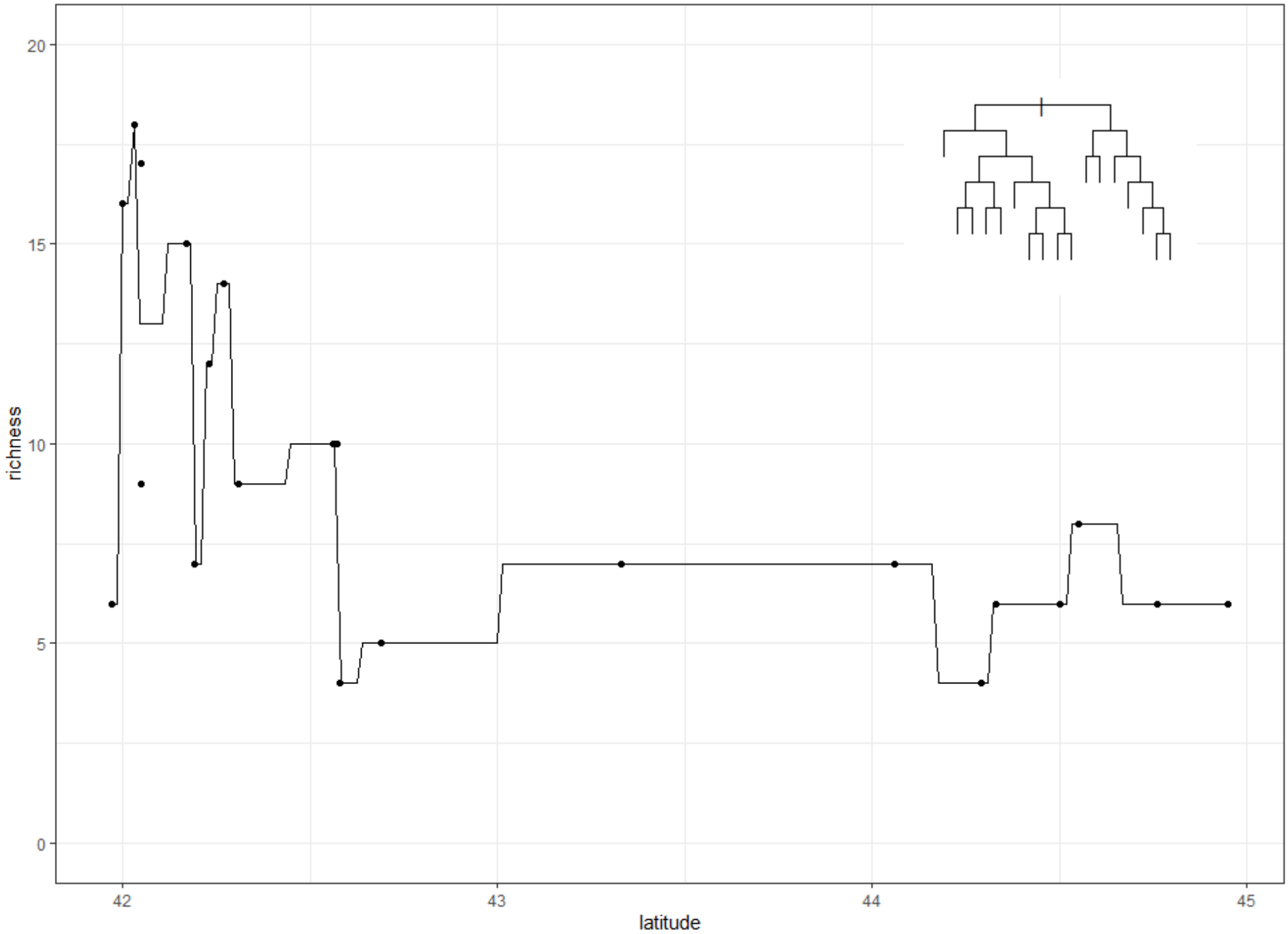
- data per node
- tree depth
- node variance
- error improvement

Find splits that minimize training error

- regression: SSQ
- classification: Gini index or entropy

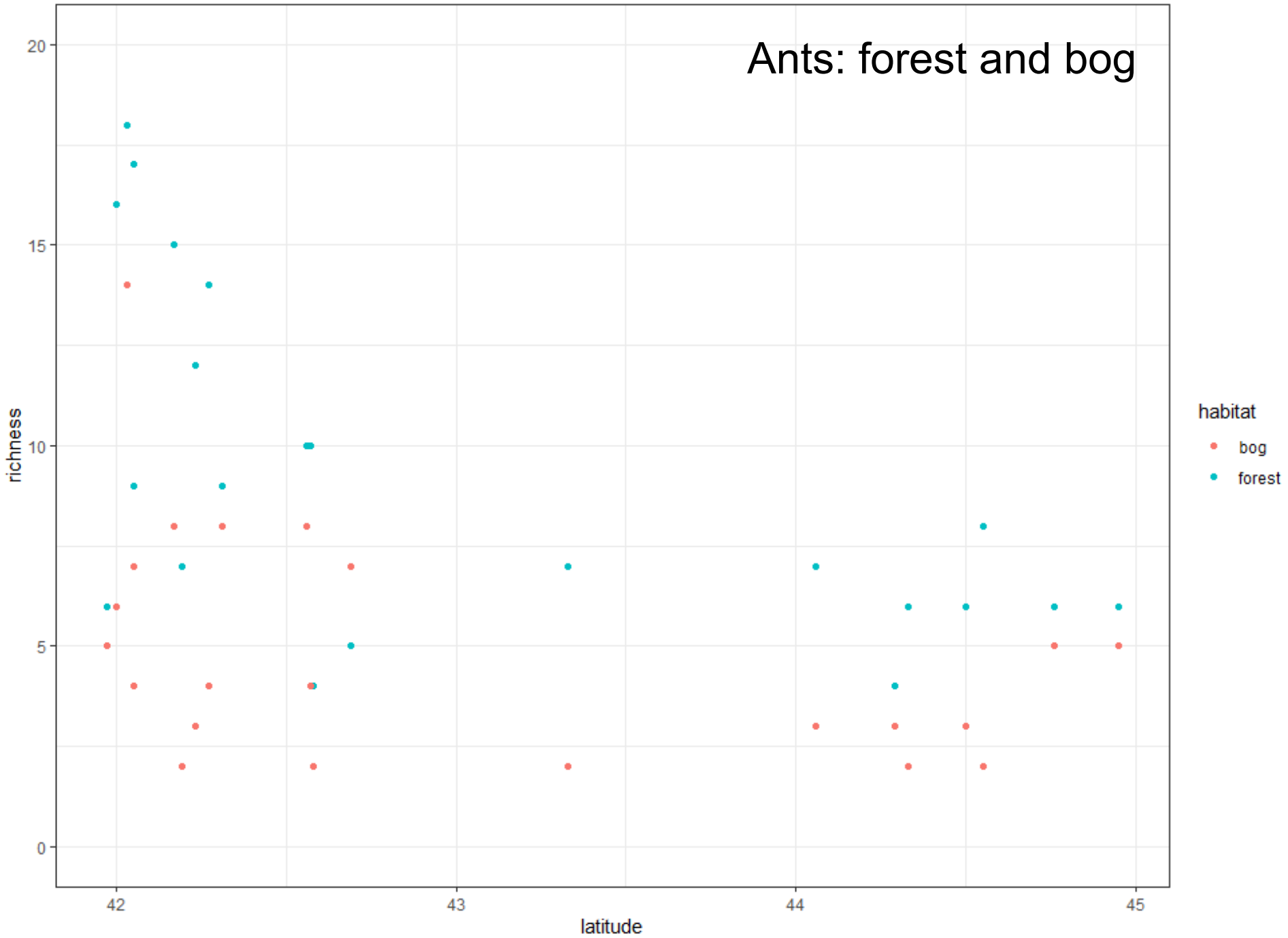
Same data, deeper tree

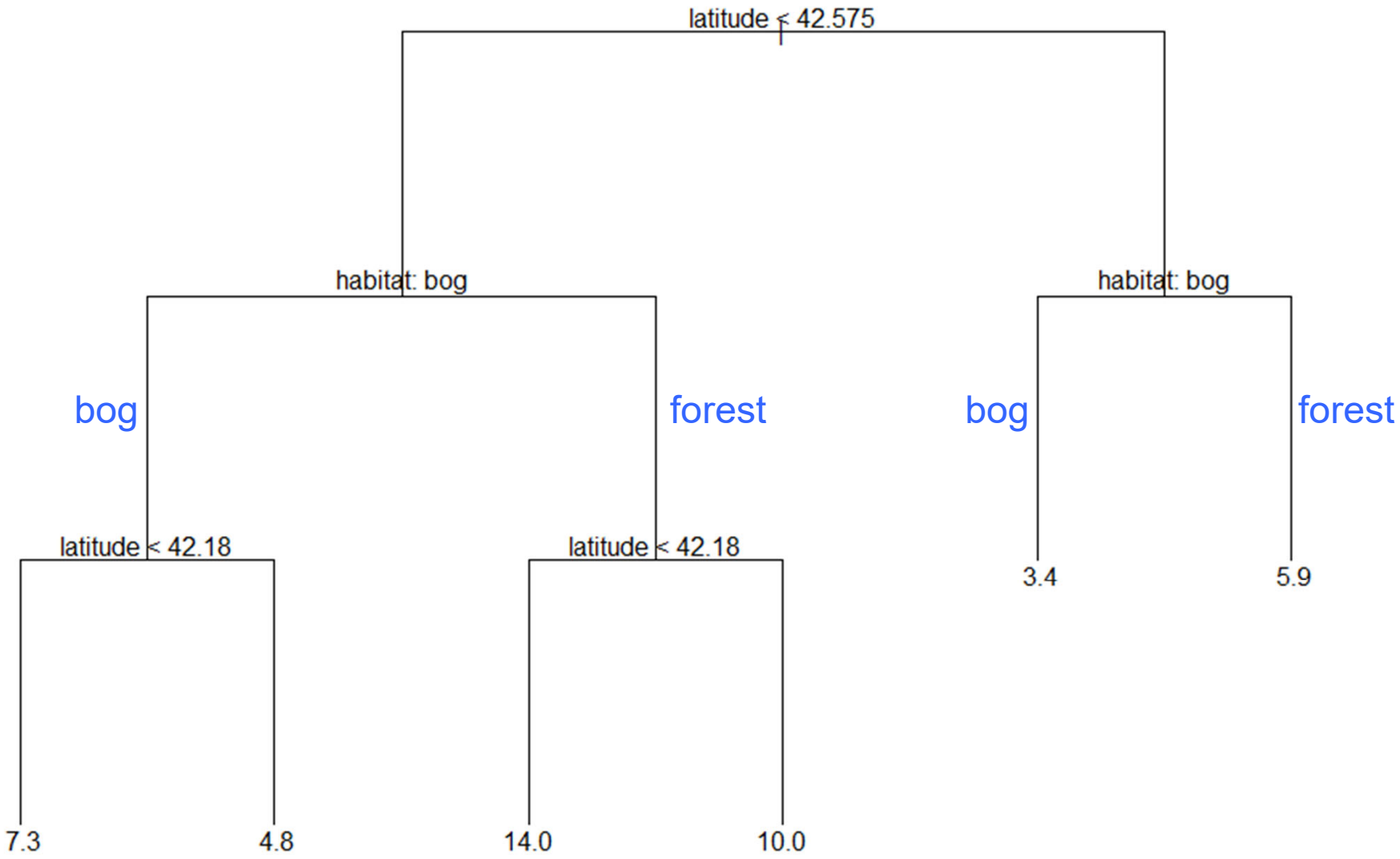


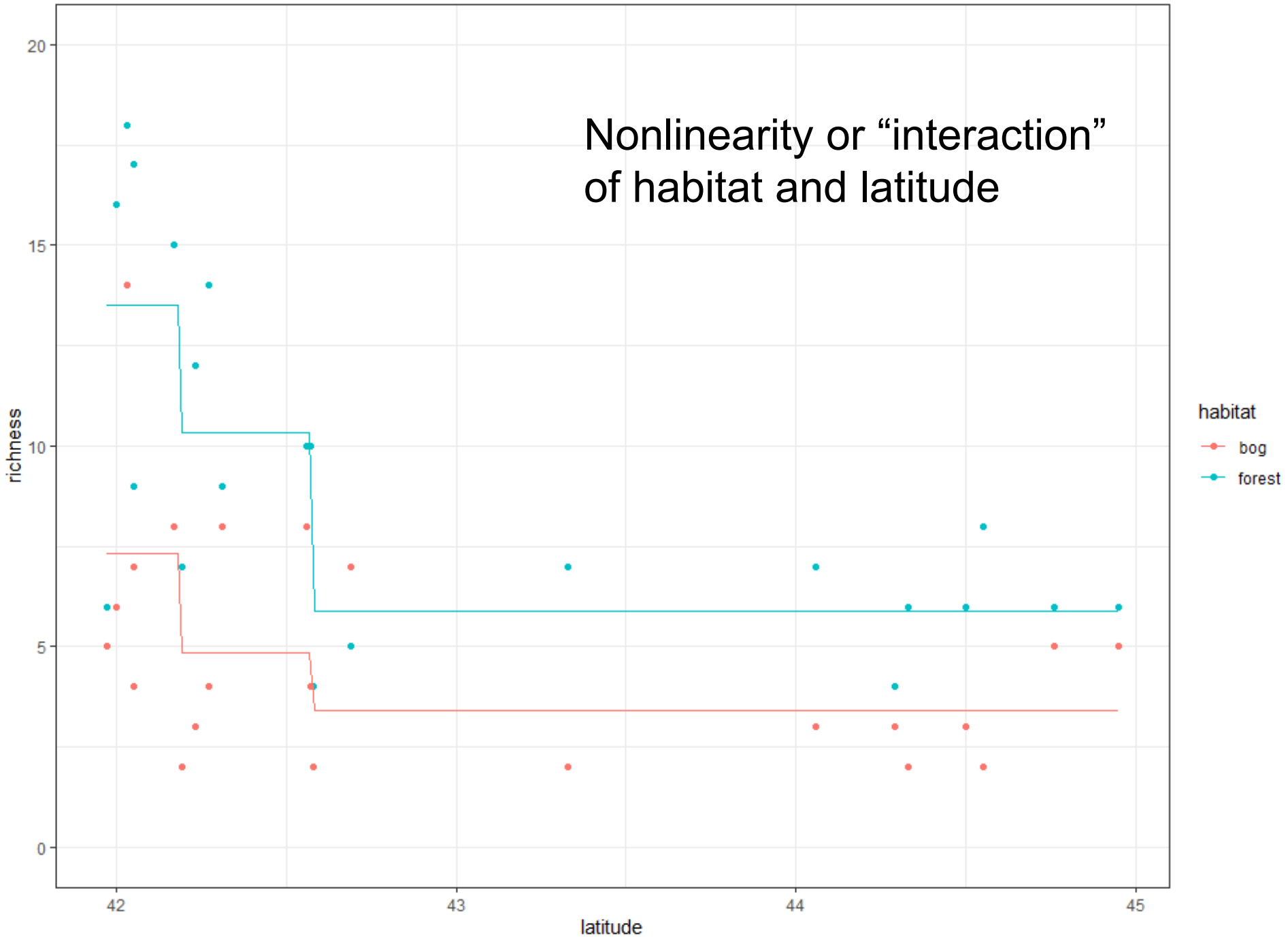

20
15
10
5
0
richness
42
43
44
45
latitude

Code

- ants_tree.R
- translate algorithm pseudocode to R

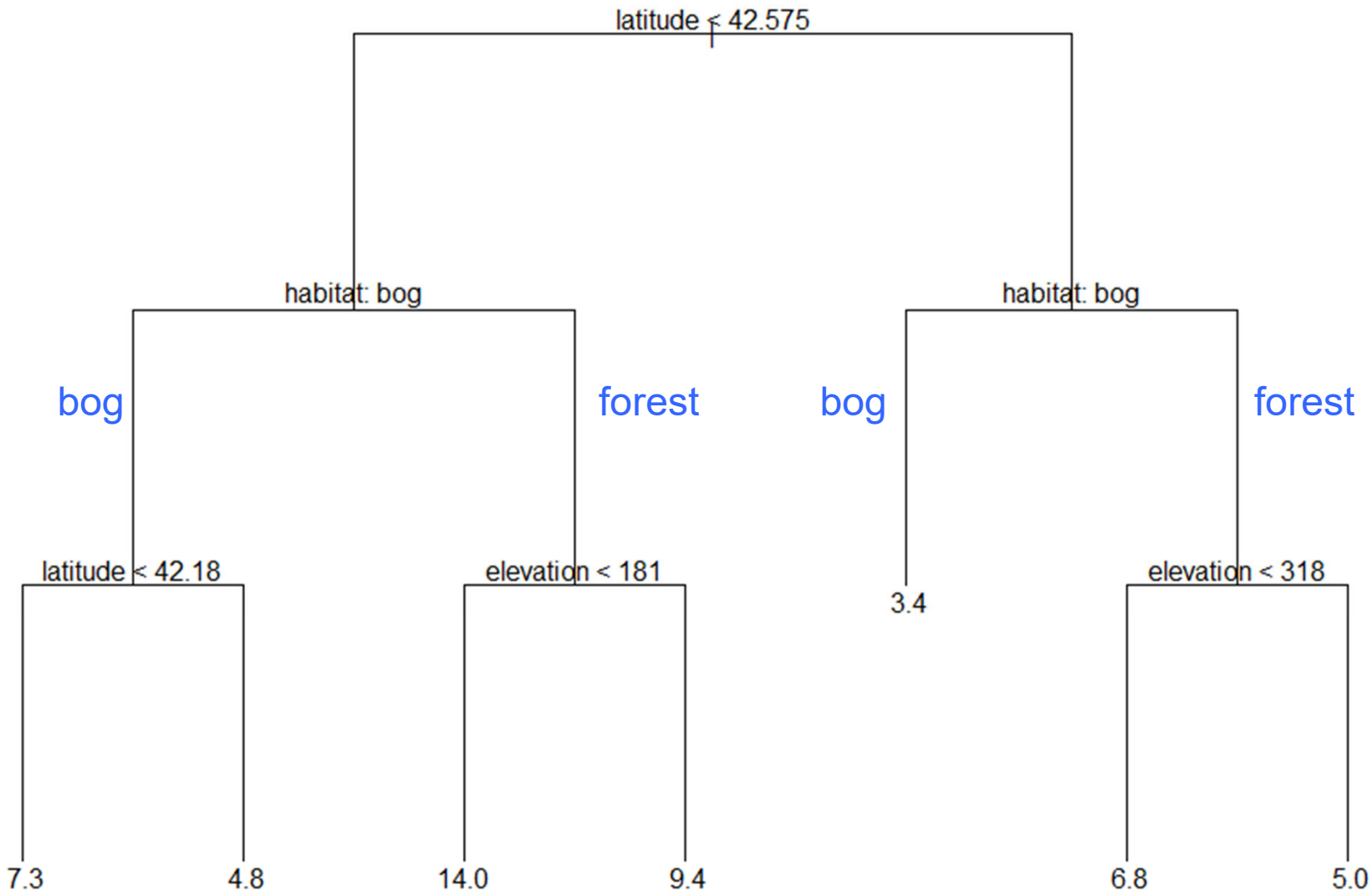

Ants: forest and bog
richness
20
15
10
5
0
42
43
44
45
latitude
habitat
bog
forest

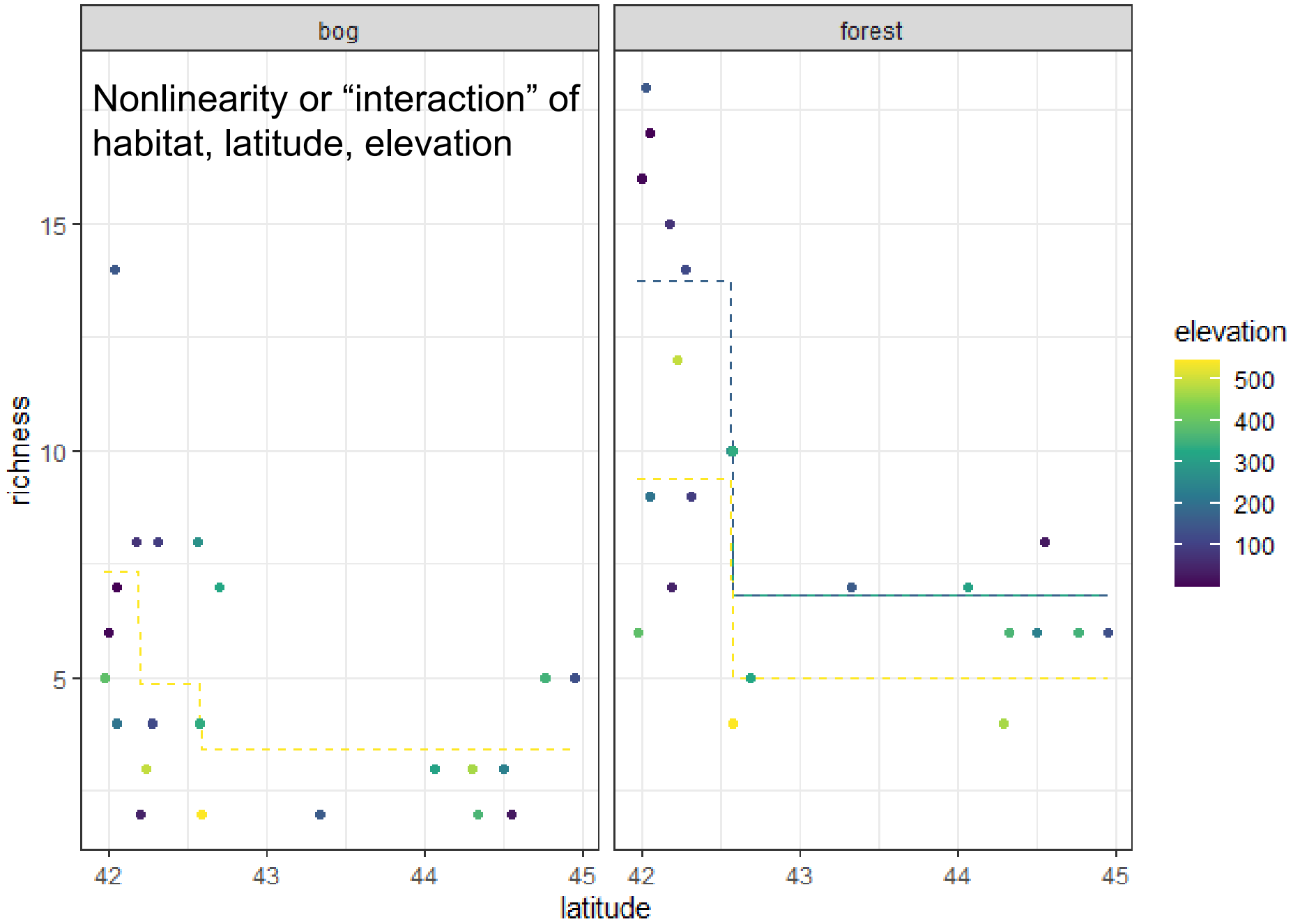

latitude < 42.575
habitat: bog
habitat: bog
bog
forest
bog
forest
latitude < 42.18
latitude < 42.18
3.4
5.9
7.3
4.8
14.0
10.0


Nonlinearity or “interaction”
of habitat and latitude
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
```

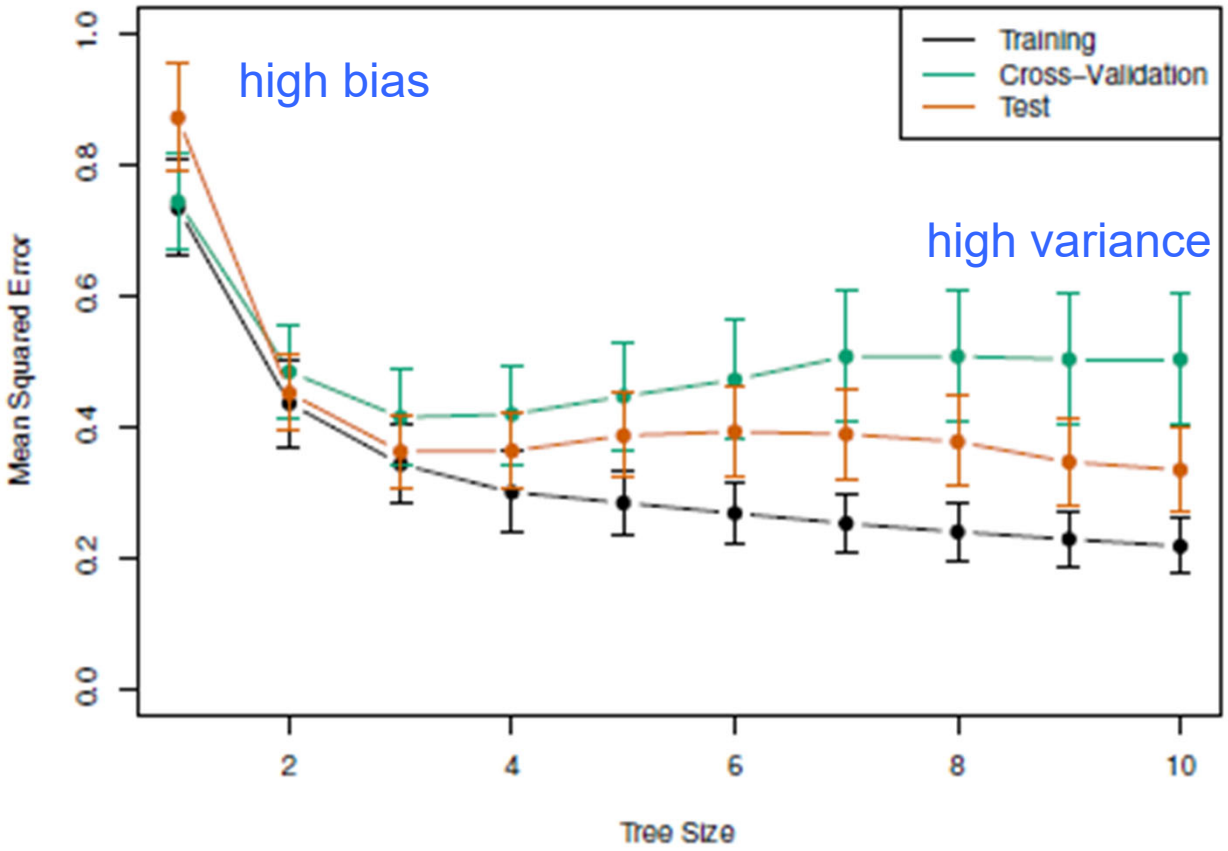
All 3 predictors




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 also use for tree complexity
  - training: complexity penalty
  - e.g. loss = SSQ + αT
  - where α is a tuning parameter, T is number of leaves
  - “pruning”


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