NRES_798_18_201501

Data mining, Regression trees

Data mining (e.g. forestry)

- Analyzing bioclimatic factors affecting species presence
- Mapping forest types from remotely sensed data
- Predicting forest attributes over large geographic areas
- Identifying suitable wildlife habitat
- Making sense of complex ecological data sets (100+ variables)
- Predicting microhabitat affecting fish species distributions
- Model the distribution of vegetation alliances
- Assessing biological indicators of environmental factors affecting fish habitat
- Identifying fuels characteristics for fire spread models

Which variables to select?

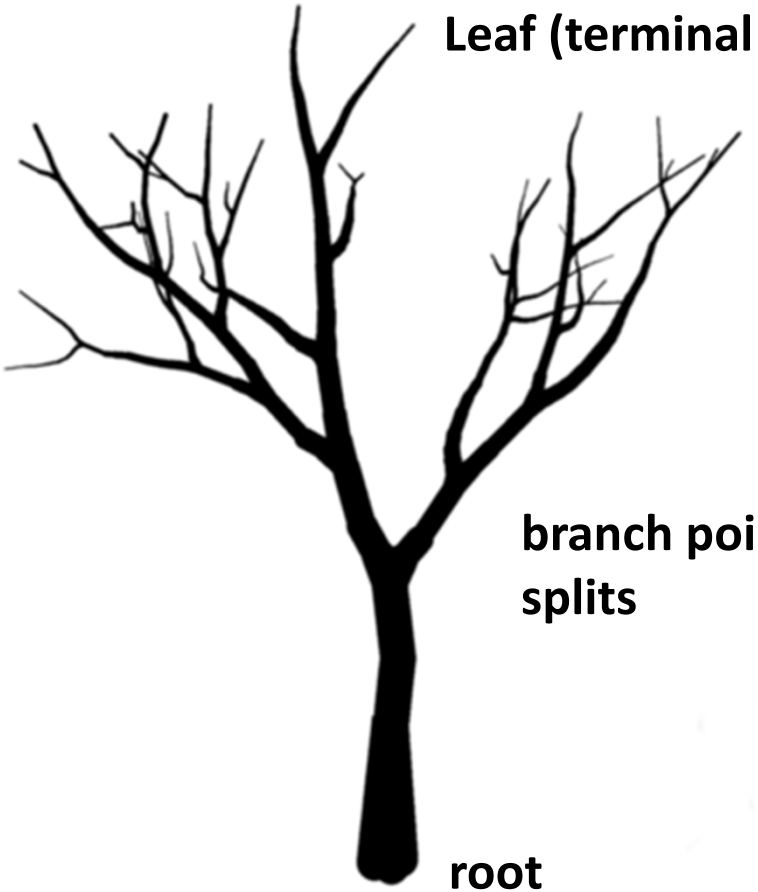
Predictor variables for determining the presence or absence of house finches

Name of predictor variable
Proportion of human housing units vacant[b]
No. half-d observation[c,d]
D from season start[c,d]
Latitude[c,d]
FeederWatch season[c,d]
Human population density[b,d]
No. bird feeders, hanging[c]
Elevation, U.S. Geological Survey National Elevation dataset[e]
Elevation, GTOPO30 Digital Elevation model[d,e]
Hr of observation effort[c,d]
No. water sources for birds[c]
Human households density[b]
No. feeders, suet[c]
30-yr annual average of monthly snowfall amt[e]
Longitude[c]
Mean no. children/human household[b]
No. bird feeders, thistle[c]
No. human family households[b]
Proportion humans 30–39 yr old[b]
No. bird feeders, platform[c]
Density of humans with multiple races[b]
No. feeders, ground[c]
Max. temp during observation period[c]
Min. temp during observation period[c]
No. deciduous shrubs and trees within site[c]
Snow depth during observation period[c]

TREE models

- Computationally intensive
- Used when there are many explanatory variables and you need to determine which are important.
- Machine-learning method
- Designed to construct prediction models from data
- Models obtained by **recursive partition** of data

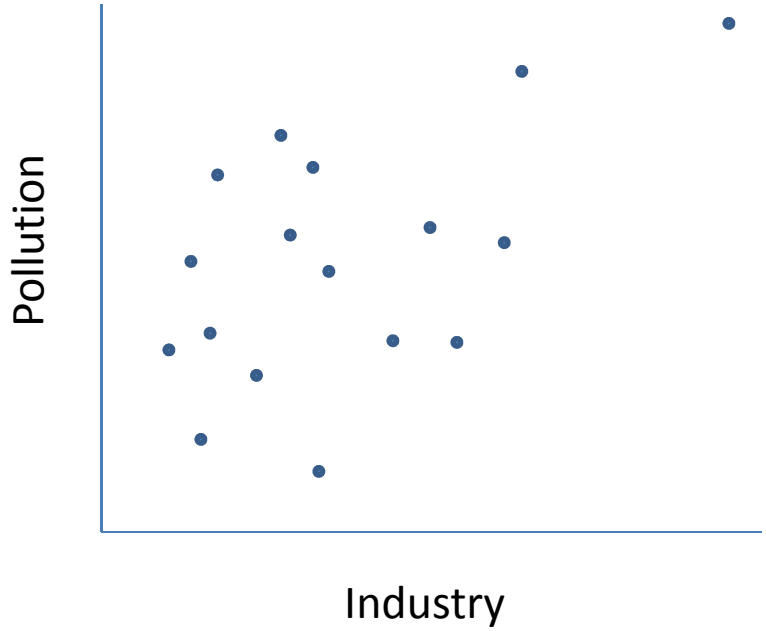
Classification tree



- Leaf: mean value for a subset of the data
  - Value is pure or data is too limited to split further
- Branch points: splitting points that are defined by given independent variables (which variable, what value)
- Root: full dependent data set

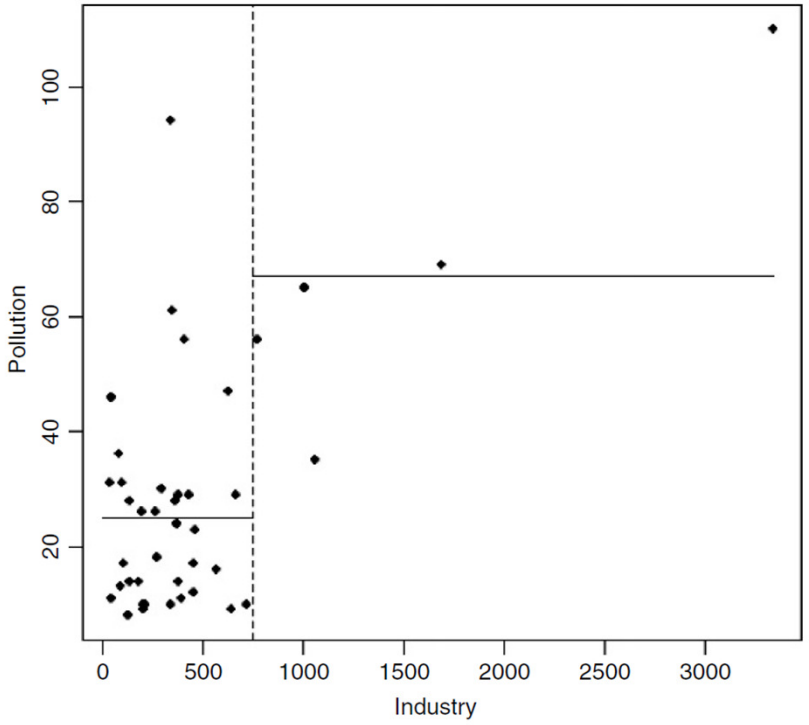
Binary recursive partitioning

- Predictors for pollution
  - Temperature
  - Industry
  - Population
  - Wind
  - Rain
  - # wet days



Binary recursive partitioning

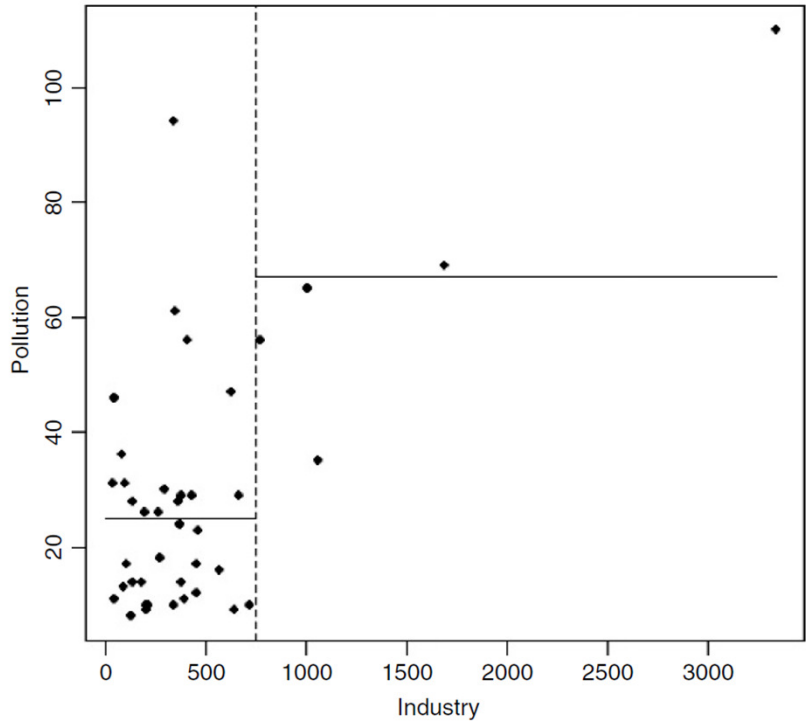
- Step through each predictor variable
- Select a threshold value for the predictor variable
- Calculate the mean of the response variable above and below the threshold
- Use these means to calculate the deviance
- Work through all possible values of the threshold
- Select the threshold that produces the lowest deviance
- Split the data based on this threshold
- Repeat for the data on each side of the threshold

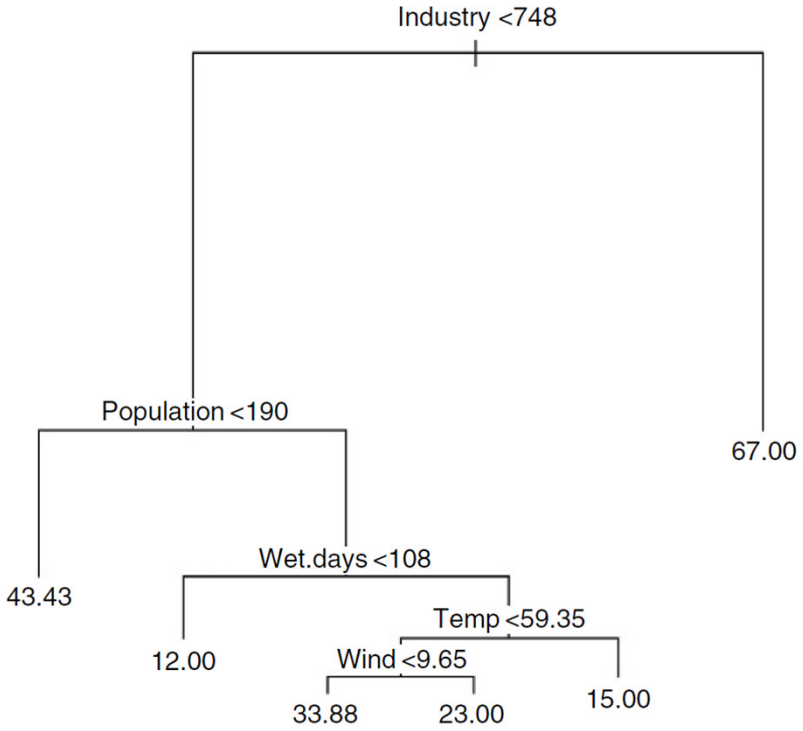


Industry

Population, temperature, wind, rain, wet days

Binary Recursive partitioning





- Select predictor variable and threshold that result in the smallest deviance
- Branch points are defined by the threshold for the predictor variable
- Tree leaf values indicate the mean of the response variable for the leaf

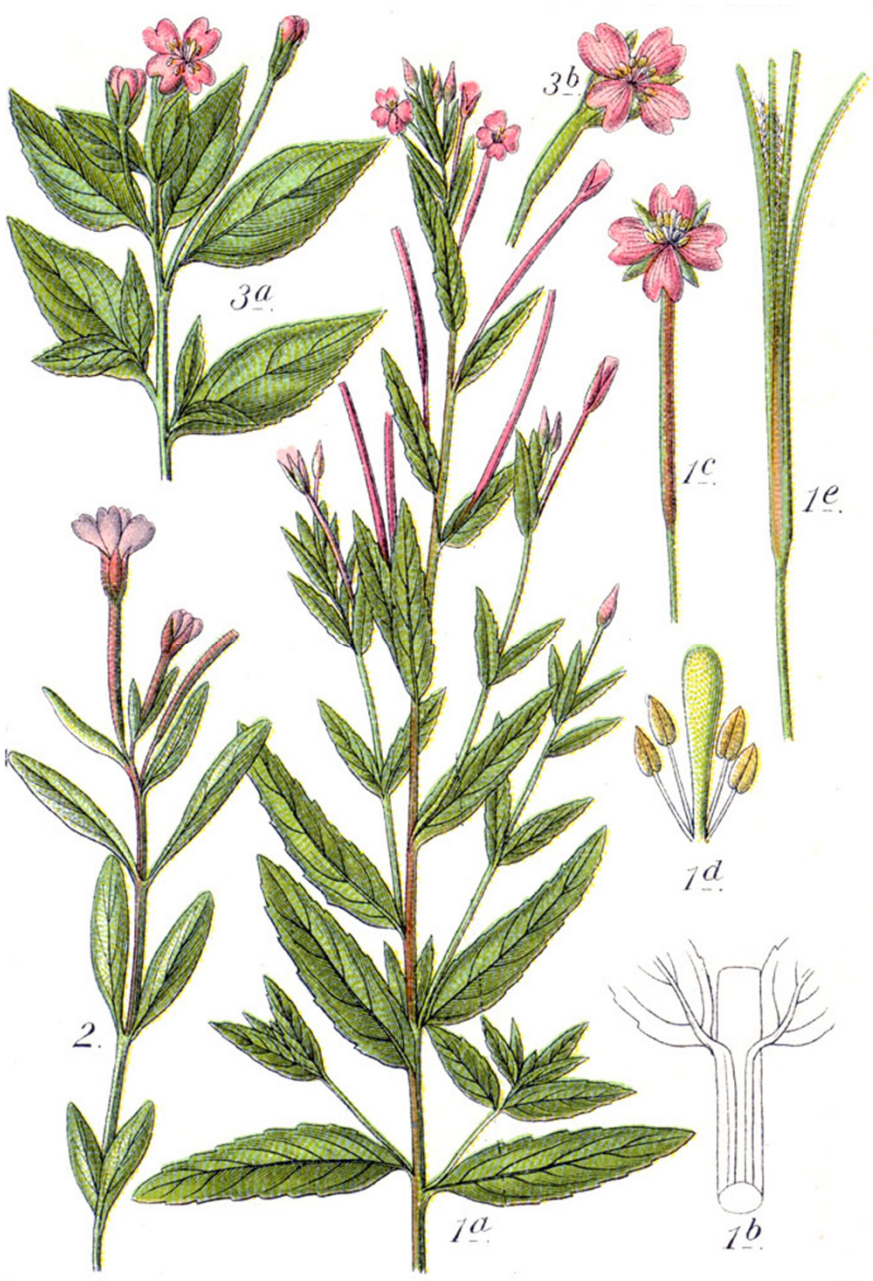
Classification and Regression TREES

- Classification tree
  - If response variable is categorical
    - E.g. species identification tree
- Regression tree
  - If response variable is continuous
    - Predictor variables can be continuous or categorical

Classification tree

- Separating species in the genus *Epilobium*
- *9 species*

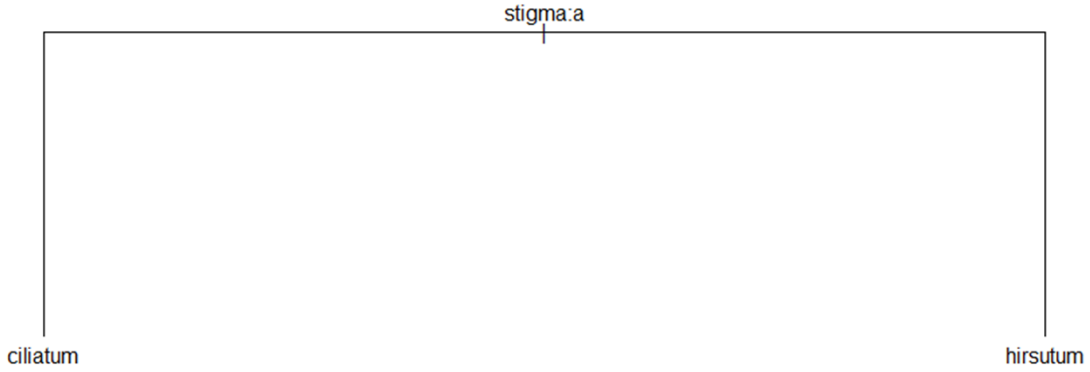
- *8 categorical traits*
  - *Stigma*
  - *Stem.hairs*
  - *Gladular.hairs*
  - *Seeds*
  - *Pappilose*
  - *Stolons*
  - *Petals*
  - *Base*



Classification tree: *Epilobium*

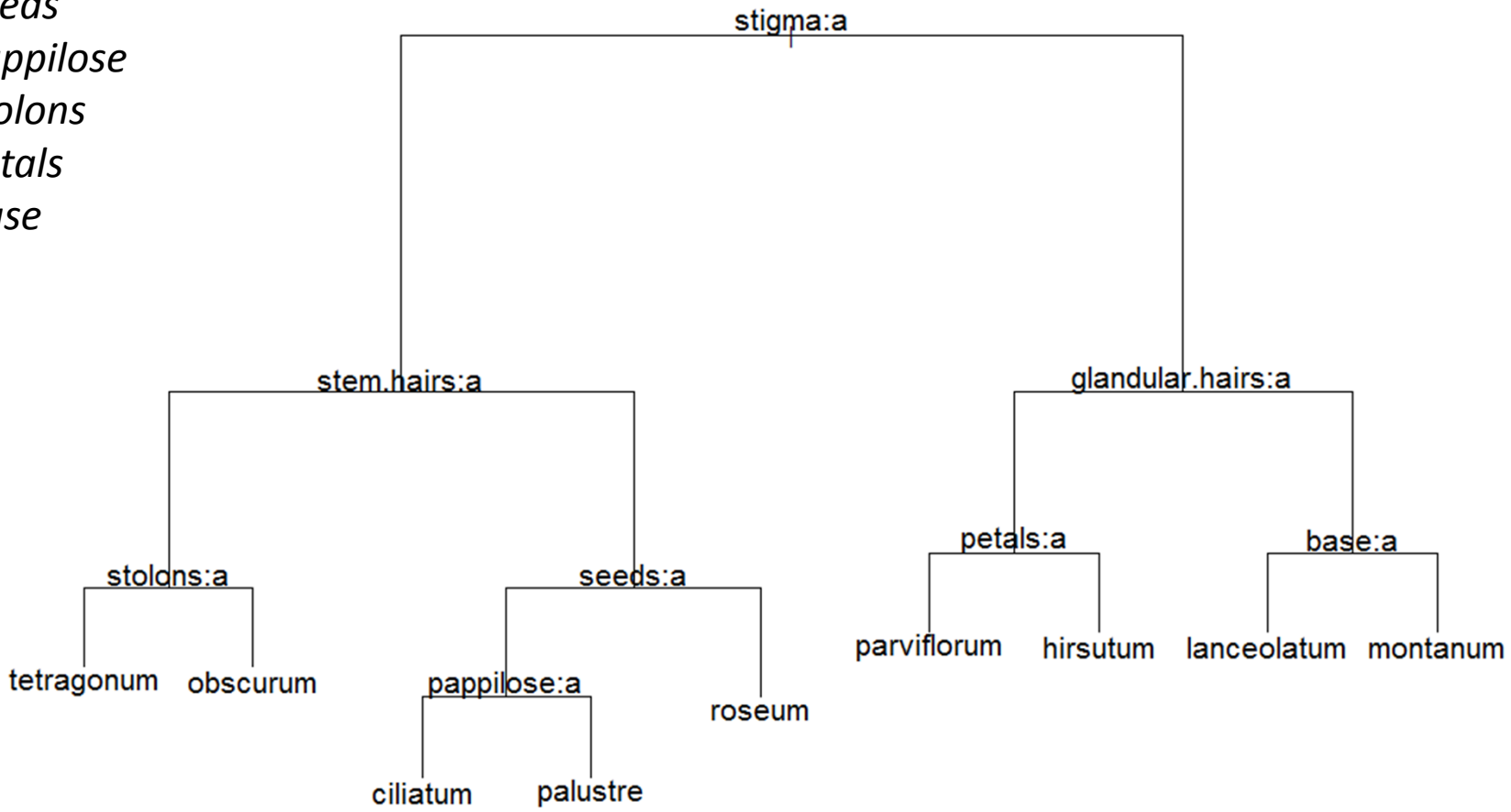
species	stigma	stem.hairs	gland.hairs	seeds	pappilose	stolons	petals	base
hirsutum	lobed	spreading	absent	none	uniform	absent	>9mm	rounded
parviflorum	lobed	spreading	absent	none	uniform	absent	<10mm	rounded
montanum	lobed	spreading	present	none	uniform	absent	<10mm	rounded
lanceolatum	lobed	spreading	present	none	uniform	absent	<10mm	cuneate
tetragonum	clavate	appressed	present	none	uniform	absent	<10mm	rounded
obscurum	clavate	appressed	present	none	uniform	stolons	<10mm	rounded
roseum	clavate	spreading	present	none	uniform	absent	<10mm	cuneate
palustre	clavate	spreading	present	appendage	uniform	absent	<10mm	rounded
ciliatum	clavate	spreading	present	appendage	ridged	absent	<10mm	rounded

- Categorical response and predictor variables



Classification tree: *Epilobium*

- *Stem.hairs*
- *Gladular.hairs*
- *Stigma*
- *Seeds*
- *Pappilose*
- *Stolons*
- *Petals*
- *Base*



TREES in R

```
install.packages("tree")
library(tree)

m1 <- tree(responce.variable ~ ., data ,mincut = 5, minsize = 10, mindev= 0.001)
```

- Worm density
  - Area
  - Slope
  - Vegetation
  - Soil pH
  - Dampness

- mincut
  - Minimum number of points to occur on each side of split (default = 5)
- minsize
  - Smallest node size possible (default = 10)
- mindev
  - Within node deviance must be greater than this value for a node to be split

Regression TREES

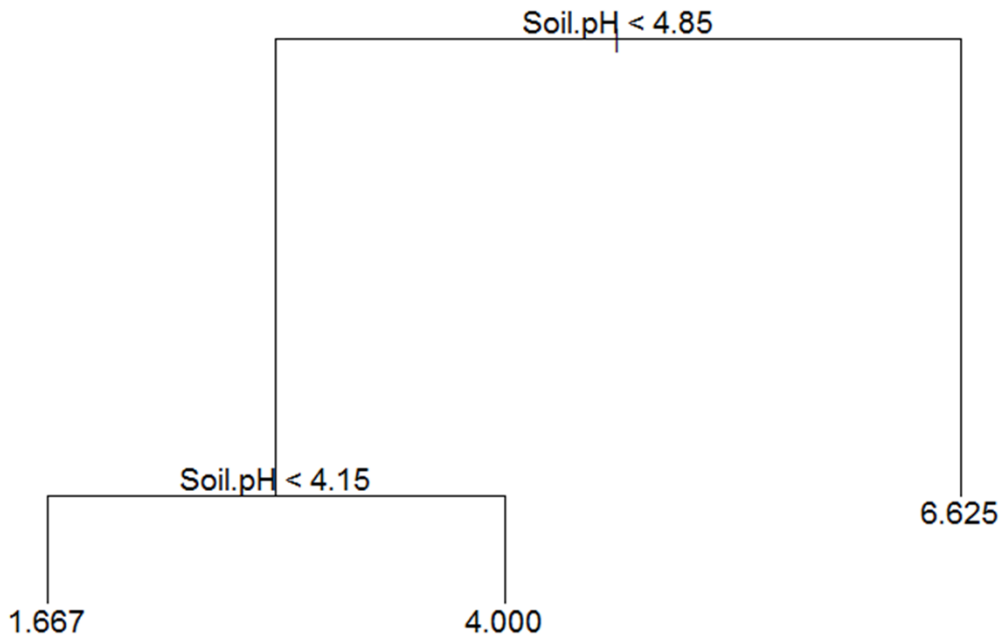
- Worm density
  - Area
  - Slope
  - Vegetation
  - Soil pH
  - Dampness

```
m1 <- tree(Worm.density ~ Area + Slope + Vegetation + Soil.pH + Damp, worms)

m1 <- tree(Worm.density ~ ., worms[,2:7], minsize=10)
```

Regression TREES

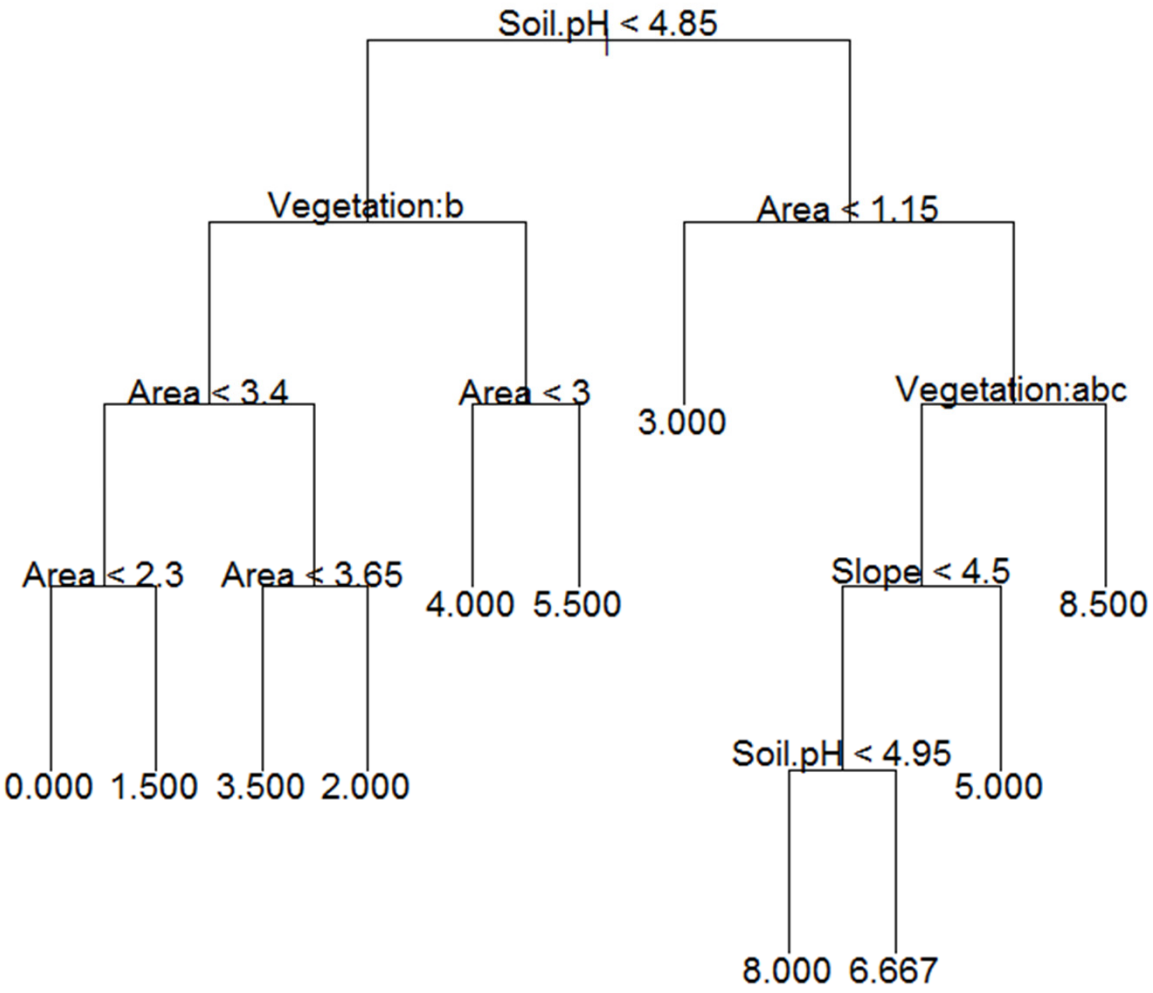
```
m1 <- tree(Worm.density ~ ., worms[,2:7], minsize=10)
plot(m1)
text(m1,cex=1.2)
```



- Worm density
  - Area
  - Slope
  - Vegetation
  - Soil pH
  - Dampness

Regression TREES

```
m1 <- tree(Worm.density ~ ., worms[,2:7], minsize=10)
plot(m1,type = c("uniform"))
text(m1,cex=1.2)
```



- Worm density
  - Area
  - Slope
  - Vegetation
  - Soil pH
  - Dampness

1) root 20 130.5000 4.350
  2) Soil.pH < 4.85 12 35.6700 2.833
    4) Vegetation: Grassland 8 10.8800 1.875
      8) Area < 3.4 5 2.8000 1.200
        16) Area < 2.3 1 0.0000 0.000 *
        17) Area > 2.3 4 1.0000 1.500 *
      9) Area > 3.4 3 2.0000 3.000
        18) Area < 3.65 2 0.5000 3.500 *
        19) Area > 3.65 1 0.0000 2.000 *
    5) Vegetation: Arable,Scrub 4 2.7500 4.750
      10) Area < 3 2 0.0000 4.000 *
      11) Area > 3 2 0.5000 5.500 *
  3) Soil.pH > 4.85 8 25.8800 6.625
    6) Area < 1.15 1 0.0000 3.000 *
    7) Area > 1.15 7 10.8600 7.143
      14) Vegetation: Arable,Grassland,Meadow 5 5.2000 6.600
        28) Slope < 4.5 4 2.0000 7.000
          56) Soil.pH < 4.95 1 0.0000 8.000 *
          57) Soil.pH > 4.95 3 0.6667 6.667 *
        29) Slope > 4.5 1 0.0000 5.000 *
      15) Vegetation: Orchard,Scrub 2 0.5000 8.500 *

Node
Split
N
Deviance
Yval

CART
(Classification And Regression TREEs)

- Strengths
  - Very simple to perform
  - Require very few assumptions or a-priori estimates
  - No assumptions regarding underlying distributions
  - Interactions are fundamental to the analysis but don't need to be pre-defined
  - Non-linear relationships between parameters do not affect tree performance
  - Interpretation of TREE can be simple and informative

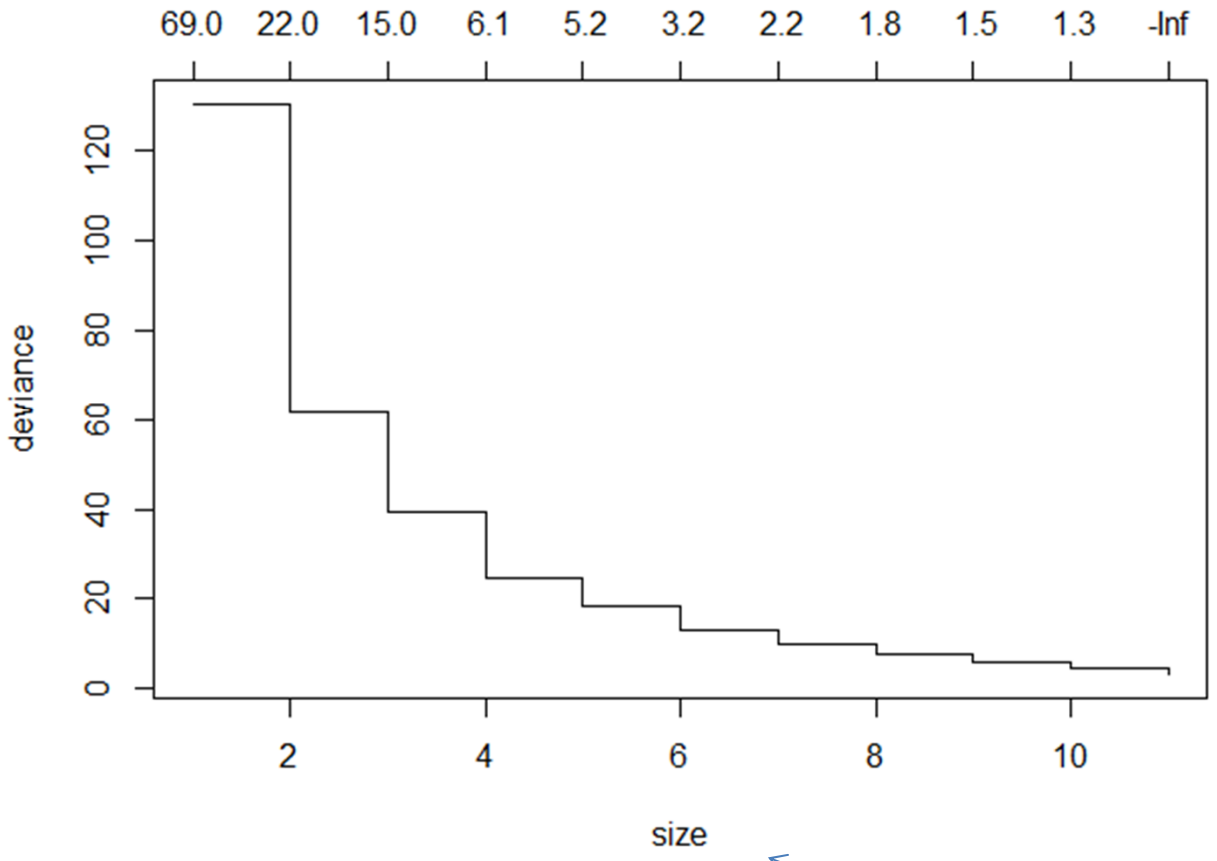
- Weaknesses
  - Computationally intensive (normally not a problem)
  - Solution not linked to biological processes (no functional form)
  - Propensity to be overfit
  - Tree structure can be sensitive to "training data" peculiarities
  - No indication of how selected tree compares to other "potential" trees

Pruning TREES

- CART models are prone to being over-fit
- Often you wan to “prune” the tree down to a more limited set of predictor variables
- Trees can be simplified by snipping of the least important splits using a cost-complexity measure

Pruning TREES

- m1 <- tree(Worm.density ~ .,worms[,2:7],minsize=3)
- m2 <- prune.tree(m1)
- plot(m2)

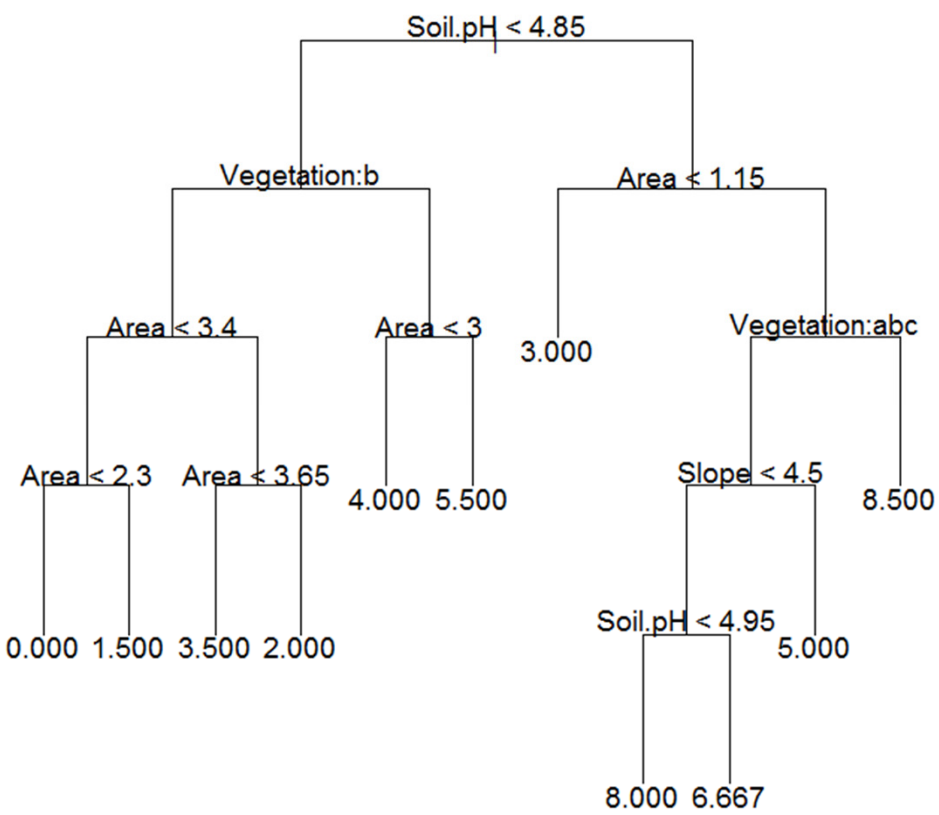


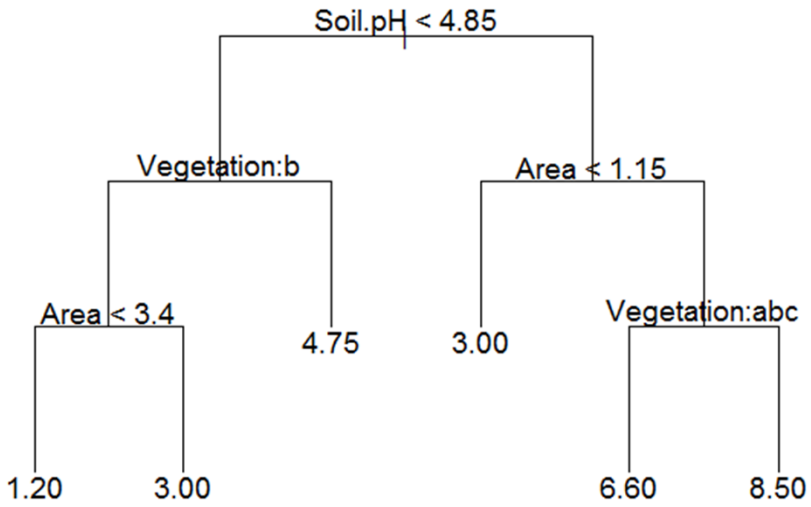
Number of nodes
(complexity of tree)

m3 <- prune.tree(m1, best = 5)

Pruning TREES

```
m3 <- prune.tree(m1,best=6)
plot(m3,type = c("uniform"))
text(m3,cex=1.2)
```





Random Forest

```
install.packages("randomForest")
library(randomForest)
```

- Aim is to improve the predictive accuracy and generality of TREE model
  - Use bootstrapping to create a large number of trees
  - This builds a “forest” of potential tree structures
  - Combine results across all trees to derive a more robust TREE model