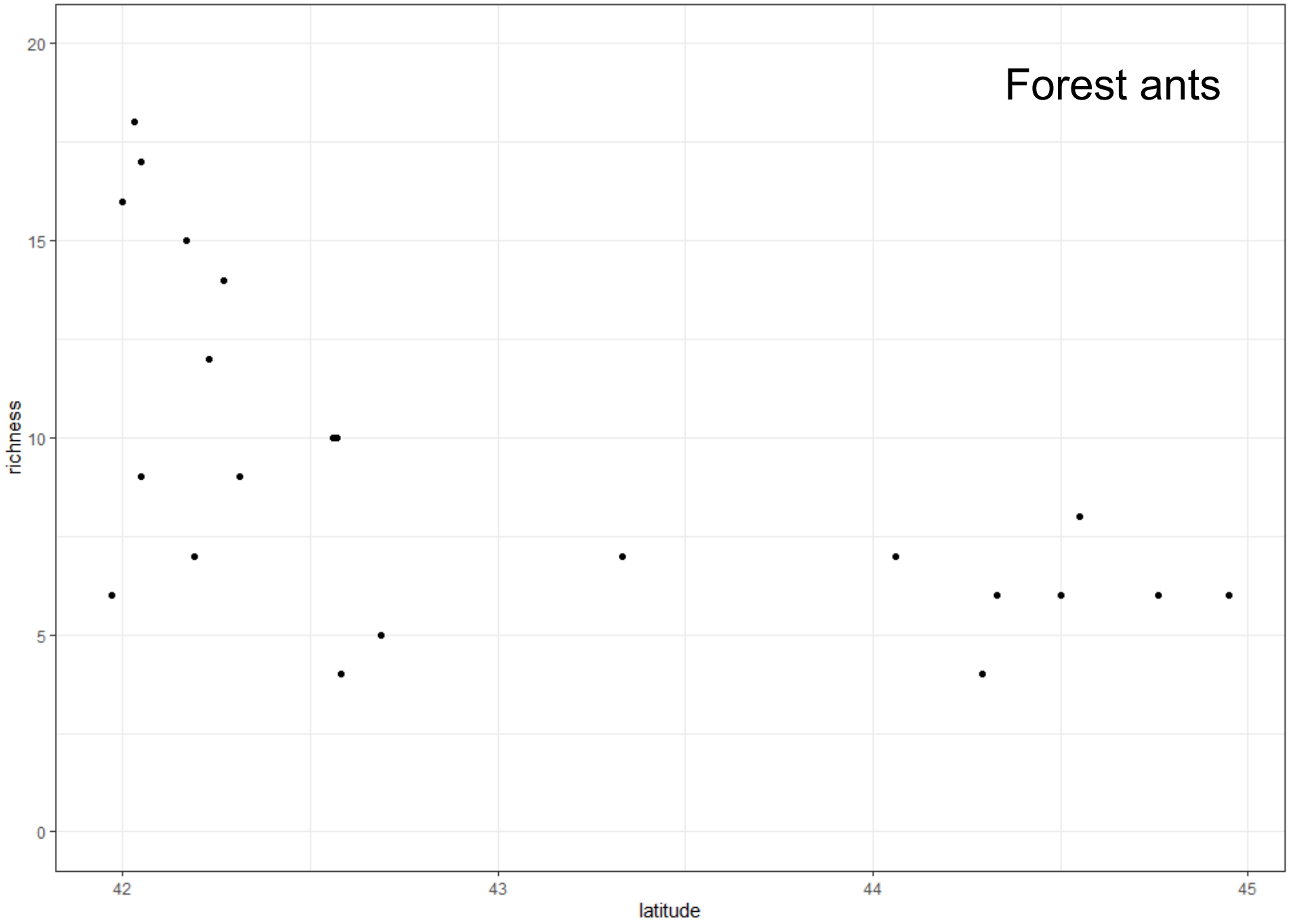
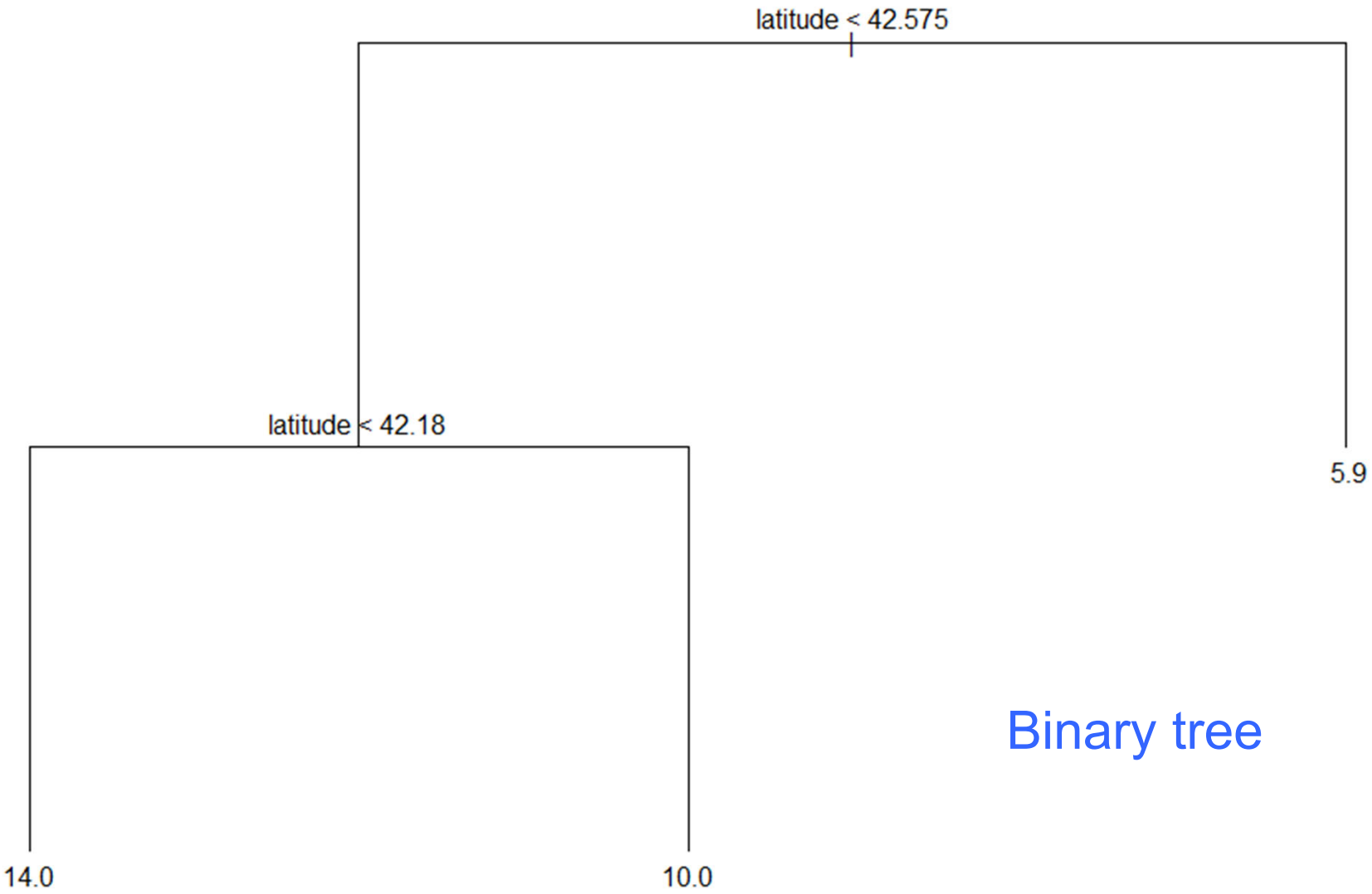
Today

- Decision tree models
- + training and inference

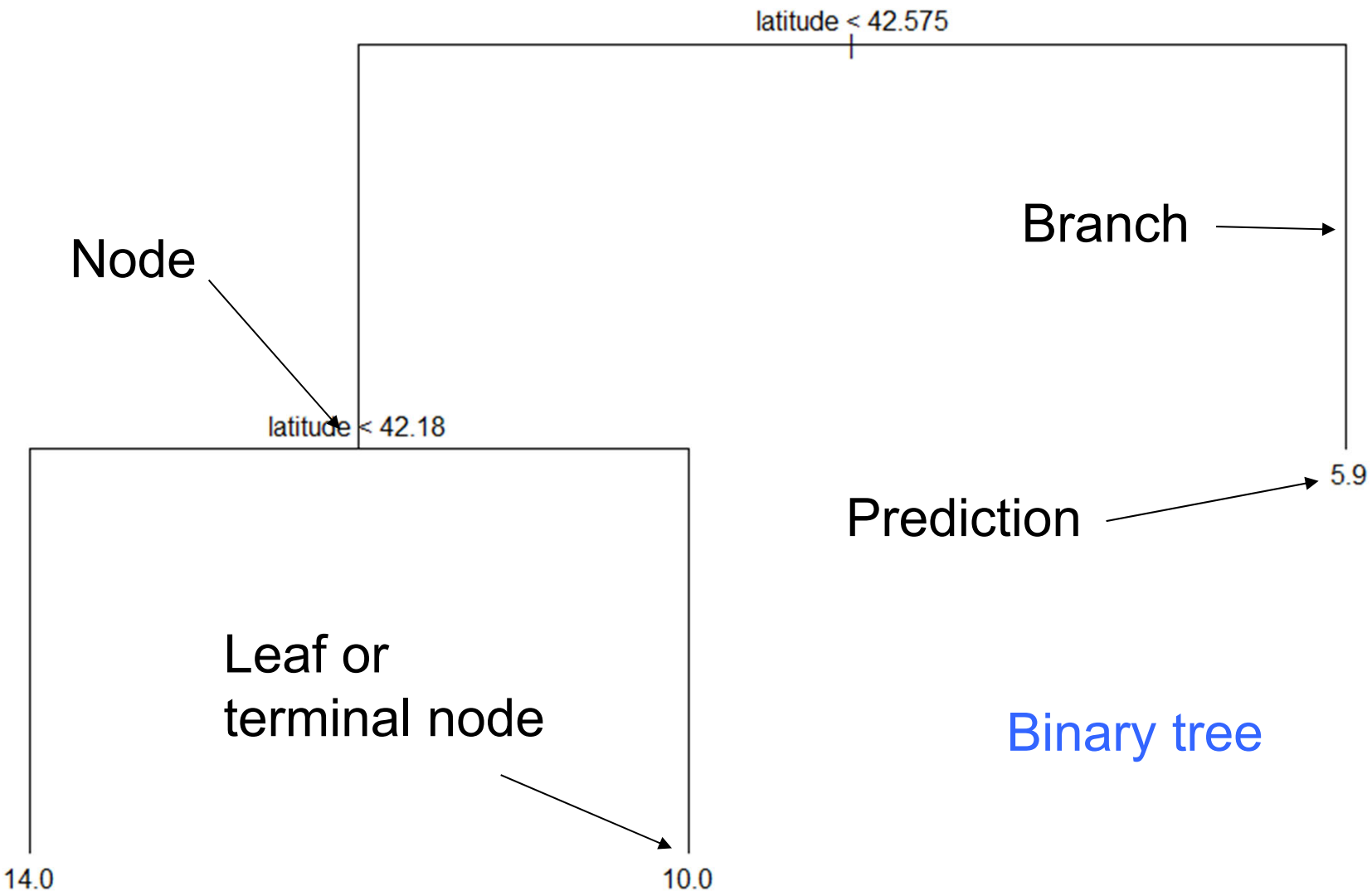

Forest ants
richness
20
15
10
5
0
42
43
44
45
latitude

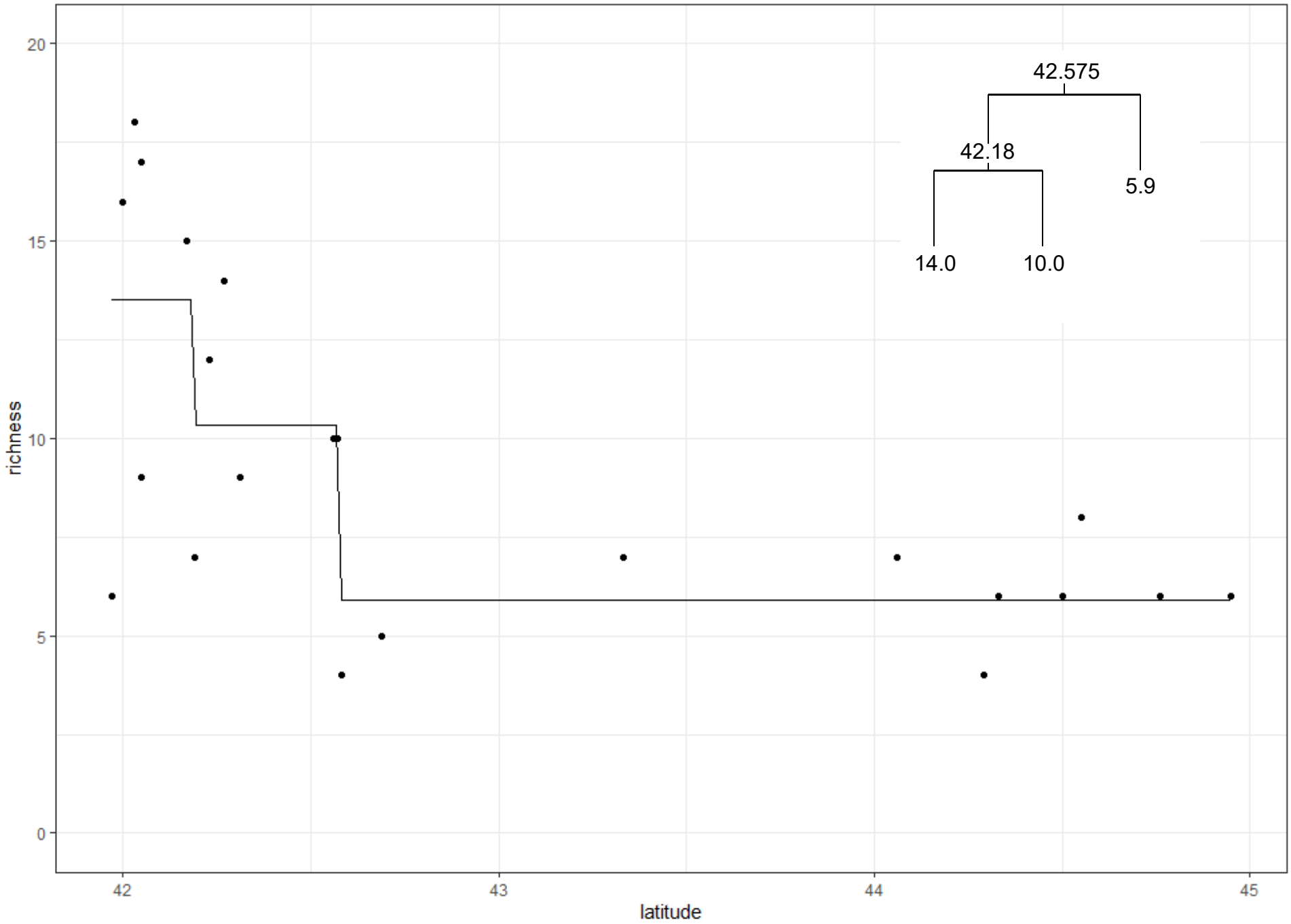
A regression tree model

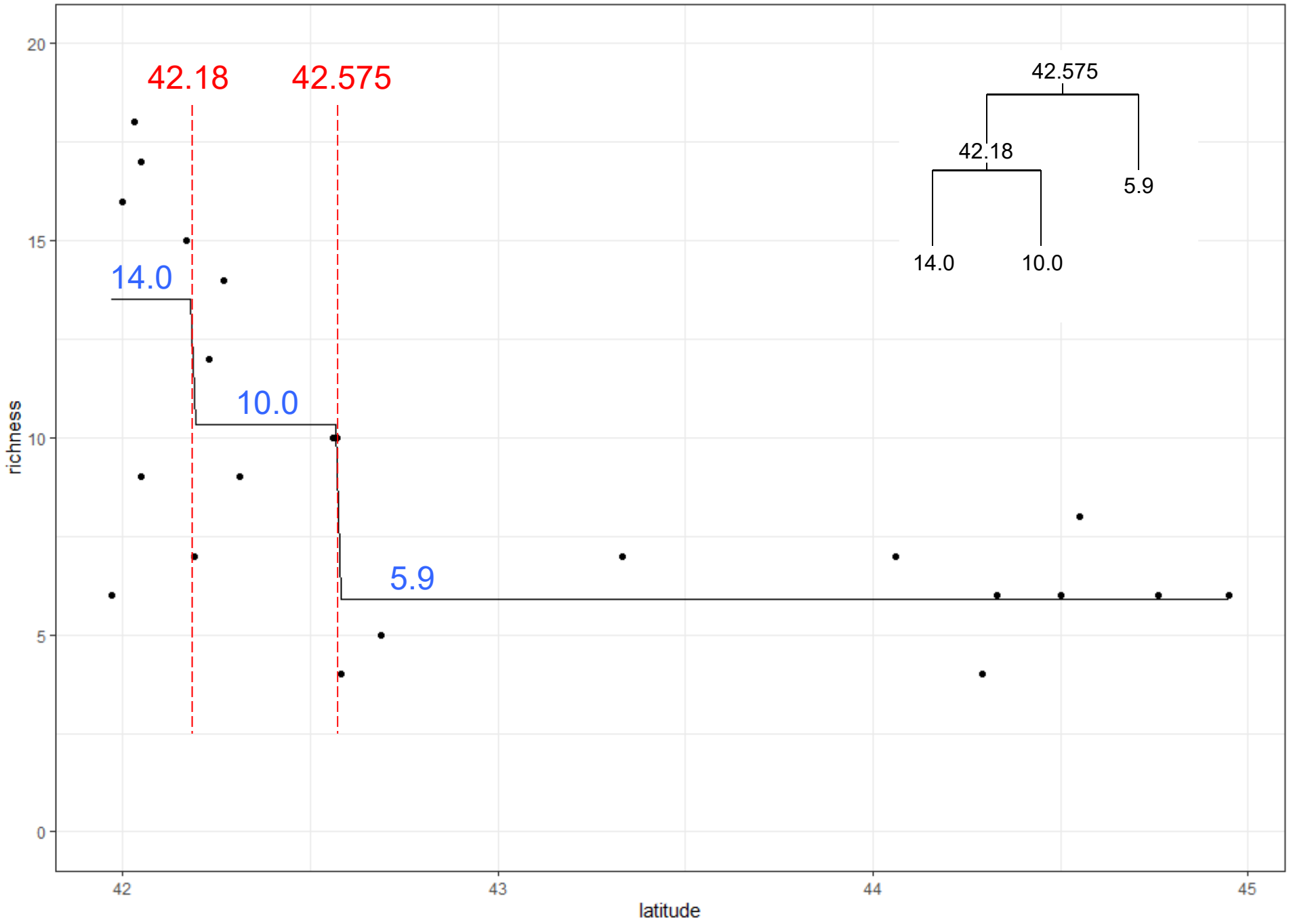


Binary tree

A regression tree model



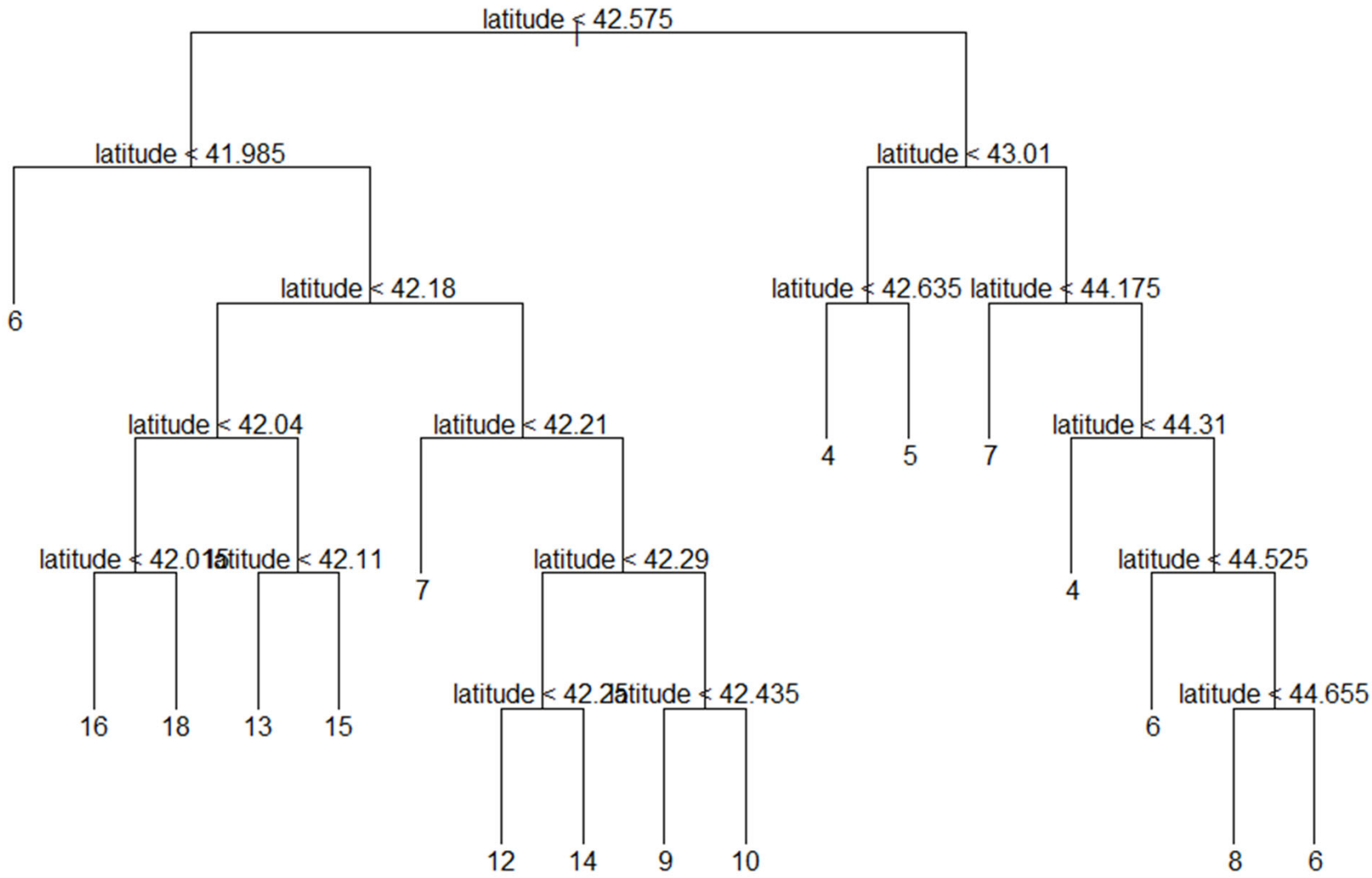

20
15
10
5
0
richness
42
43
44
45
latitude
42.575
42.18
5.9
14.0
10.0

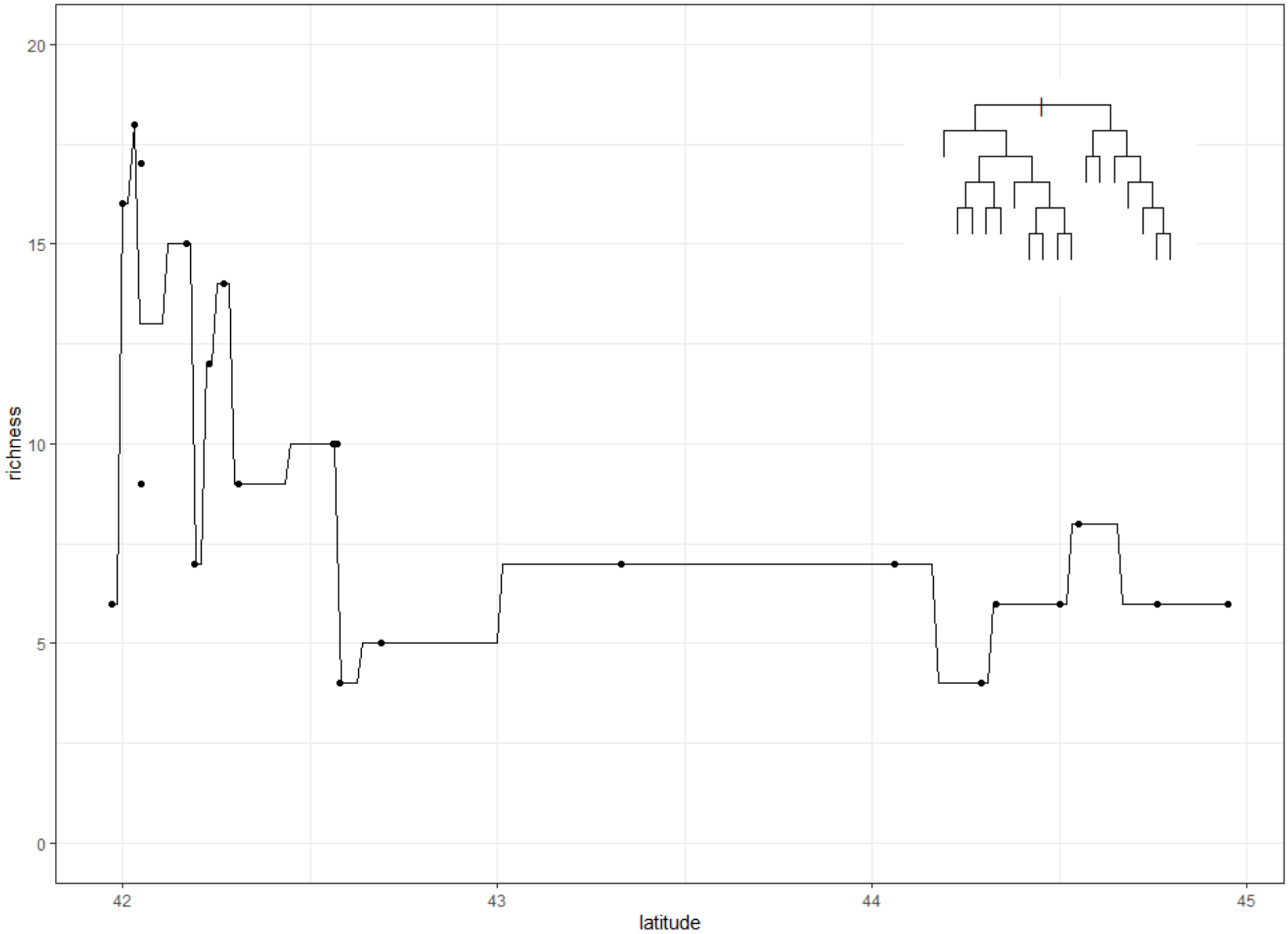

42.18
42.575
14.0
10.0
5.9
richness
latitude
0
5
10
15
20
42
43
44
45
42.575
42.18
5.9
14.0
10.0

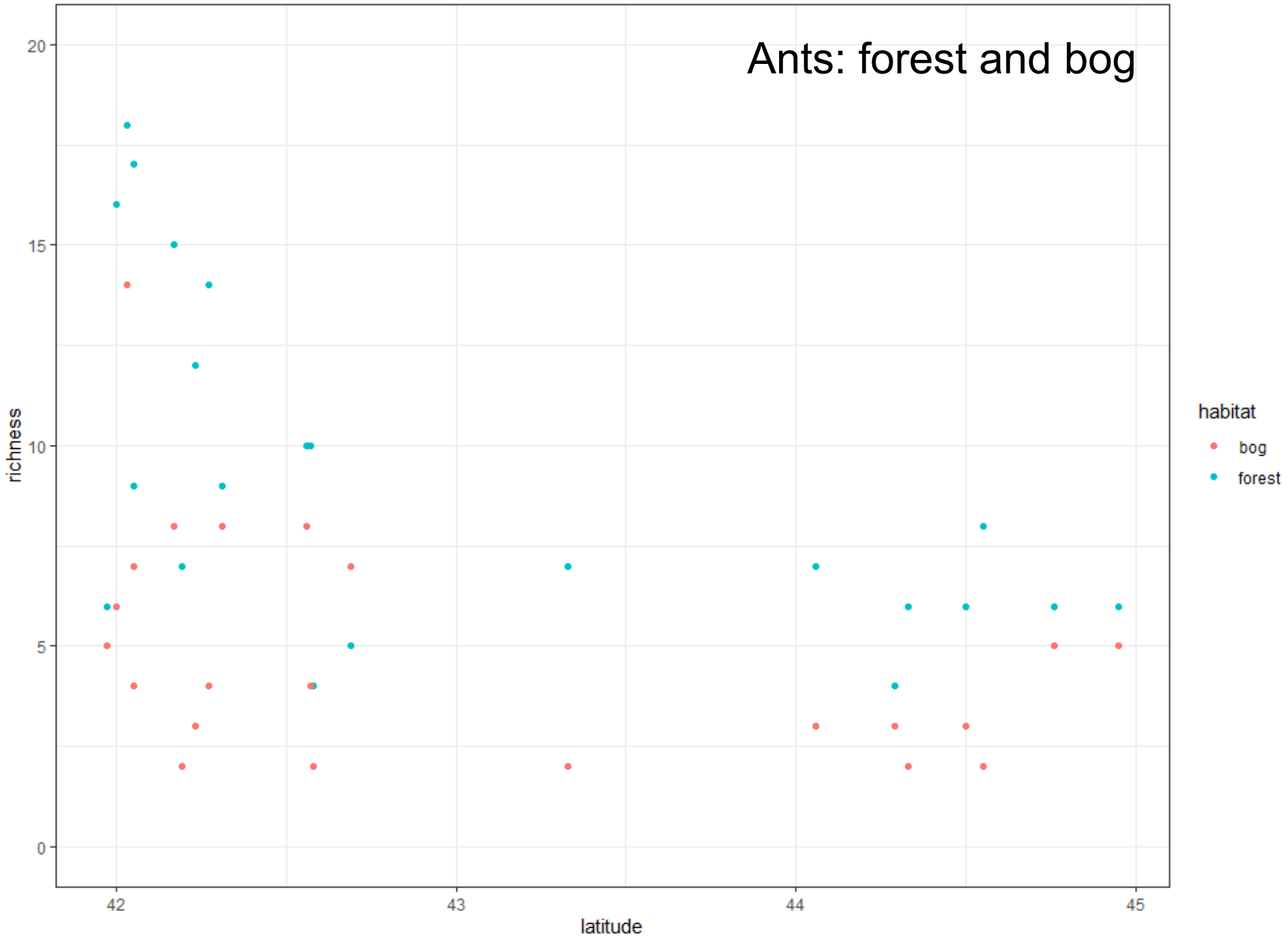
Training

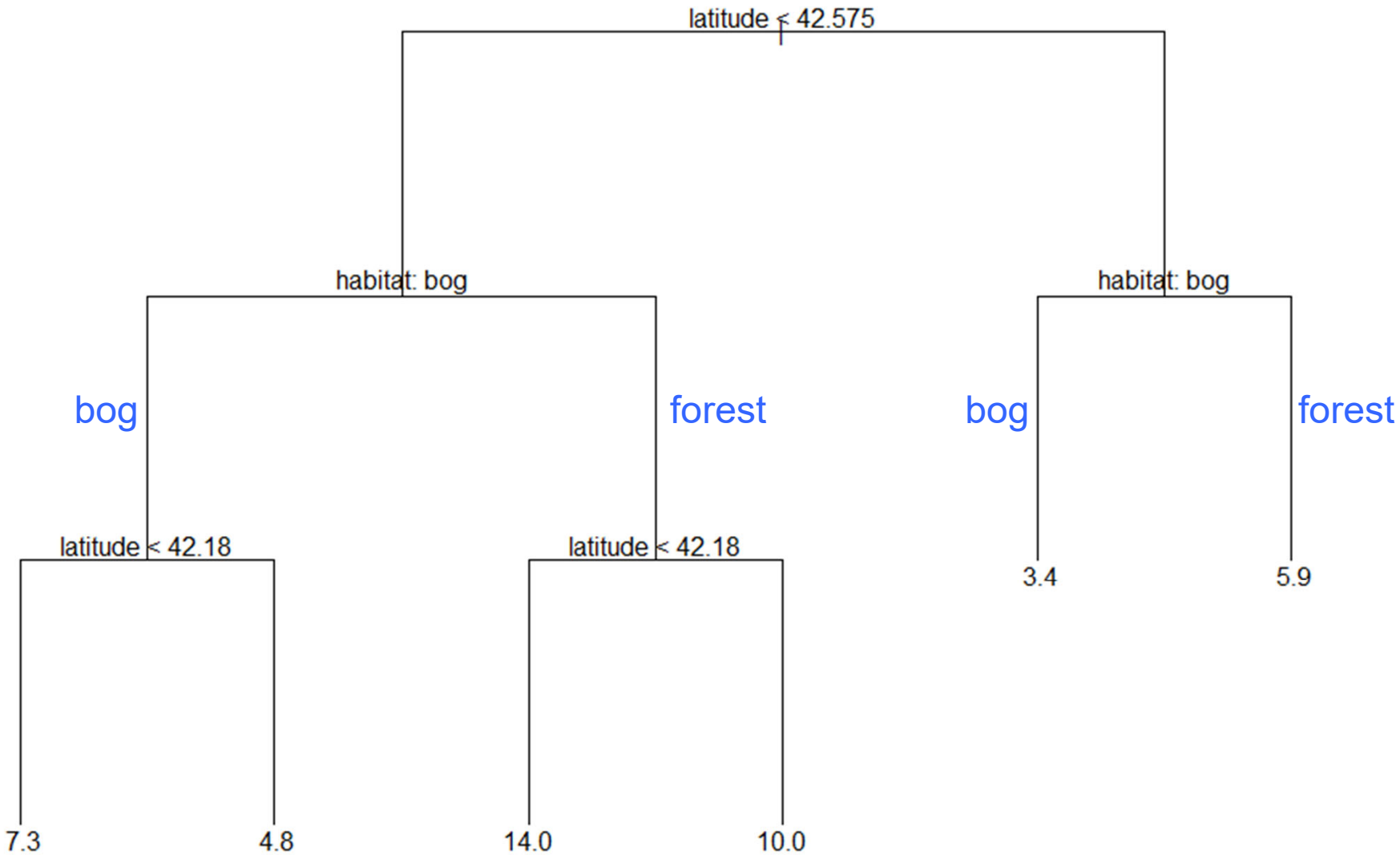
- Find splits that minimize training error
  - regression: SSQ
  - classification: Gini index or entropy
- Algorithm: binary recursive partitioning
  - split (minimize error)
  - if stop = FALSE, split again, ...
- Stopping rules
  - e.g. degree of error improvement, data per node

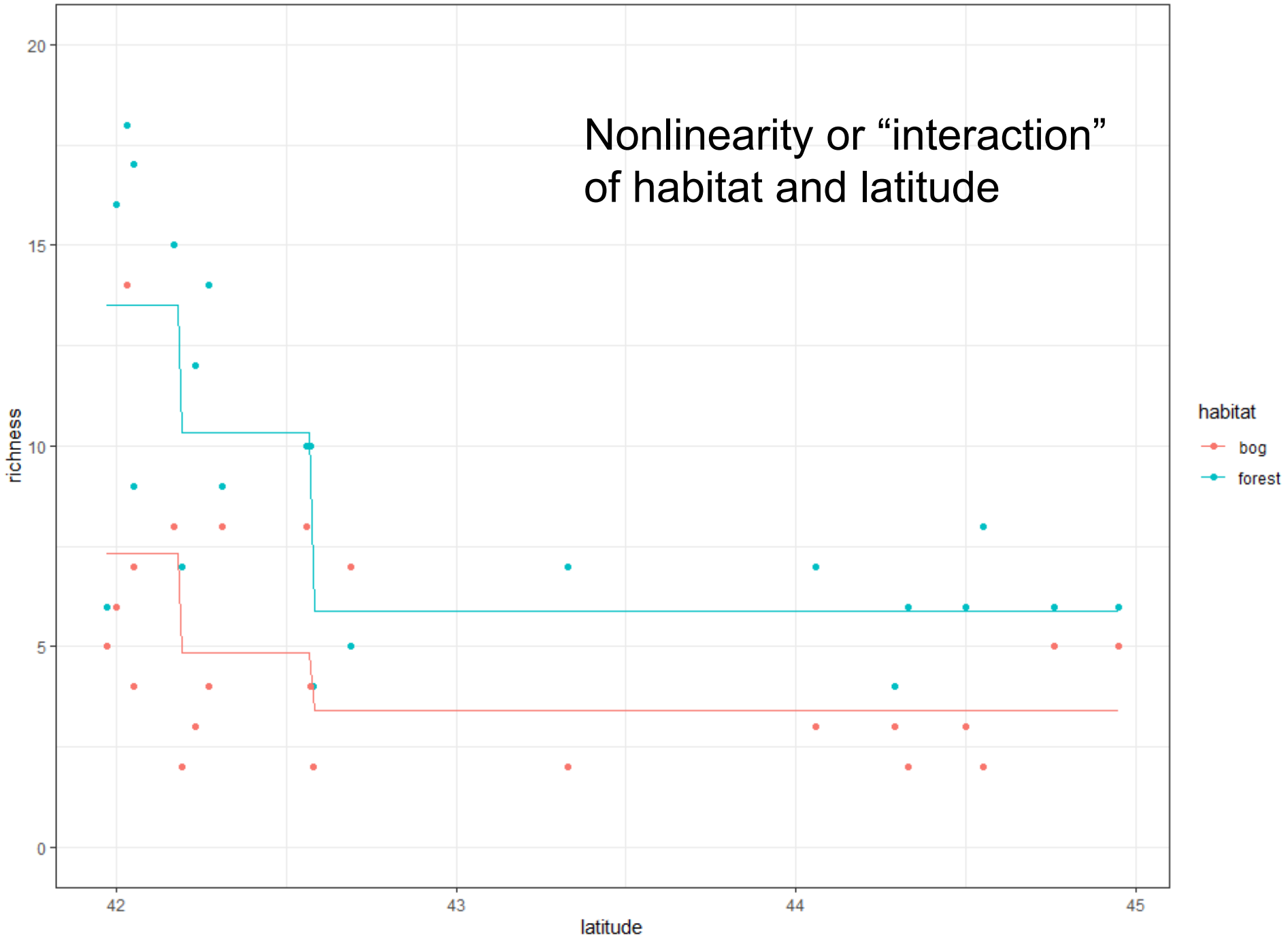
Same data, deeper tree




20
15
10
5
0
richness
42
43
44
45
latitude

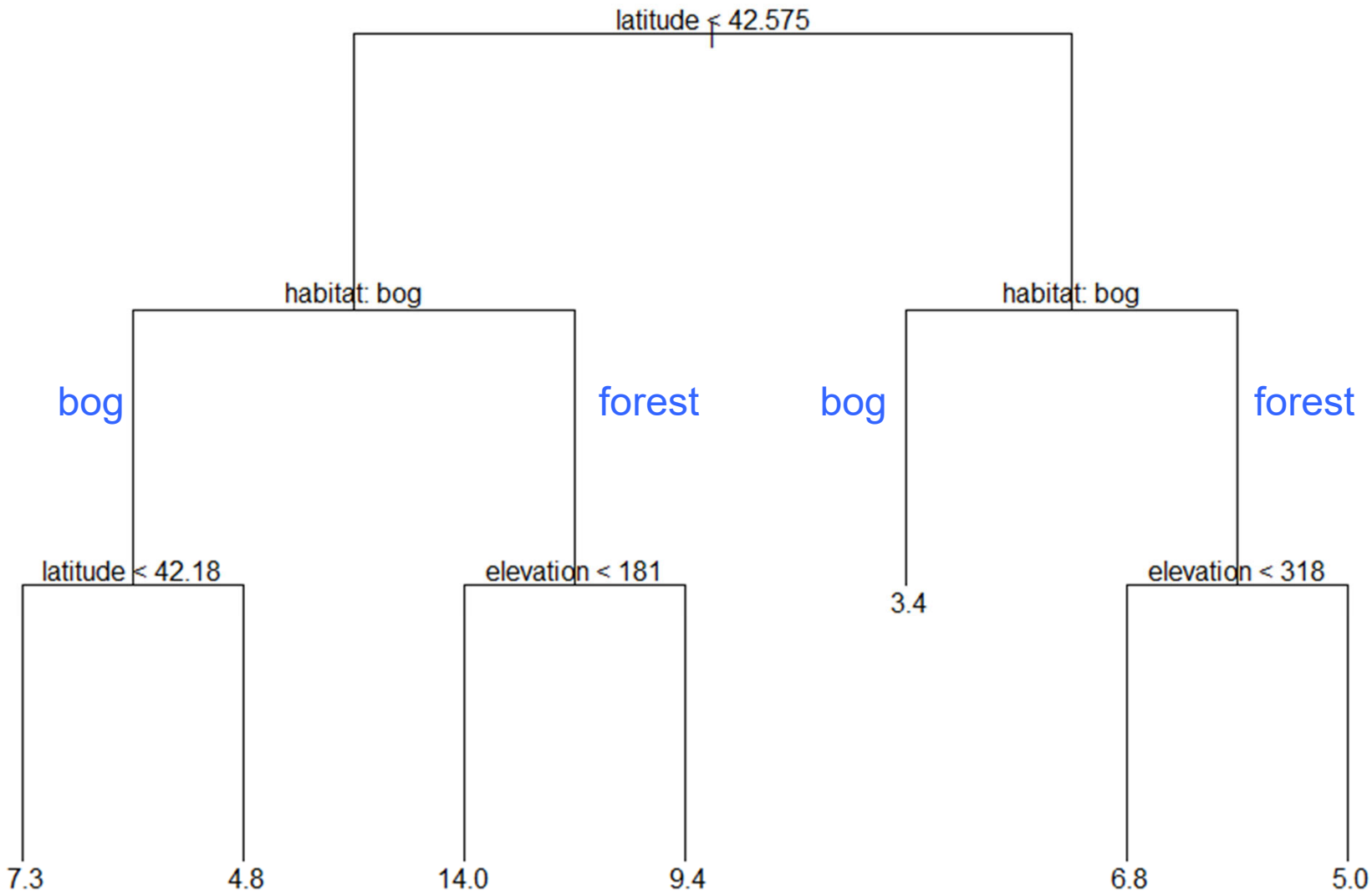

Ants: forest and bog
richness
20
15
10
5
0
42
43
44
45
latitude
habitat
bog
forest

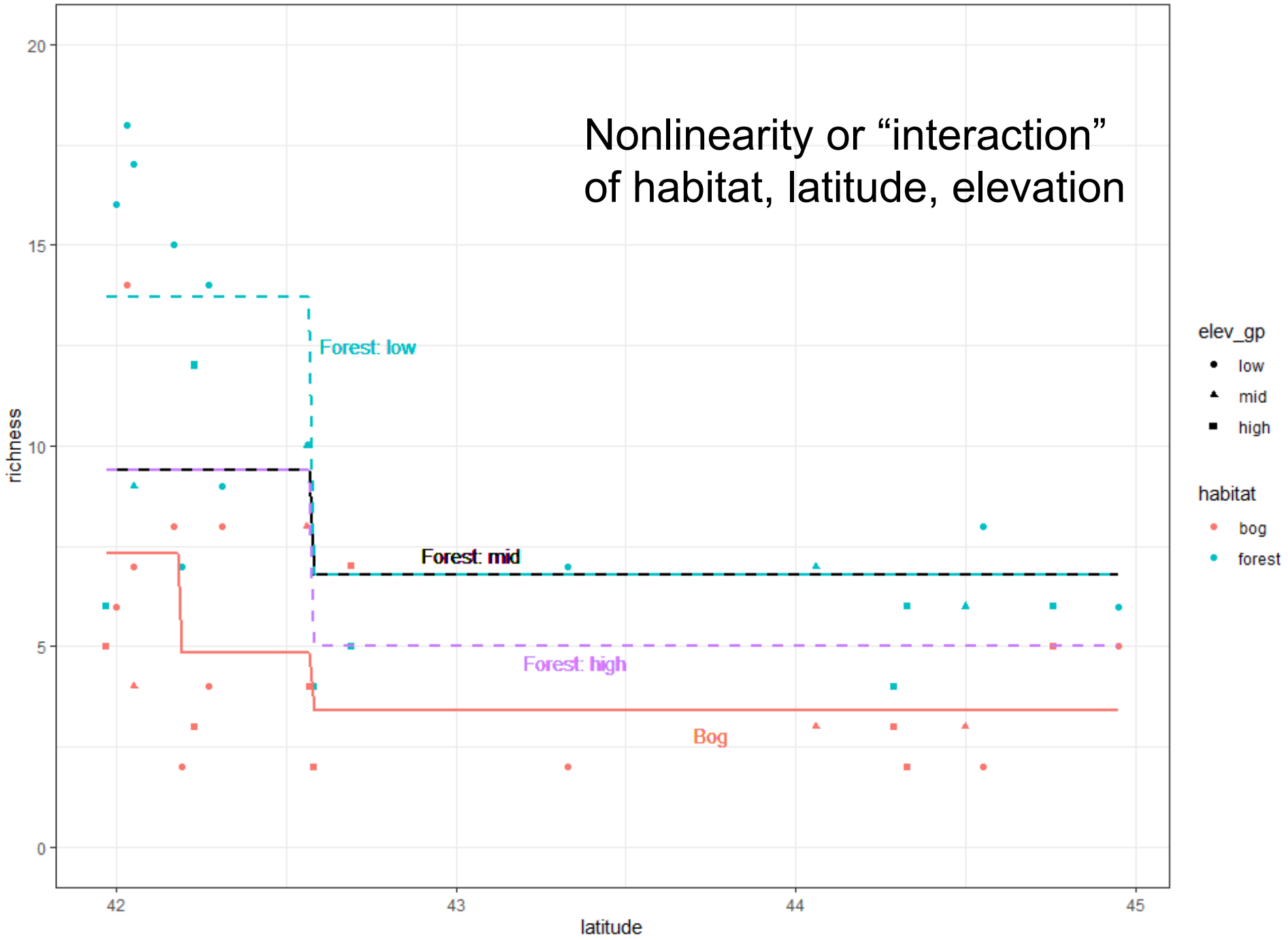

latitude < 42.575
habitat: bog
habitat: bog
bog
forest
bog
forest
latitude < 42.18
latitude < 42.18
3.4
5.9
7.3
4.8
14.0
10.0


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
```

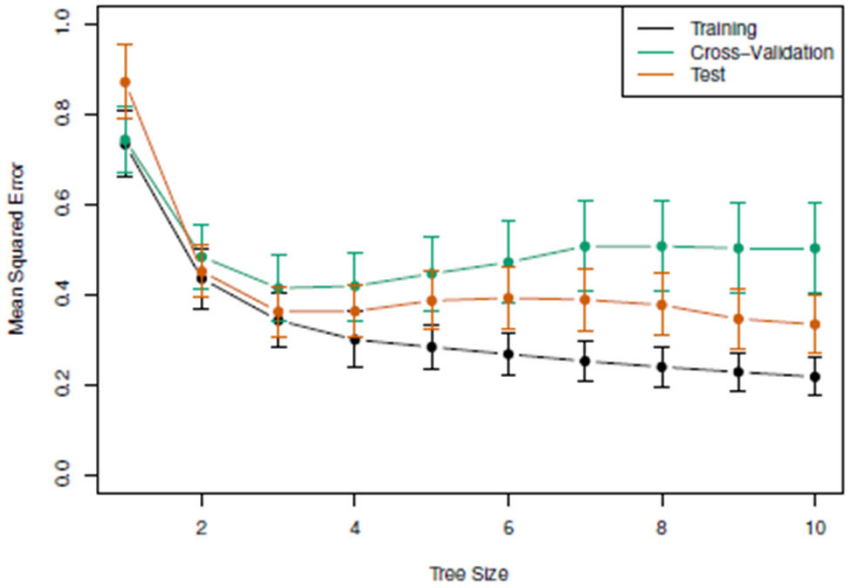
All 3 predictors




Nonlinearity or “interaction”
of habitat, latitude, elevation
Forest: low
Forest: mid
Forest: high
Bog
richness
latitude
20
15
10
5
0
42
43
44
45
elev_gp
low
mid
high
habitat
bog
forest

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”


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