An alternative to preference mapping: Regression trees and Random forests. Application to preferences for ciders

Léa Koenig[1], Adèle Lepetit[1], Ronan Symoneaux[2] & Philippe Courcoux[1]

[1] *ONIRIS, Sensometrics and Chemiometrics Lab. BP43222 Nantes, France*
E-mail : *lea.koenig@oniris-nantes.fr ; adele.lepetit@oniris-nantes.fr ; philippe.courcoux@oniris-nantes.fr*

[2] *Groupe ESA, UR GRAPPE – Angers – France*
E-Mail : *r.symoneaux@groupe-esa.com*

Abstract
Regression trees and Random forests are recursive partitioning methods which allow explaining a response from a set of explanatory variables. In the context of a preference study, these techniques are evaluated for the identification of drivers of liking and the prediction of preference.
Keywords: multivariate regression tree, random forest, drivers of liking, cider.

Résumé en Français
Les arbres de régression et les forêts aléatoires sont des méthodes de partitionnement récursif qui permettent de modéliser une réponse soumise à un ensemble de variables explicatives. Dans le contexte d'une étude hédonique, ces techniques sont utilisées pour la sélection d'attributs sensoriels et la modélisation de la préférence.
Mots-clés : arbre de régression multivarié, forêt aléatoire, préférence, cidres

1. Introduction

The aim of this study is to evaluate the use of regression trees and random forests for modelling preference data as an alternative method to preference mapping.
The goal is to model and predict preference for products from sensory attributes.

2. Methodology

2.1 Univariate regression trees

The CART (Classification and Regression Trees) methodology (Breiman, 1984) is used to explain and predict a response variable as a function of several explanatory variables. CART produces a decision tree allowing the modelization of the response as a set of decision rules. Each node involves a single explanatory variable and a cutoff point. In the context of a

hedonic study, the preference is explained by a subset of sensory attributes and the tree allows describing the effects of attributes on the rating of acceptability (Romano, 2014).

2.2 Random forests

The Random Forest methods (Breiman, 2001), a complementary method to CART, are used to determine the importance of the different explanatory variables in order to select only the most important ones. A succession of regression trees are built using a double randomization (bootstrapping on products and randomized selection of variables). This selection of attributes is used for eliciting the main drivers of liking.

2.3 Multivariate regression trees

Multivariate regression trees are an extension of univariate regression trees for multivariate responses (De'ath, 2002). For a consumer study, the hedonic scores given by the different subjects of the panel are described by a single tree. At each node, the variable is chosen as the most discriminant for all consumers.

3. Application

3.1 Presentation of the data

Data came from a CASDAR project on the preference for ciders funded by the French Ministry of Agriculture. This project involved ESA Angers, Adria of Normandie, INRA and IFPC. A sensory profile was performed on 19 ciders (different origins) by a panel of 10 trained subjects on 47 attributes (aroma, smell and color). On the same ciders, a hedonic study was realized by 341 consumers on an 11 points scale.

3.2 Selection of drivers of liking using random forest

Random forests were used for assessing the importance of the sensory attributes on the mean liking score of the panel. In order to measure the variability of these importances, a bootstrapping of subjects is performed and allows giving a standard deviation of the importance of each attribute. This allows selectingthe most important ones. Importances of the sensory attributes are presented on figure 1:

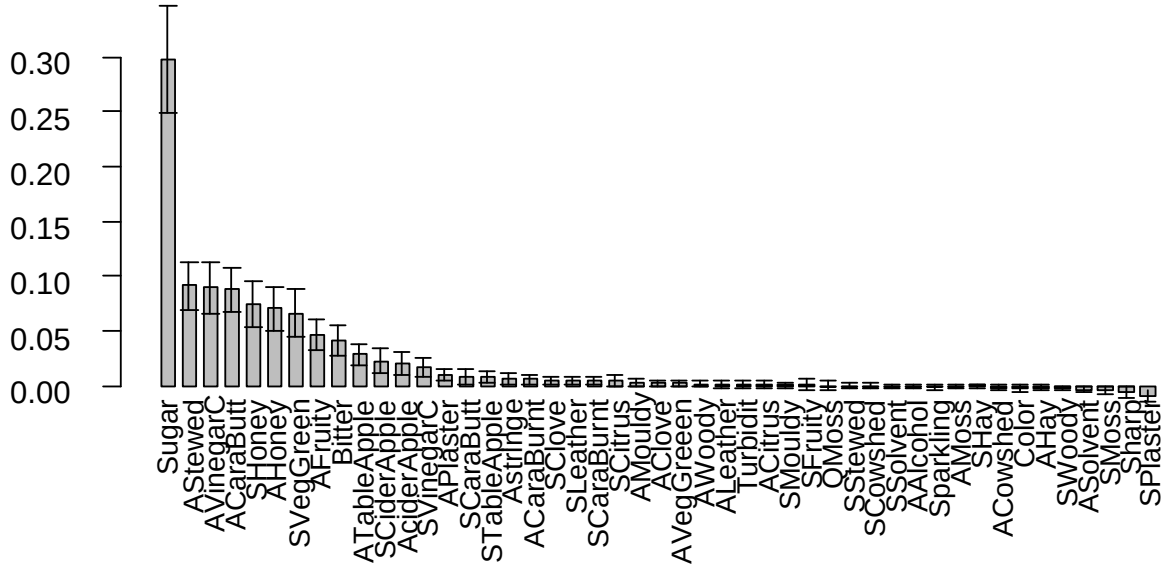

Figure 1: Importance of the variables

The 19 first variables were selected and used for building regression trees.

3.2 Multivariate regression trees

As the response is multivariate (different consumers), multivariate regression trees are adapted for describing the individual preferences.
The regression tree obtained for the whole panel is presented on figure 2:

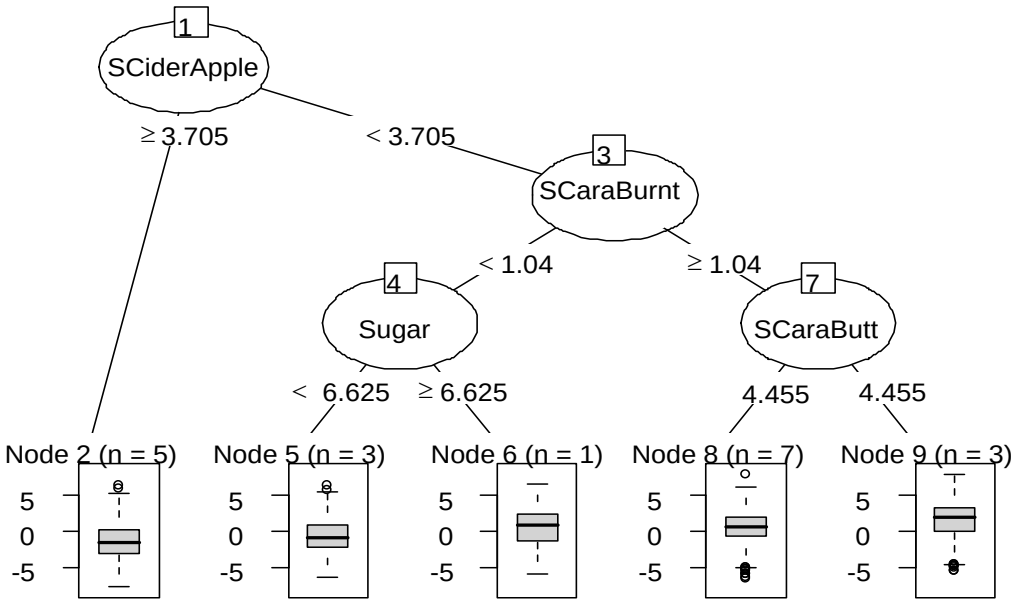

Figure 2: Multivariate regression tree

The variable which best discriminate the preference for the 19 ciders is the smell of cider apple with a cutoff point of 3.705. Boxplots represent the mean and the dispersion of the preference of the 341 consumers. Clearly, the mean preference is increasing from the left to the right of the tree.

In this study (not shown here), the same methodology was used after a segmentation of the panel in three different classes of preference. This allows to highlight the drivers of liking in each class of consumers.

4. Conclusion

The use of regression trees is clearly suitable in the context of a preference study. Random forests give a way to select reliable attributes for modeling preference. In addition, multivariate trees allows an easy understanding and interpretation of the preference of a panel. It also can deal with non-linearity.

References

Breiman L. (1984). Classification and Regression Trees. Taylor & Francis, 368p.

Breiman L. (2001). Random Forest. *Machine Learning*. Vol 45, pp.5-32

De'ath G. (2002). Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. *Ecology*, Vol 83, No 4, pp 1105-1117

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