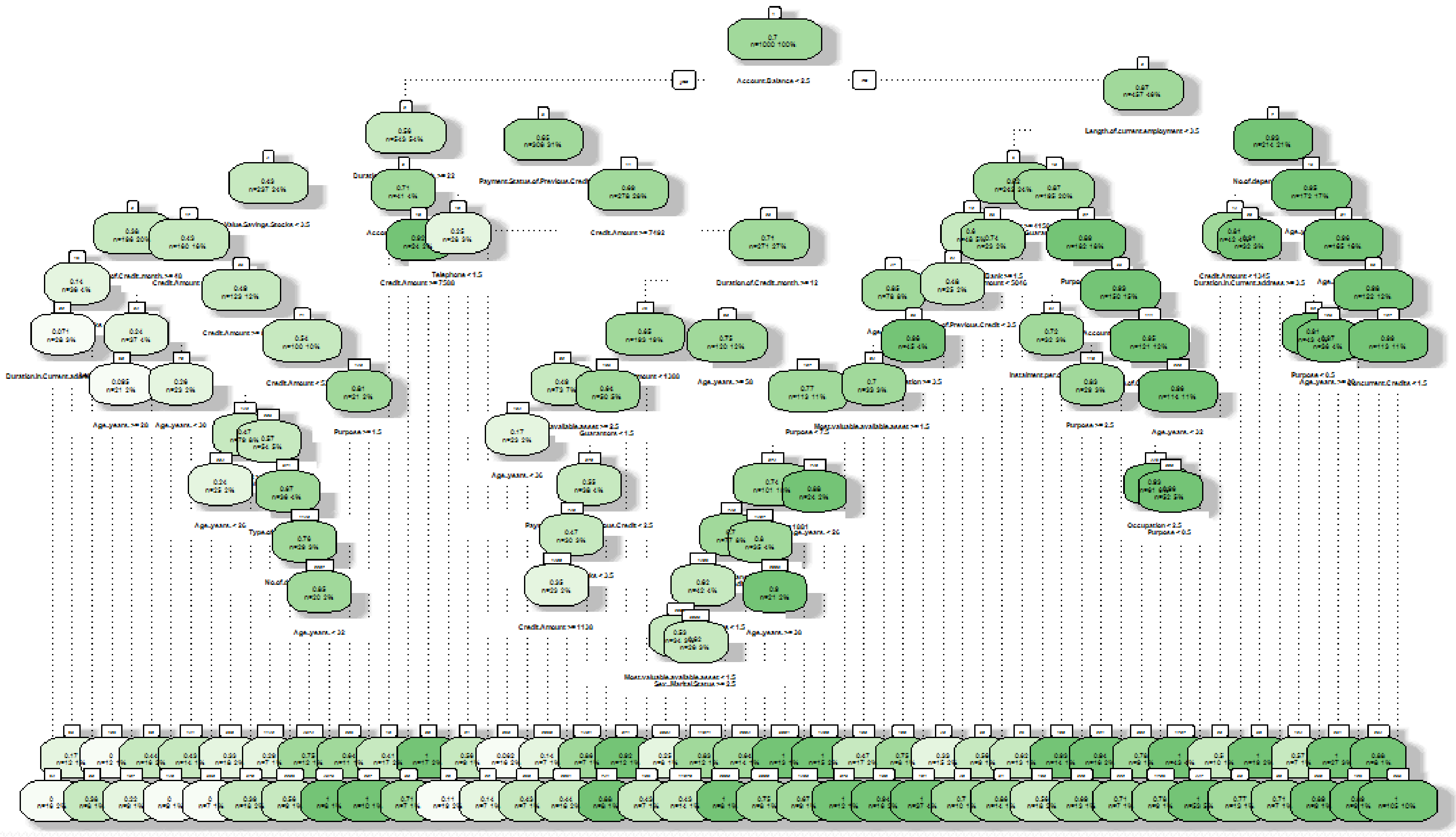
Decision Trees with R

Patrick Collins

Sample Decision Tree



Metrics

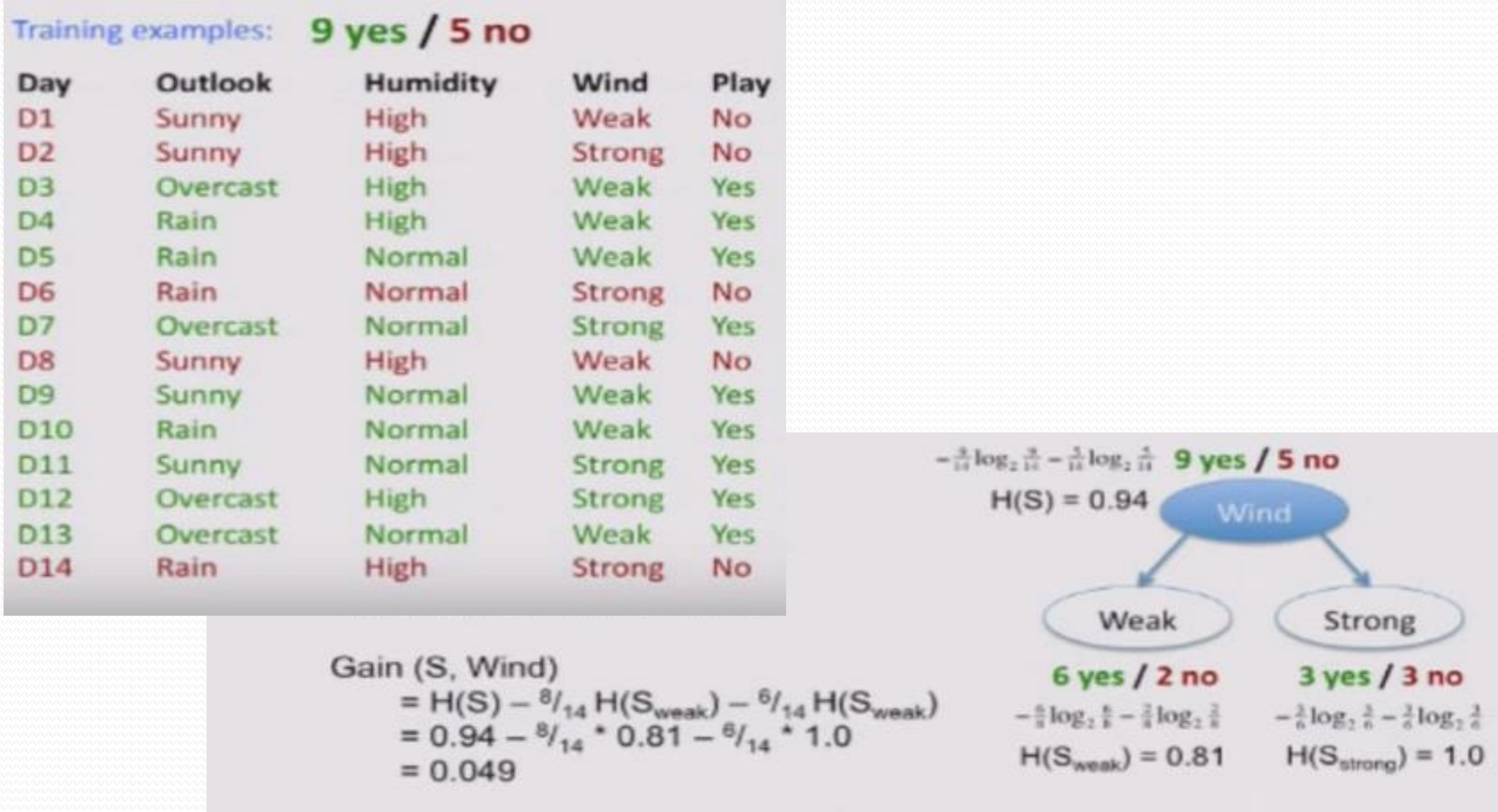
- **Gini impurity**
  - “how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset”[wiki]
- **Information gain**
  - **Entropy**
- **Variance reduction**

Information gain

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Categorical Variables



Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

$-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}$ 9 yes / 5 no

$H(S) = 0.94$

Wind

Weak

Strong

6 yes / 2 no

$-\frac{6}{8}\log_2\frac{6}{8}-\frac{2}{8}\log_2\frac{2}{8}$

$H(S_{weak}) = 0.81$

3 yes / 3 no

$-\frac{3}{6}\log_2\frac{3}{6}-\frac{3}{6}\log_2\frac{3}{6}$

$H(S_{strong}) = 1.0$

Gain (S, Wind)

$$= H(S) - {}^{8}/_{14}\, H(S_{weak}) - {}^{6}/_{14}\, H(S_{weak})$$
$$= 0.94 - {}^{8}/_{14} * 0.81 - {}^{6}/_{14} * 1.0$$
$$= 0.049$$

https://www.youtube.com/watch?v=nodQ2soCUbI

Continuous variables

- Recursive Partitioning
- Using mean as a piecewise constant for splits

- http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Rpart

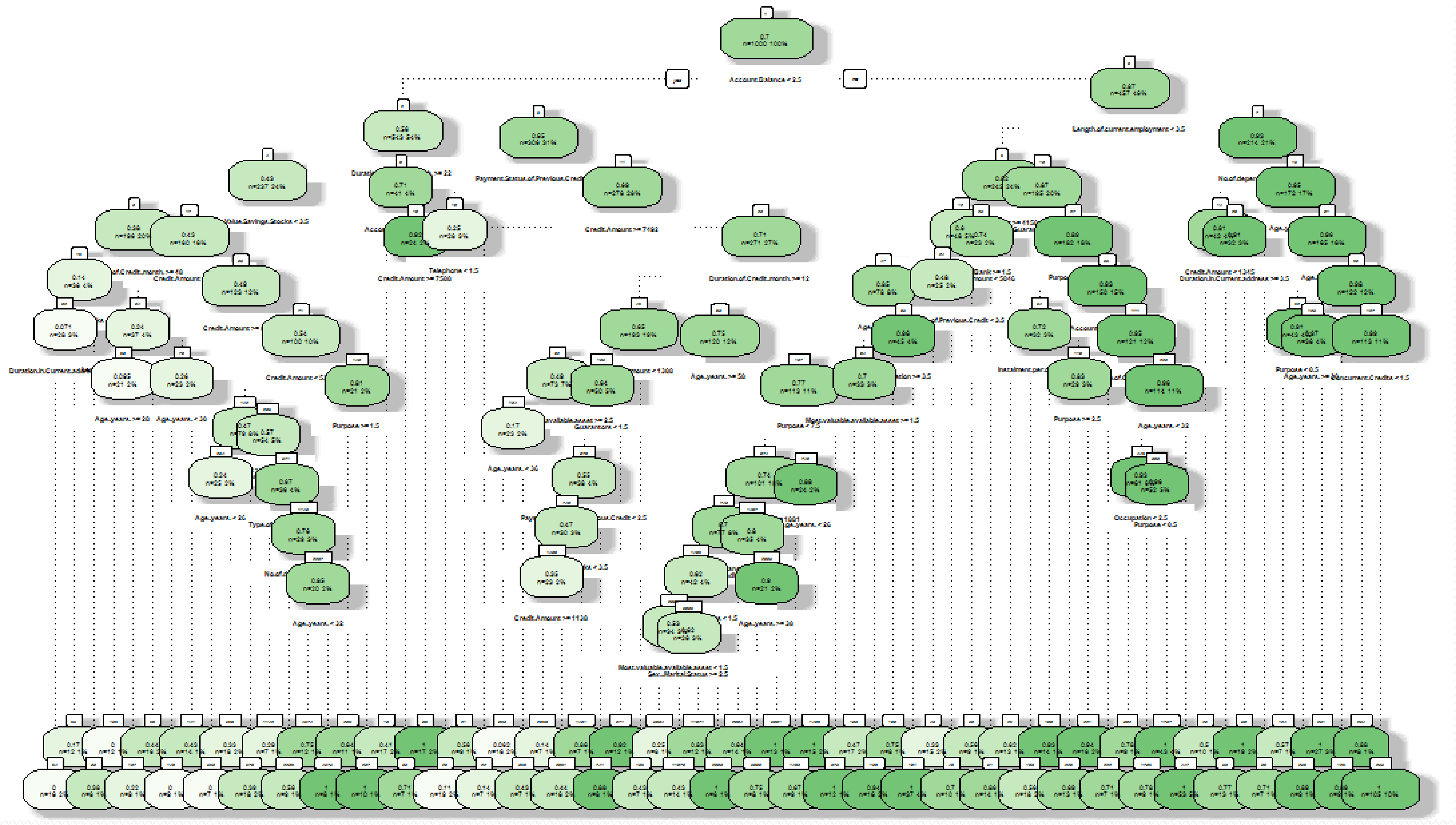
- Primarily tuned with the complexity parameter CP

- Also has
  - Minsplit
  - Minbucket
  - Maxcompete
  - Maxsurrogate
  - Usesurrogate
  - Xval (number of cross-validations.
  - Surrogatestyle
  - maxdepth

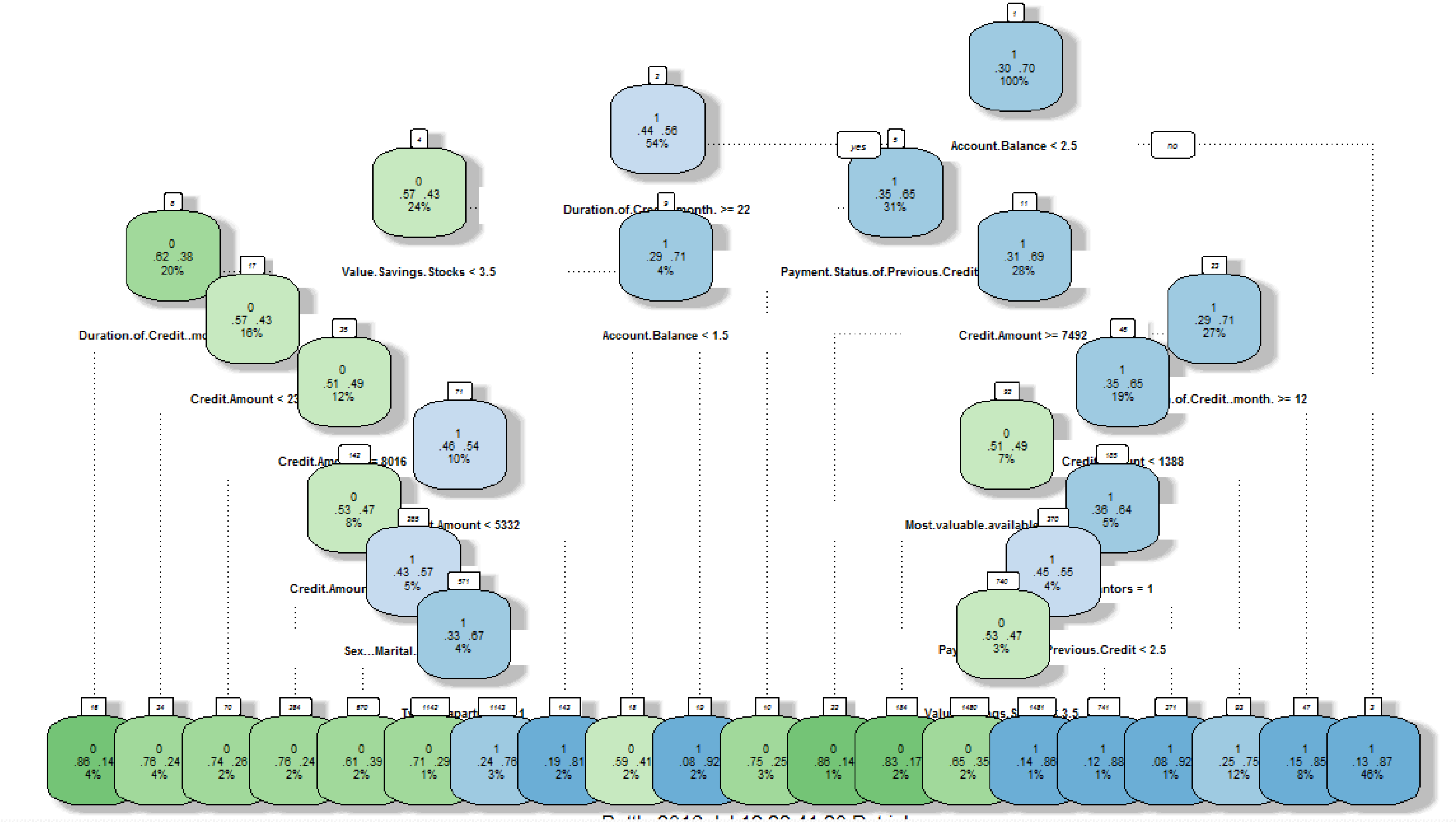
Variable importance

- Good for variable selection for other models
- Current understanding
  - Sum of information gain for each variable
  - Variables with many levels tend to rank highly but is an artefact of too many levels. >20 levels seems best (RoT)

Compare tree size - 147 leaves

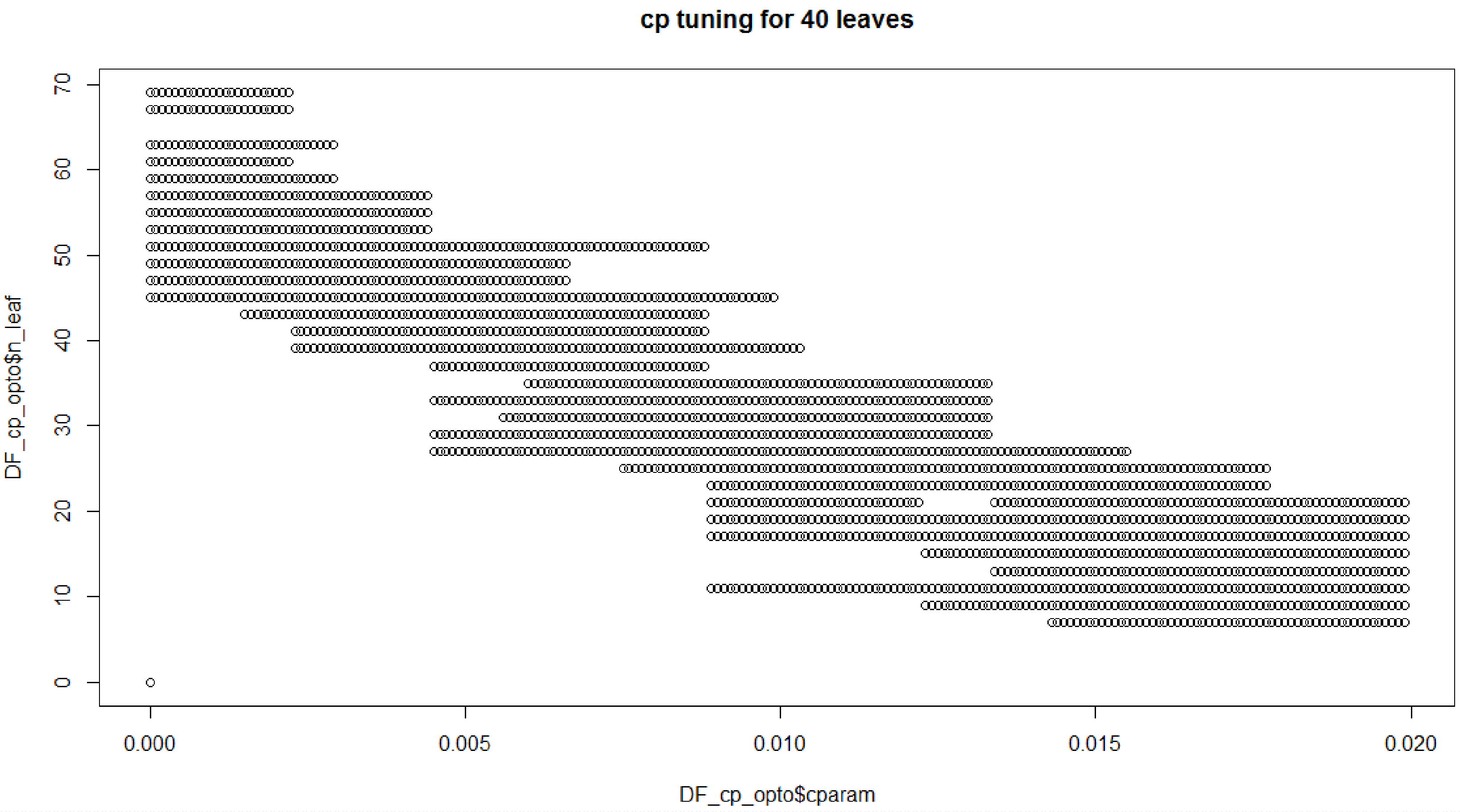


Compare tree size - 40 leaves



Programme Live

Good luck to ya


cp tuning for 40 leaves
DF_cp_opto$n_leaf
0
10
20
30
40
50
60
70
0.000
0.005
0.010
0.015
0.020
DF_cp_opto$cparam