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) [2] and Random Forests (RF) [3]. 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 partitions 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:

\[
\Delta l(i) = l(l) - p_l l(l) - p_{\text{null}} l(n)
\]

Where \( p_l = N_0 / N_t \), and \( p_{\text{null}} = N_{\text{null}} / N_t \)

### Cl trees

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

At each node:

1. **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.

2. **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 levels.
- CI trees favor splits in categorical variables and continuous variables with few levels.

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.

### 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-rdaloc”
  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-rdaloc” measure provides an unbiased variable selection and allows variables with only one non-null value to have a significant measure of importance.

### References


