Recursive partitioning methods for identifying relevant variables in a sparse unbalanced data set

Orlane Briand [1], Ivanne Debec [1], Arnaud Montet [2], Lorraine de Malleray [2] & Evelyne Vigneau [2]

1 Sensometrics and Chemometrics Laboratory, ONIRIS, La Géraudière, Nantes

2 IFF, Neuilly

Introduction

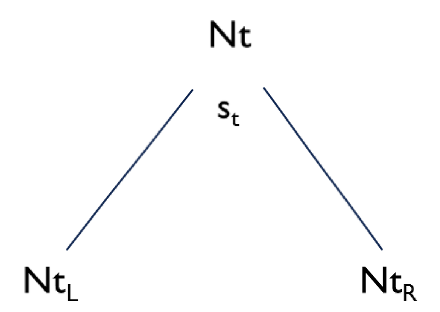
The aim of this study was to explore various recursive partitioning methods when dealing with a sparse and unbalanced data set.

→ It is proposed to use Classification and Regression Trees (CART) [1], Conditional Inference trees (CI trees) [4] and Random Forests (RF) [2]. Besides their predictive ability, these methods are easy to interpret, providing and efficient way to identify relevant variables.

→ In this case study, a quantitative quality response *y*, was to be related to a large number of quantitative predictors (*X* matrix), most of them being sparse (with zero values). Moreover, about one third of these predictors included only one non-null observation, giving rise to a very unbalanced dataset.

CART

The CART algorithm partition the initial subset of observations at each node into two groups in order to maximize a measure related to the variation of the node impurity.



Maximizes :

$$\Delta i(t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Where $p_L = N_{tR}/N_t$, and $p_R = N_{tR}/N_t$

CI trees

In order to overcome the bias selection problem known with CART.

At each node :

► **Step 1 :**

The association of each predictor to the response is assessed by a permutation test framework.

The predictor showing the strongest relationship to the response (lowest *p-value*) is chosen. If none of them reach the predefined significance level, the actual node is not further split.

► **Step 2 :**

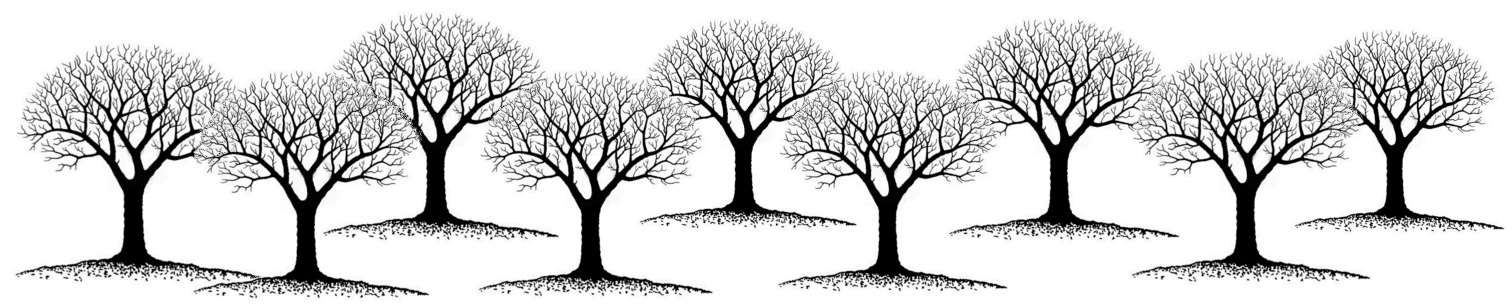
Choice of splitting threshold.

Random Forests

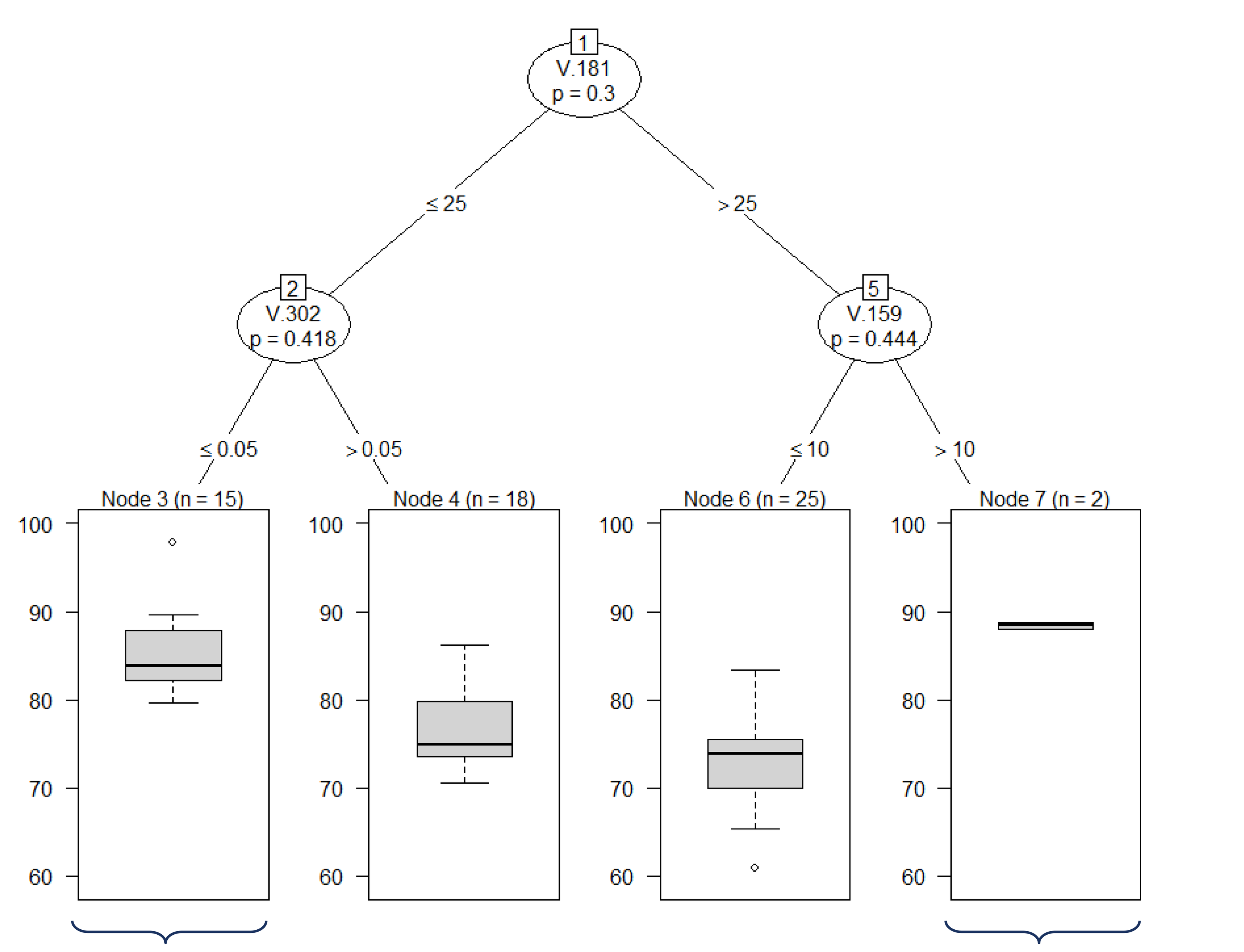
Random Forests are collections of trees (CART or CI trees) for more robustness.

► CART favors splits in continuous variables and variables with numerous categories [5].

Caution : In our case study, this is to be considered because, even if the variables *X* are all quantitative, the more they are sparse, the less there are choices in the cut-off points.



CI tree



► Relatively high level of response when :
V.181 ≤ 25 and V.302 ≤ 0.05

► High level of response when :
V.181 > 25 and V.159 > 10

Variable importance measures

► Useful tool for ranking.

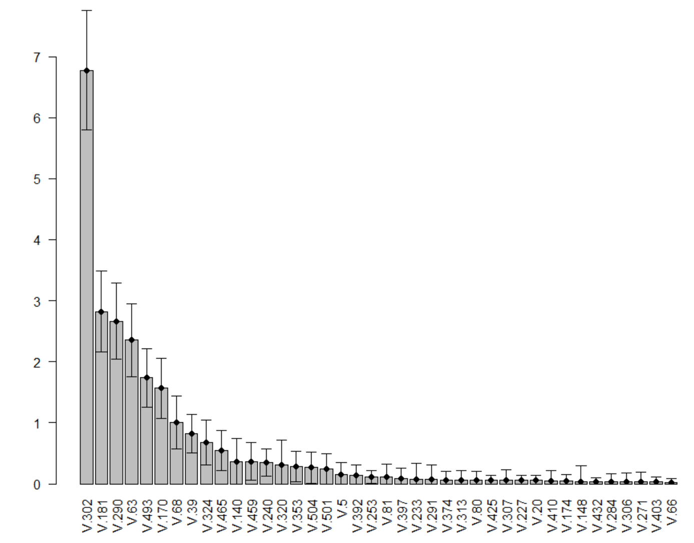
► Most common criterion : **Mean Decrease in Accuracy** (MDA).

► The three types give almost the same ranking of variables.

→ **by permutation**

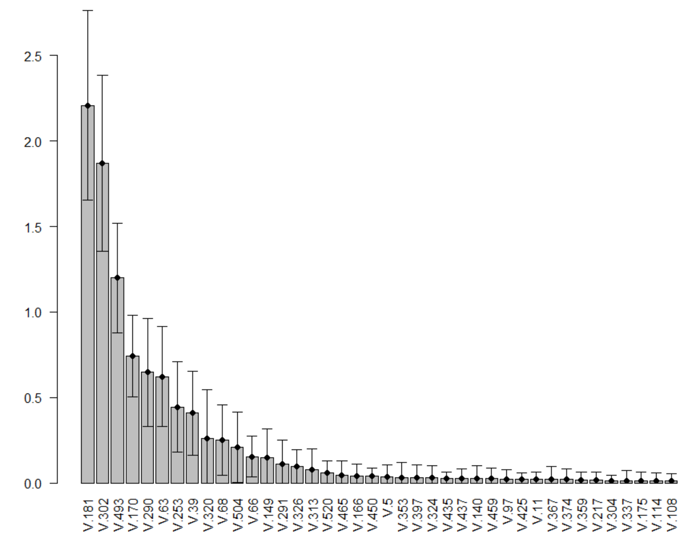
● **"MDA-CART"**

Determined by permuting the values of each variable and measuring how much the permutation decreases the accuracy of the model.



● **"MDA-CI-perm"**

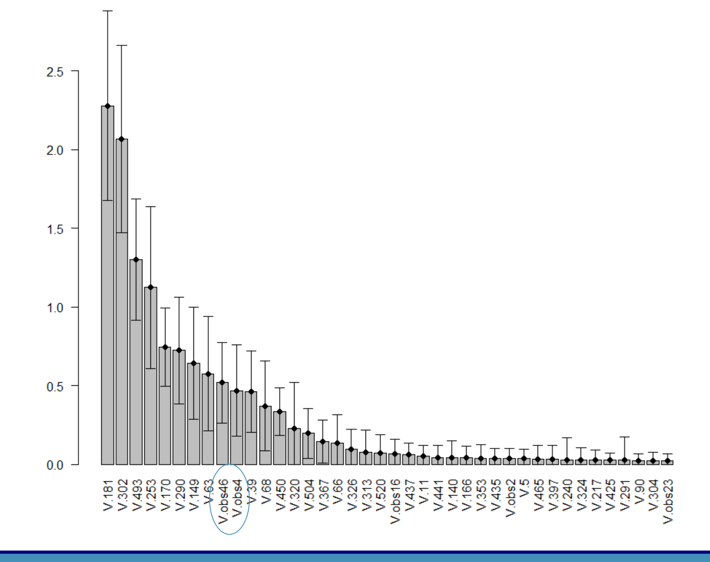
Evaluated following the permutation principle of the MDA importance in 'RandomForest' but based on CI trees, instead of CART trees.



→ **by random allocation**

● **"MDA-CI-rdalloc"**

Each observation is randomly allocated to the child nodes if the split of their parent node is conducted in the variable of interest.



Conclusion

→ Two frameworks of procedures were proposed in order to solve the variable selection problems caused by the sparse and unbalanced data set. The conditional inference trees [4] seem to be an appropriate solution to this kind of regression problems.

→ In the Conditional Inference trees framework, the "MDA-CI-rdalloc" measure provides an unbiased variable selection and allows variables with only one non-null value to have a significant measure of importance.

References

[1] Breiman, L., Friedman, J. H., Olshen R.A. & Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

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

[3] Hapfelmeier, A., Hothorn, T., Ulm, K. & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Stat Comput*, 24:21-34.

[4] Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased Recursive Partitioning: A conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

[5] Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. (2007). *Biais in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics*, 1-21.