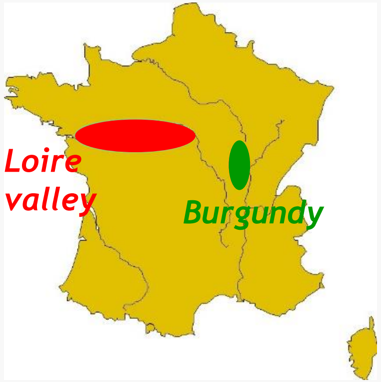
Regression trees and random forests as a tool for identifying the volatile organic compounds implied in the olfactory perception of wines

Evelyne Vigneau, Philippe Courcoux, Rémy Lefebvre,
Ronan Symoneaux, Angélique Villière


Pôle Agronomique Ouest
Oniris
IFV
INSTITUT FRANÇAIS DE LA VIGNE ET DU VIN
ESA
iFPC
Innovaroma Project 2012-2015

DATA: The wines

- 8 wines — Pinot Noir (PN) — Burgundy
- 8 wines — Cabernet Franc (CF) — Loire Valley



selected according to their scores of exemplarity

Loison A., Symoneaux R., Deneulin P., Thomas-Danguin T., Fant C., Guérin L., Le Fur Y. (2015); *Food Quality and Preference*, 40, 240-251.

DATA: Sensory analysis

- 16 wine trained panelists

- 33 aroma attributes (orthonasal perception)
  e.g.
  Cherry Stone,
  Fresh Black Currant,
  Cooked Cherry,
  Fresh Strawberry,
  (bell) Pepper,
  Woody,
  …

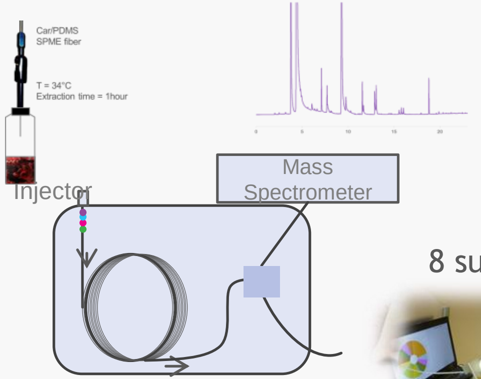
- Scores of intensity (0-10)

DATA: Volatile Organic Compounds (V.O.C.)

47 Volatile Organic Compounds

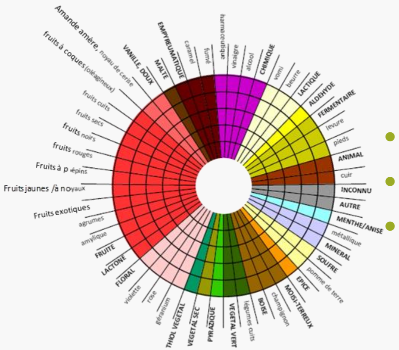
e.g.
3-Isobutyl-2-methoxypyrazine (IBMP),
3-mercapto-hexan-1ol (3MH-1ol),
Whyskeylactone,
4-Ethyl guaïacol,
…..

DATA: Olfactometry



8 subjects

68 olfactory compounds

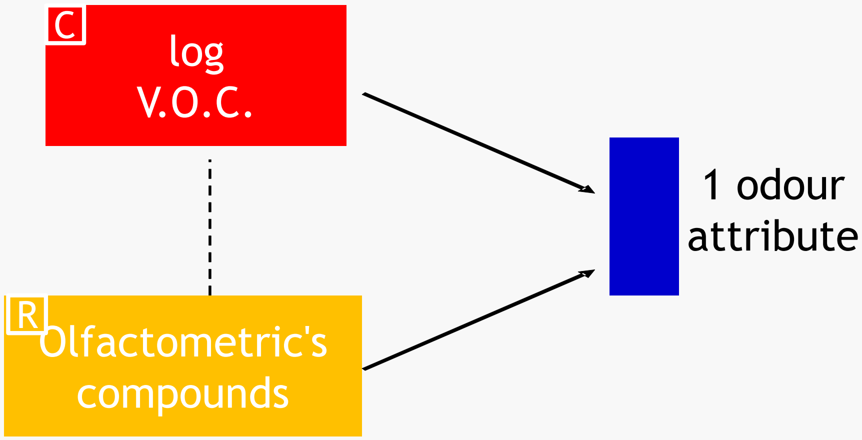


- Perception
- Description
- Intensity

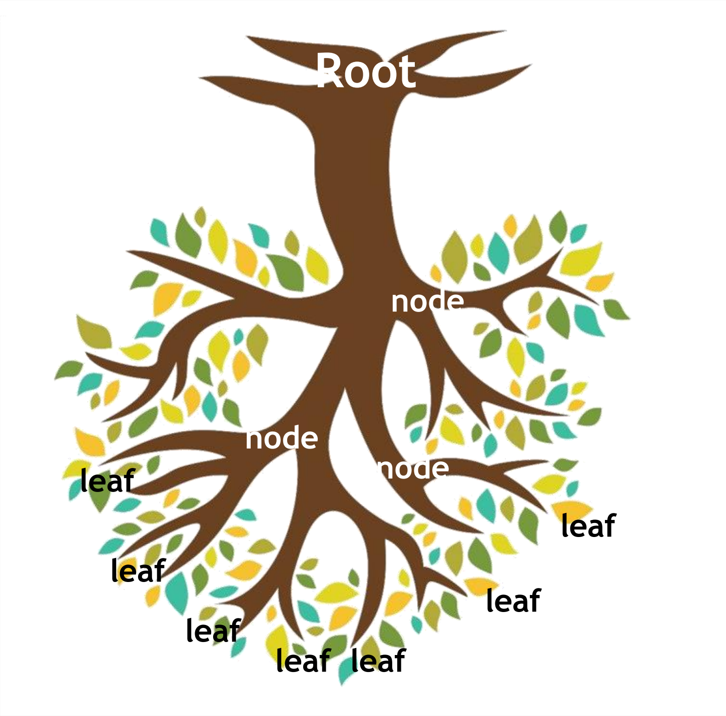
Villière, A., S. Le Roy, C. Fillonneau, F. Guillet, H. Falquerho, S. Boussely, C. Prost (2015). *Flavour journal*. DOI 10.1186/s13411-015-0034-0
Villière, A., C. Fillonneau, Prost, C. (2015). *IXth In Vino Analytica Scientia Symposium*, Trento (Italy),14-17 July 2015.

OBJECTIVE

Modelization and prediction of
the key aromatic characteristics of the wines (orthonasal perception)
by taking account of their V.O.C.s profiles and olfactory compounds.

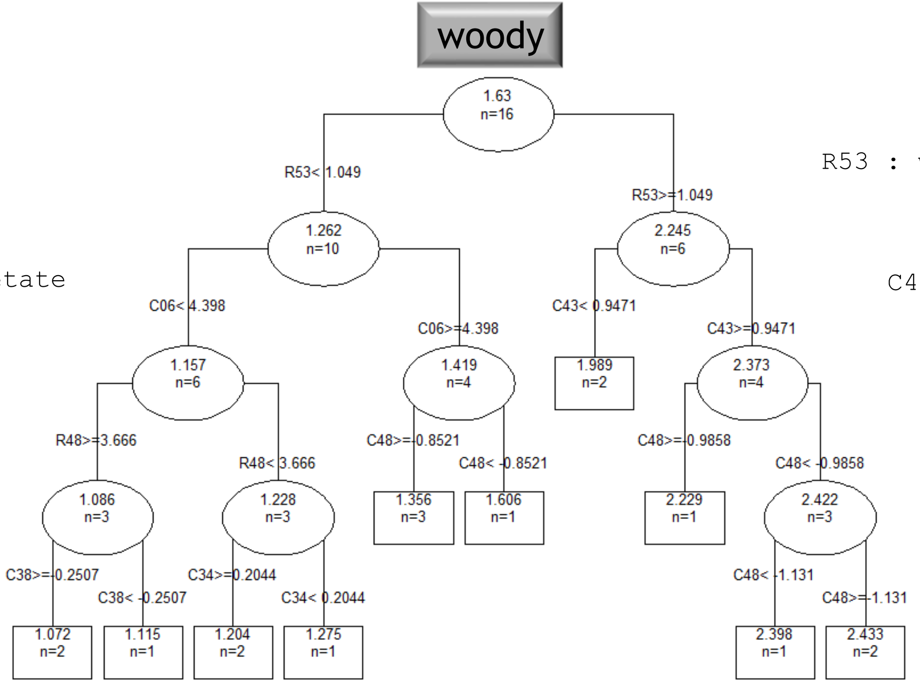


METHOD :Classification and Regression Trees (CART)
How ?



Breiman L., Friedman J., Olshen R., Stone C. (1984). Classification And Regression Trees. Chapman & Hall

METHOD :Classification and Regression Trees (CART) Illustration



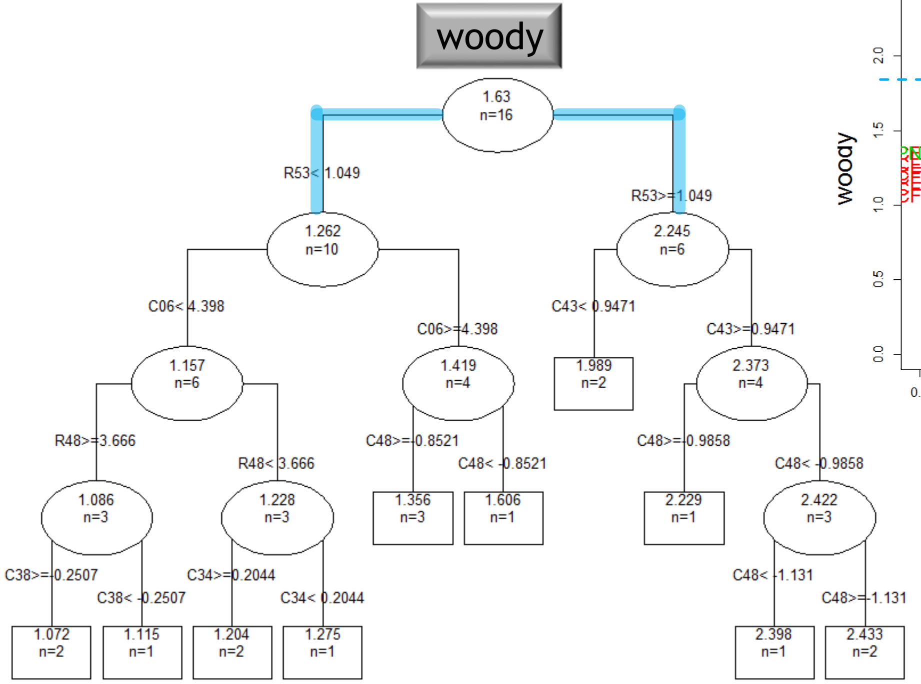
C06 : Ethyl acetate

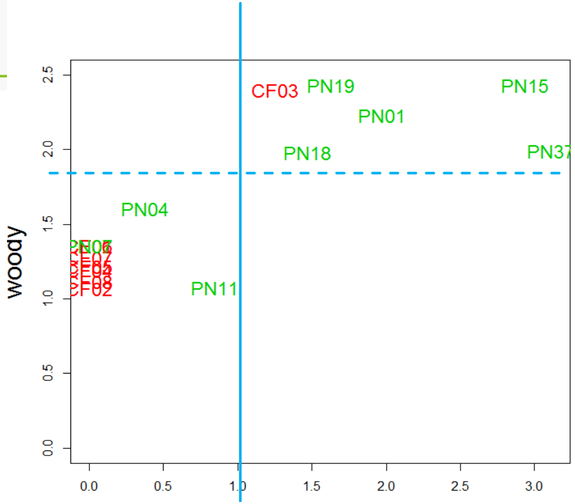
R53 : whyskeylactone

C43 : phenol

Here - uniform branch length

METHOD :Classification and Regression Trees (CART) Illustration



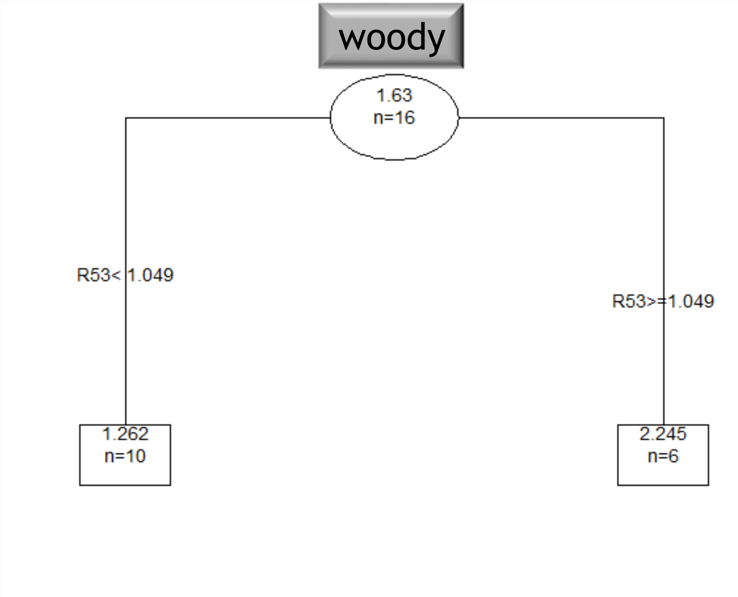


R53 : whyskeylactone

METHOD :Classification and Regression Trees (CART)

Pruning

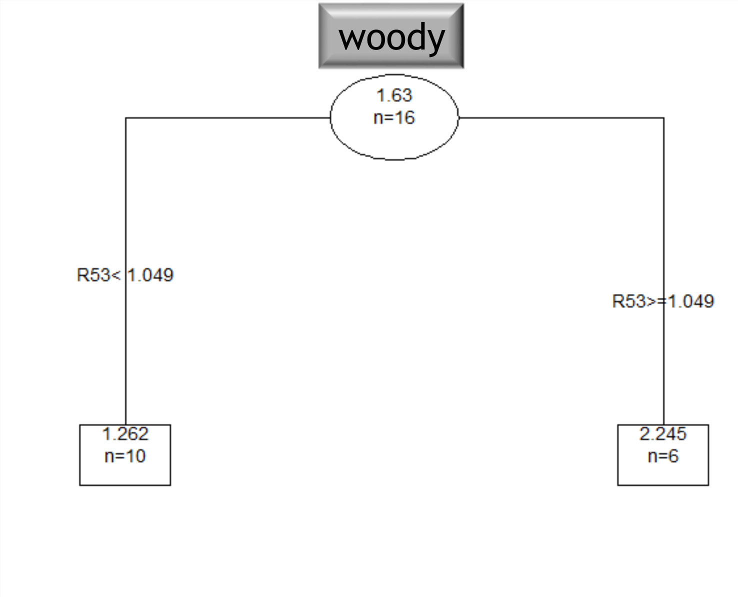
based on Cross-Validation (here LOO)



METHOD :Classification and Regression Trees (CART)

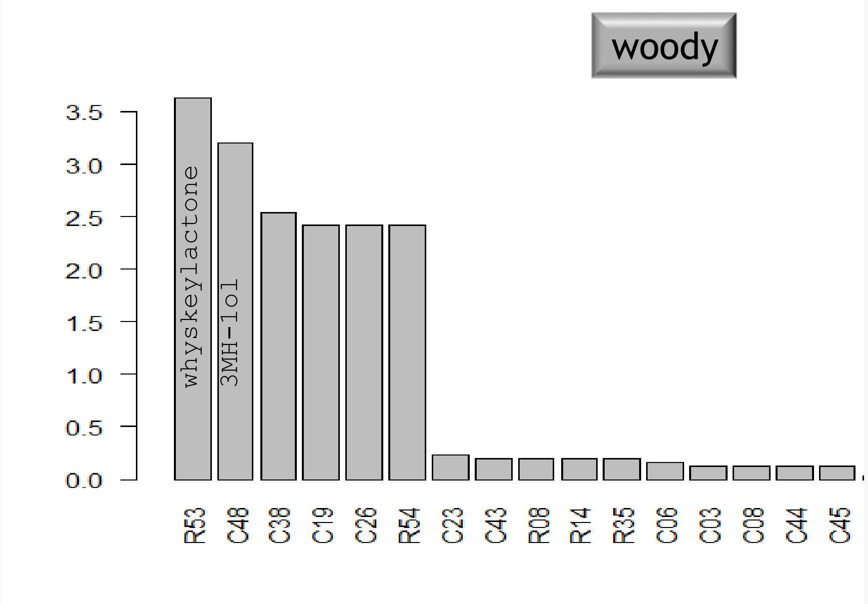
Pruning

based on Cross-Validation (here LOO)



Surrogate variables

impurity decrease allowed by each variable



METHOD :Classification and Regression Trees (CART)

Why ?

- Output providing decision rules,
- Non linear links can be handled,
- No distributional hypotheses
- …

However

when there are correlated variables,

- caution must be taken in terms of interpretation
- lack of robustness of the obtained tree.

METHOD :Classification and Regression Trees (CART)

Why ?

- Output providing decision rules,
- Non linear links can be handled,
- No distributional hypotheses
- …

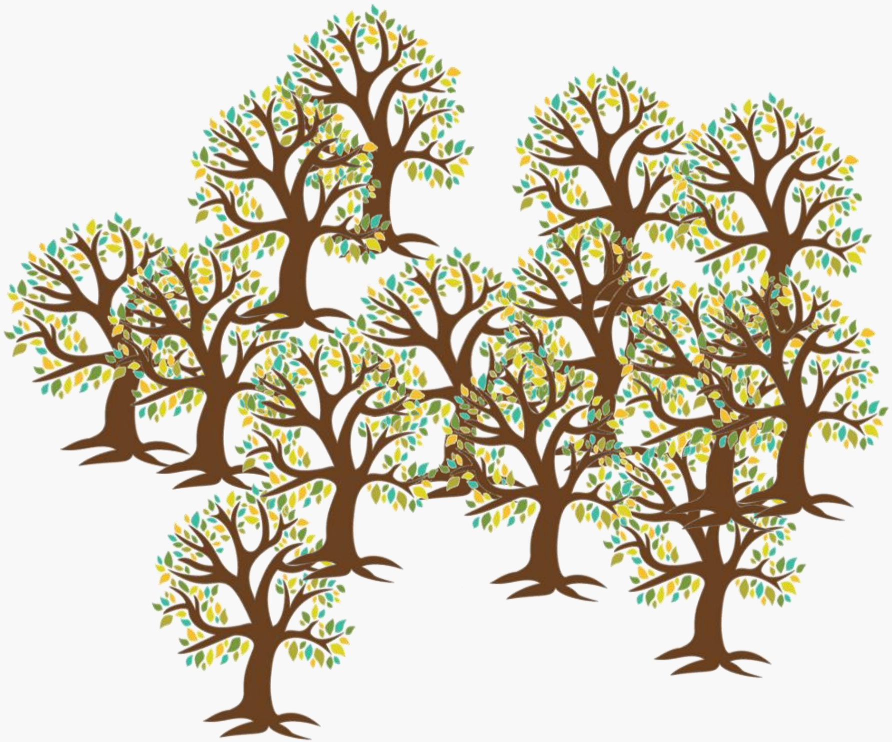
However

when there are correlated variables,

- caution must be taken in terms of interpretation
- lack of robustness of the obtained tree.

construct a lot of trees with randomization :
Random forests

METHOD: Random Forests



Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32

- Randomization of the observations (bootstrap samples / bagging)
- Random selection of variables at each node (*mtry* variables selected at random at each node)

Determination of the Variable Importance (VI) of each variable (Breiman, 2001) based on the mean increase of the error obtained with a tree after permutation of the values for the Out-Of-Bag samples.

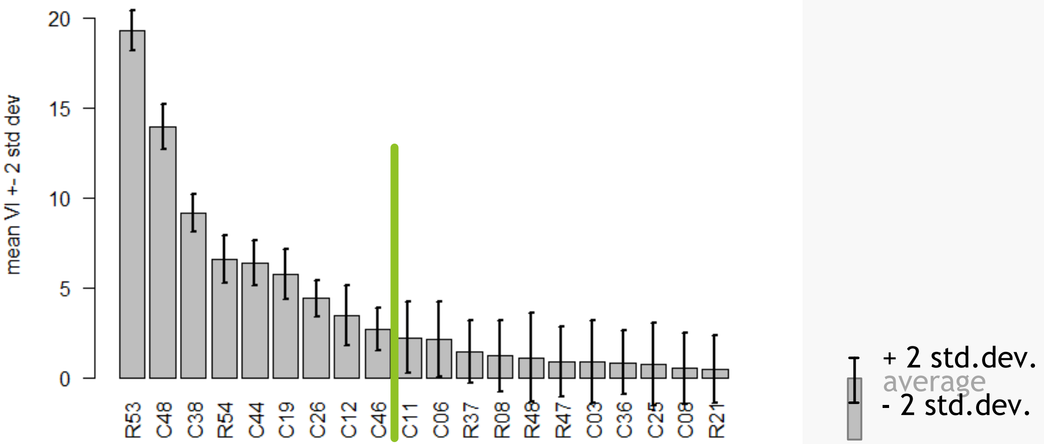
Strategy of analysis / Step ➊

- Forest with 2000 trees *(ntree=2000, mtry=p/3)*
- Repetition of the procedure => 50 forests are built

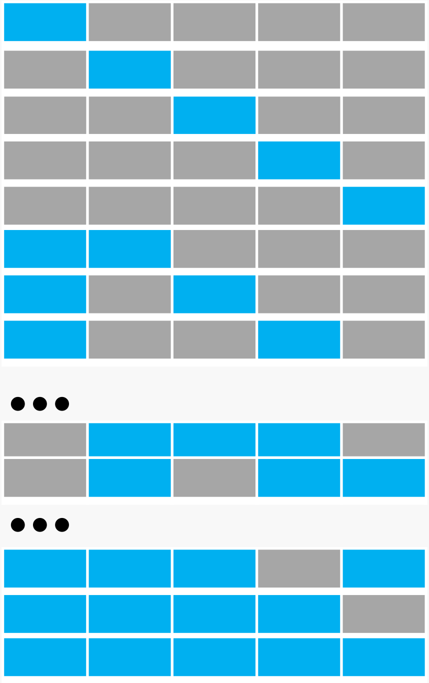
Ranking of the compounds according to their Variable importance (VI)

(here the first 20 are shown)

woody



Strategy of analysis / Step ❷



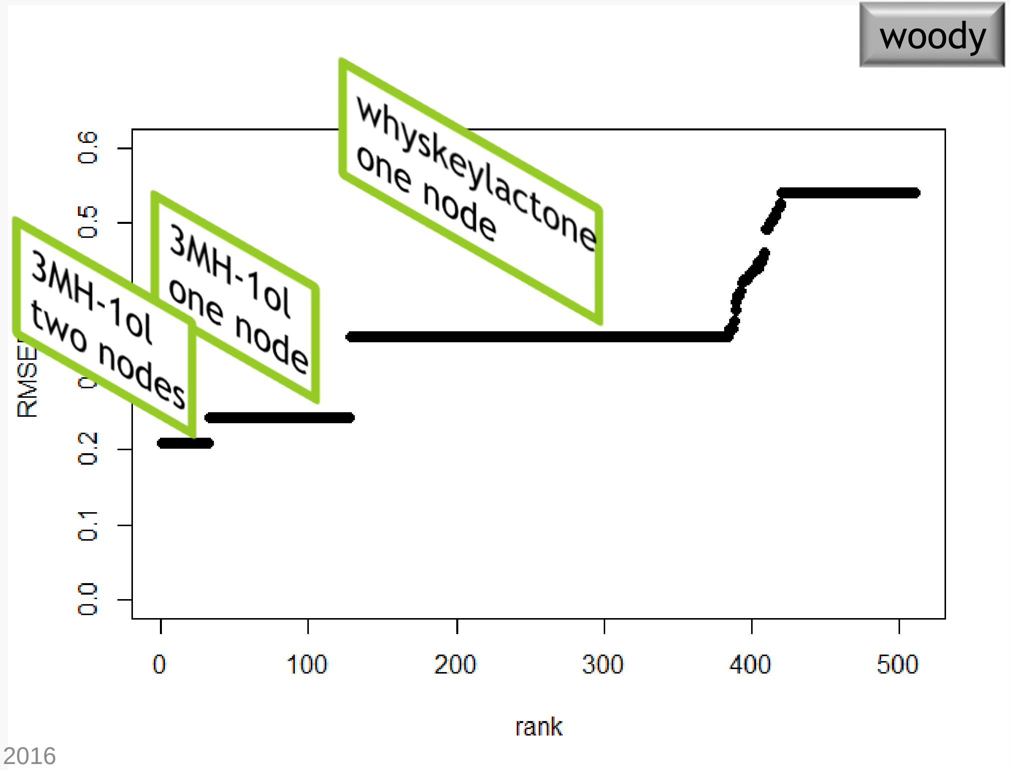
For each subset "s" of 1, 2, 3…compounds among pre-selected compounds

- Construction of the regression tree for "s"
- Pruning based of LOO cross-validation
- $RMSEP_{LOO,s}$

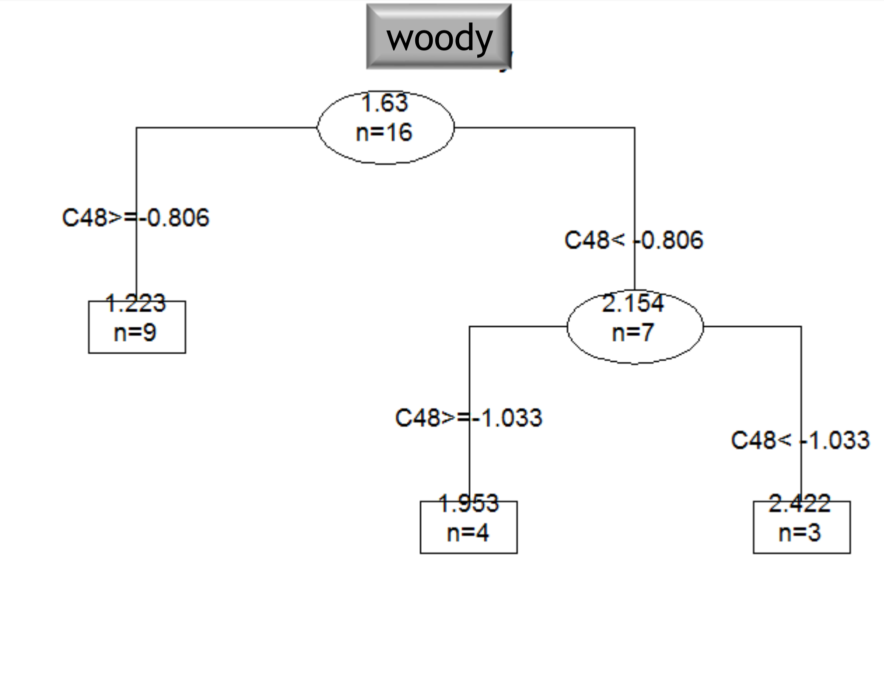
Strategy of analysis / Step ❷

\# pre-selected compounds = 9
⇨ 511 pruned regression trees

The $RMSEP_{LOO}$ are sorted

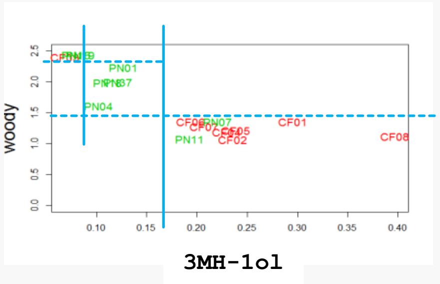


RESULTS

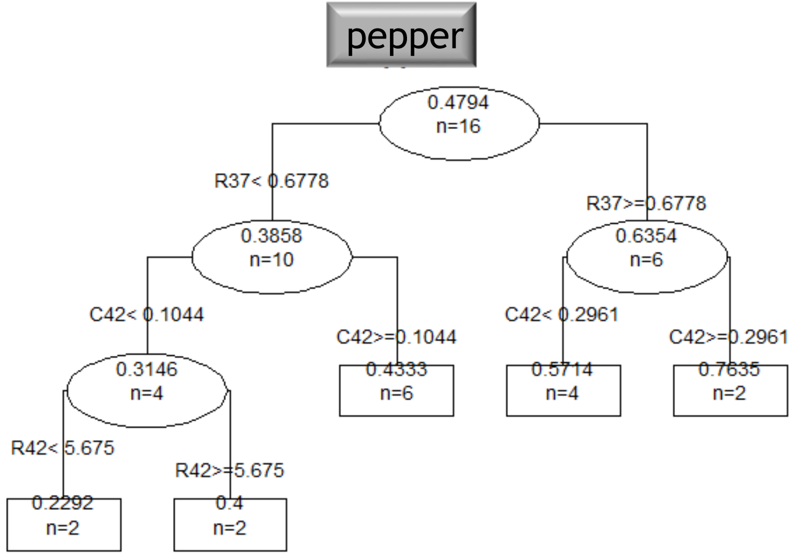


Decision rule

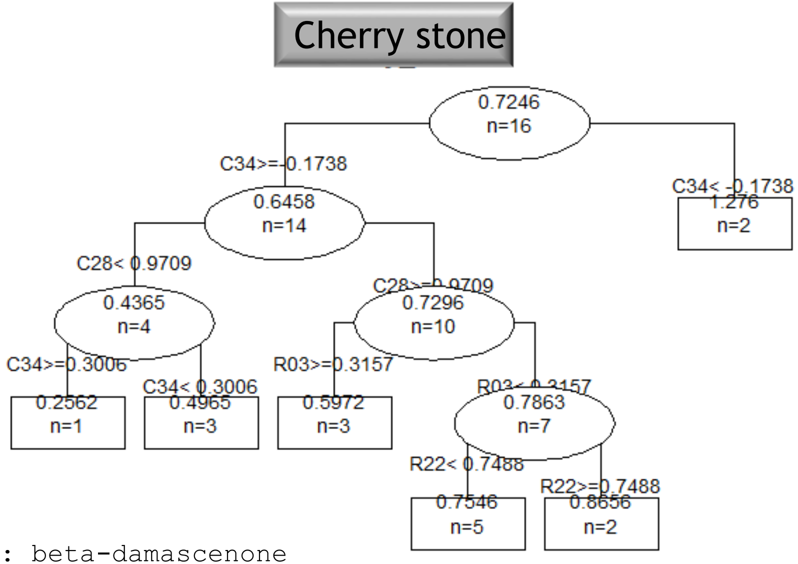
- If `C48=log10(3MH-1ol)>-0.806`
  i.e. `3MH-1ol>0.16`
  ⇨ The woody odor should be rather low (expected value= `1.22`)
- If `0.09<3MH-1ol<0.16`
  ⇨ The woody odor should be intermediate (expected value= `1.95`)
- If `3MH-1ol<0.09`
  ⇨ The woody odor should be rather high (expected value= `2.42`)



RESULTS: Others examples



R37: 3-isobutyl-2-methopyrazine
C42: 1-phenoxy-2-propanol
R42: isovaleric acid



C34: beta-damascenone
C28: 1-octanol
R03: acetaldehyde
R22: not identified

RESULTS: Comparison of the prediction ability with PLS regression

Sensory attribute	method	# of compounds in model	model complexity	$RMSEP_{LOO}$
Woody	Reg. Tree	1	2 nodes	0.208
	PLS	115	2 PLS comp.	0.438
	PLS-VIP (>1.5)	6	2 PLS comp.	0.253
Pepper	Reg. Tree	3	4 nodes	0.093
	PLS	115	1 PLS comp.	0.107
	PLS-VIP (>1.5)	13	1 PLS comp.	0.077
Cherry stone	Reg. Tree	4	5 nodes	0.157
	PLS	115	3 PLS comp.	0.212
	PLS-VIP (>1.5)	12	4 PLS comp.	0.066

To conclude

Regression trees

Interpretation

- model easy to interpret,
- providing a set of decision rules.
- parsimonious (in our case study),
- able to handle non-linear relationships and interaction between the predictors.

Prediction

Satisfactory prediction ability.

Thank you for your attention

Some references on CART and Random Forests

Breiman L., Friedman J., Olshen R., Stone C. (1984). Classification And Regression Trees. Chapman & Hall

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32

Genuer R. Poggi , J.-M. , Tuleau-Malot C. (2010). Variable Selection using Random Forests. *Pattern Recognition Letters*, 31 (14), 2225-2236.

Romano R., Davino C., Næs T. (2014). Classification trees in consumer studies for combining both product attributes and consumer preferences with additional consumer characteristics. *Food Quality and Preference*, 33, 27–36.