Food Quality and Preference 68 (2018) 135–145

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

# Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception

E. Vigneau[a,*], P. Courcoux[a], R. Symoneaux[b], L. Guérin[c], A. Villière[d]

[a] StatSC, Oniris, Univ Bretagne Loire, INRA, 44322 Nantes, France
[b] USC 1422 GRAPPE, INRA, Ecole Supérieure d'Agricultures, Univ. Bretagne Loire, SFR 4207 QUASAV, 49100 Angers, France
[c] Institut Français de la Vigne et du Vin, 37400 Amboise, France
[d] GEPEA, Flavor group, Oniris, Univ. Bretagne Loire, CNRS, 44322 Nantes, France

ARTICLE INFO

*Keywords:*
Random forest
CART
Olfactory perception
Volatile organic compounds
Wine

ABSTRACT

The purpose of this paper is to discuss the application of the Random Forest methodology to sensory analysis. A methodological point of view is mainly adopted to describe as simply as possible the construction of binary decision trees and, more precisely, Classification and Regression Trees (CART), as well as the generation of an ensemble of trees or, in other words, a Random Forest. The interest of the permutation accuracy criterion, as a measure of variable importance, is specifically emphasized as a way of identifying the most predictive variables and selecting a subset of these variables for parsimonious and efficient predictive models. A two-step procedure is proposed for choosing this subset of variables. The principle of the method is illustrated in a case study in which the aim was to better understand and predict the olfactory characteristics of red wines made of the Cabernet Franc grape variety, from their Volatile Organic Compound (VOC) content. For two main olfactory attributes, the bell pepper odor and the leather odor, it was possible to list the most important compounds and to highlight a very small number of compounds useful for estimating each of the olfactory attributes considered. For the latter, it was also observed that Random Forest models had a better predictive ability than Partial Least Squares (PLS) Regression models.

## 1. Introduction

Regression models are commonly used for linking a response variable to a set of predictors, usually with two aims: explanation and prediction. Methods such as neural networks or support vector machines are known to be very efficient for prediction but appear as "black box" systems and are unsuccessful for interpretation. Yet, identifying the relevant predictors and understanding the underlying mechanism are of interest in many applications. Using (generalized) linear methods, the coefficient estimates are used to judge the relevance of the predictors. However, in order to deal efficiently with the multi-collinearity between the predictors, complementary strategies (*e.g.* stepwise procedures) or specific statistical methods, such as Partial Least Squares (PLS) regression, Least Absolute Selection and Shrinkage Operator (LASSO) model (Tibshirani, 1996) or Elastic Net Regression (Zou and Hastie, 2005) may be required. Another family of models, stemming from machine learning methodologies, has also been developed. These are based on the recursive partitioning of the dataset, as in Classification and Regression Trees (CART) introduced by Breiman, Friedman, Olshen, and Stone (1984). For more efficiency, individual trees can be combined into an ensemble of trees or forest. The Random Forest (RF) approach suggested by Breiman (2001) is the most popular algorithm for the construction of an ensemble of CART. Classification and Regression Trees can be used for classification (in the case of a categorical response variable) or regression (when the response variable is quantitative). This article focuses on regression. In the field of sensory analysis, this corresponds to a large number of situations in which, for instance, the response variable can be a sensory attribute (to be related to instrumental or chemical predictors), or a hedonic score (as in external preference mapping).

In the last decade, RF and CART have been extensively used in multiple applications for microarray, epidemiological, medical or psychological data, mainly for classification problems. In food science, and more specifically in studies on wine, the RF methodology has successfully been applied by Brillante et al. (2015) to predict the skin flavonoid content in wine grapes from berry physico-chemical characteristics and by Gómez-Meire, Campos, Falqué, Díaz, and Fdez-Riverola (2014) to classify six grape origins of Galician (Spain) white wine taking account of their volatile compounds profiles. In sensory and preference consumer data analysis, CART and RF are seldom used. To our knowledge,

* Corresponding author.
*E-mail address:* evelyne.vigneau@oniris-nantes.fr (E. Vigneau).

https://doi.org/10.1016/j.foodqual.2018.02.008
Received 22 May 2017; Received in revised form 19 February 2018; Accepted 21 February 2018

RF was first applied to sensory analysis by Granitto, Gasperi, Biasioli, Trainotti, and Furlanello (2007) in the classification of six varieties of locally-made cheeses from their sensory attributes. Bi and Chung (2011) and Bi (2012) pointed out that the variable-importance measure based on the RF model can be considered as a valid measure to identify drivers of liking even in presence of multicollinearity among the sensory attributes. Romano, Davino, and Naes (2014) compared the CART approach with PLS regression for relating segments of consumers obtained from a liking experiment to consumer characteristics including attitudes, habits and demographic dependent variables. More recently, Kuesten, Dang, Nakagawa, Bi, and Meiselman (2016) adopted a strategy based on RF importance measure (Yao et al., 2013) in order to make a selection among many psychographic scales and other consumer data used as covariates in a Propensity Score Analysis (PSA) for causal inference of outcomes such as consumers' overall liking. In our case study, the aim is to relate and predict the sensory olfactory characteristics of wines made of the Cabernet Franc variety to their Volatile Organic Compound (VOC) patterns. In this regard, this study was intended to contribute to the general topic of correlating sensory measurements to instrumental measurements, an old and fundamental problem in the sensory field (Powers and Moskowitz, 1976).

The main reasons for investigating alternative approaches to the ordinary least squares method or PLS regression are (a) the fact that recursive partitioning approaches are not restricted to a predetermined functional form of the relationship between the response variable (dependent variable) and the explanatory variables (independent variables or predictors), (b) the ability of the tree to take account of potential complex interactions between the predictors, (c) the ability to give a simple interpretation of a tree as a decision rule, instead of linear combinations of a large number of predictors. In addition, RF provides a statistic, the variable importance (V.I.), which is very useful for assessing the significance of each explanatory variable in the model. This metric, also called the permutation accuracy importance measure, has been used as a variable selection criterion and discussed by several authors, as by Liaw and Wiener (2002), Strobl, Boulesteix, Zeileis, and Hothorn (2007), Archer and Kimes (2008), Grömping (2009), Genuer, Poggi, and Tuleau-Malot (2010) and Bi (2012). One of its key advantages is that it takes into account each predictor individually as well as in multivariate interactions with other predictors, even when the number of variables is much larger than the number of samples, which is a common situation in sensory analysis (Kuesten et al., 2016).

In the case study considered herein, the aim is to illustrate the interest of RF in a sensory experiment to identify the most relevant predictors, from a large number of correlated quantitative ones (i.e. VOCs), to explain a sensory olfactory attribute of wines. For this purpose, the ranking of the predictors based on the V.I. metric is specifically addressed. The interpretability of recursive splitting models is also highlighted, as well as their predictive ability compared to the PLS regression approach.

## 2. Material and methods

### 2.1. Wine samples

In order to illustrate the CART and RF approaches, a study of red wines of the Cabernet Franc variety is considered, which was part of the INNOVAROMA research program (with support from the Pays de la Loire Region, France).

Wines were produced in the Loire Valley area (France) by IFV (Institut Français de la Vigne et du Vin). Different land parcels located in the vineyards of five Protected Geographical Indications (Bourgueil, St Nicolas de Bourgueil, Chinon, Anjou and Saumur) were chosen. For each harvested parcel, the same wine making process was applied. Twenty wines produced in 2011 and ten in 2012 were considered.

The wines samples were assessed, on the one hand, by a sensory trained panel regarding their olfactory attributes and, on the other hand, by an analytical laboratory for the quantification of a range of VOCs.

### 2.2. Sensory analysis

For each vintage, the set of wines was subjected to sensory evaluation by a trained panel, at UR GRAPPE, Angers (France). The panel consisted of the same 14 panellists in both years. The panellists were asked to assess the odorant characteristics of the wines, perceived directly in an orthonasal way, in two replicates, during five and three sessions for the first and second set of wines, respectively. Wines were presented to subjects according to a William Latin-square arrangement.

The wines were stored in a climate-controlled dark cellar maintained at 11 °C ( ± 1 °C). The day before the session, the wine bottles were stored at room temperature (20 °C ± 1 °C). The day of the session, they were poured into dark INAO-approved wine glasses (30 mL) and the panel leader checked that the wines were free of cork taint.

The panel evaluated 33 olfactory sensory attributes obtained from a previous experiment about the sensory characterization of Cabernet Franc and Gamay wines (Coulon-Leroy, Symoneaux, Lawrence, Mehinagic, & Maitre, 2017) with some added attributes generated by consensus with the panellists at the beginning of the study. Scores for all the attributes were collected on a 14-cm linear scale by FIZZ (version 2.10; Biosystems, Courtenon, France) and converted into scores from 0 to 10. One of the specific features of these sensory data is the large number of rating scores equal to zero. In fact, with an extended list of olfactory attributes, elicited for a whole description of the wine odorant complexity, it is not uncommon that a given attribute is not perceived by one or several assessors, for a given sample. Only the sensory attributes with a citation percentage (percentage of non-null scores, over all the assessors, samples and repetitions) higher than 15% were retained. This led to a list of fifteen orthonasal sensory attributes: "Leather", "Musc", "Black Pepper", "Cherry Stone", "Blackcurrant", "Cherry_fresh", "Cherry_cooked", "Strawberry_fresh", "Blackberry", "Prune", "Cut Grass", "Bell Pepper", "Woody", "Hay" and "Ethanol".

### 2.3. Volatile organic compounds

Gas Chromatography (DB-FFAP column, 30 m × 0.32 mm × 0.25 µm) coupled with Mass Spectroscopy (GC-MS) analysis was performed by a specialized company after solid phase micro-extraction (SPME, fiber CAR-DVB-PDMS, Supelco, Bellefonte, PA, USA; extraction 60 min at 45 °C) in order to quantify thirty-eight VOCs in each wine (Table 1).

The choice of these compounds was based on wine knowledge (Ebeler, 2001; Ferreira, 2010) as well as on the specific features of wines made from the Cabernet grape variety (Jiang, Xi, Luo, & Zhang, 2013; Synos, Reynolds, & Bowen, 2015). Compounds were identified by comparison of their mass spectra to those of standard databases (Wiley, NIST, internal database) and injection of standard reference compounds. The quantification of compounds was based on a calibration method using these reference compounds.

### 2.4. Statistical method

In order to understand better and predict the olfactory sensory characteristics of the products in the light of their VOCs, the Random Forest (RF) method was investigated. RF takes advantage of aggregating multiple different CART models from a single training dataset.

In this work, CART and RF were implemented using the R packages *rpart* (Therneau et al., 2015) for construction of the CART trees, *partykit* (Hothorn and Zeileis, 2016) for their graphical representation and *randomForest* (Liaw and Wiener, 2015) for RF.