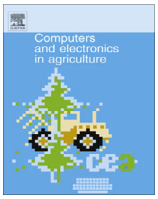
Computers and Electronics in Agriculture 117 (2015) 186–193

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes

Luca Brillante [a,*], Federica Gaiotti [a], Lorenzo Lovat [a], Simone Vincenzi [b], Simone Giacosa [c], Fabrizio Torchio [c], Susana Río Segade [c], Luca Rolle [c], Diego Tomasi [a]

[a] CRA-VIT Council for Agricultural Research and Economics, Viticulture Research Center, Conegliano, TV, Italy
[b] University of Padova, Centro Interdipartimentale per la Ricerca in Viticoltura ed Enologia, Legnaro, PD, Italy
[c] University of Turin, Dipartimento di Scienze Agrarie, Forestali e Alimentari, Grugliasco, TO, Italy

ARTICLE INFO

Article history:
Received 14 May 2015
Received in revised form 1 July 2015
Accepted 5 July 2015
Available online 24 August 2015

Keywords:
Random forest
Gradient Boosting Machine (GBM)
Wine-grape
Flavonoids
Texture analysis

ABSTRACT

Flavonoids are a class of bioactive compounds largely represented in grapevine and wine. They also affect the sensory quality of fruits and vegetables, and derived products. Methods available for flavonoid measurement are time-consuming, thus a rapid and cost-effective determination of these compounds is an important research objective. This work tests if applying machine learning techniques to texture analysis data allows to reach good performances for flavonoid estimation in grape berries.

Whole berry and skin texture analysis was applied to berries from 22 red wine grape cultivars and linked to the total flavonoid content. Three machine-learning techniques (regression tree, random forest and gradient boosting machine) were then applied. Models reached a high accuracy both in the external and internal validation. The R^2 ranged from 0.75 to 0.85 for the external validation and from 0.65 to 0.75 for the internal validation, while RMSE (Root Mean Square Error) went from 0.95 mg g^{-1} to 0.7 mg g^{-1} in the external validation and from 1.3 mg g^{-1} to 1.1 mg g^{-1} in the internal validation.

1. Introduction

Flavonoids are a group of secondary metabolites widely distributed in plants, which greatly affect the sensory and nutritional quality of fruits and vegetables (Harnly et al., 2006). They represent a huge portion of soluble phenols present in grapevine (Braidot et al., 2008). Flavonoids are among the most important compounds for the quality of red wine grapes because of their effect on wine sensory attributes (Ristic et al., 2010 is an example) and aging. The concentration of these compounds in wine depends, among other factors, on the quantity originally present in grapes (González-Neves et al., 2004). In the last ten years, flavonoids have received a very great attention from both researchers and the general audience because of their beneficial effect on human health (Yao et al., 2004). They have shown antioxidant (Lourenço et al., 2008), hypocholesterolemic (Gonzalez et al., 2015) and anti-inflammatory effects (Noll et al., 2009). Their nutraceutical properties are exploited in fresh table grapes, in pharmaceutical and cosmetic products derived from grape, and are a very appealing argument for wine marketing purposes.

Red grapes are richer in flavonoids than white ones, but their biosynthesis and concentration greatly depend on cultivar, vineyard practices, soil and climate (Koundouras et al., 2006). Grape maturity, and therefore the harvest date, is also another very important parameter because quantitative and qualitative modifications of tannins and anthocyanins (the two most represented flavonoid families in grape) happen during ripening (Kuhn et al., 2013).

Different methods based on spectrophotometry, chromatography, and mass spectrometry are usually used for the determination of flavonoids in fruits and vegetables (see Ignat et al., 2011 for a generic review and Lorrain et al., 2013 for the case of grapes and wine). Regarding grape analysis, these methods are all very accurate but they often require sample preparation and long analysis times. The problem is especially the time required for the extract preparation and purification, which has to be made by hand and can require berry peeling, solvent extractions, and

Abbreviations: BW, Berry Weight; CART, Classification And Regression Tree; Co, cohesiveness; Ch, chewiness; E_{sk}, skin young's modulus; F_{sk}, skin break force; *G*, gumminess; GBM, gradient boosting machine; *H*, hardness; *R*, resilience; RF, Random Forest; RFE, Recursive Feature Elimination; RT, Regression Tree; *S*, springiness; Sp_{sk}, skin thickness; SW, skin weight; W_{sk}, skin break energy; TA, Texture Analysis; TF, total flavonoids index.

[*] Corresponding author.
E-mail address: brillanteluca@live.it (L. Brillante).

http://dx.doi.org/10.1016/j.compag.2015.07.017
0168-1699/© 2015 Elsevier B.V. All rights reserved.

other manipulations that strongly increase costs and limit the number of acquirable data. Industry and research will greatly benefit from a rapid and cost effective method to obtain a faster screening of flavonoids in grapes. Such a method is at today lacking, although recently great advances have been made in this field by the use of Near InfraRed (NIR) spectroscopy coupled to chemometrics, in particular using partial least squares (PLS) regression models (Ferrer-Gallego et al., 2011; Rolle et al., 2012a; Cozzolino 2015).

During grape ripening, berries change not only their chemical composition, but also their mechanical properties: they soften, become less resilient, and the skin generally harden (Rolle et al., 2012b). In industry, these textural modifications are currently evaluated by sensory panels to help in the choice of the harvest date. Texture Analysis (TA) has shown to be an effective instrumental technique for an accurate evaluation of physical–mechanical characteristics of grapes (Letaief et al., 2008; Giordano et al., 2013; Battista et al., 2015). It is cost-effective as it does not require long times and reagents for sample preparation and analysis.

Although flavonoids and texture parameters belong to different grape properties, their values are both influenced by the berry ripening process. The phenolic ripeness of grape skin was found to be well assessed when the TA values were used (Río Segade et al., 2008), but the possibility of a predictive model has been never investigated, and neither an evaluation of possible chemometrics approaches to these parameters exists. A model linking the differences in berry mechanical properties and chemical composition induced by the grape ripeness could be an alternative to NIR methods for rapidly assessing the flavonoid contents at the berry level.

TA data are different from those obtained with NIR. In the first method, the number of measured parameters available as predictors is limited, and it is generally lower than the number of observations, i. e. the dataset is in a long format. Conversely, NIR datasets are wider, the number of wavelengths available as predictors is large and therefore PLS, a regression algorithm well suited to these situations, has been extensively applied (Cozzolino, 2015). With the reduced number of predictors present in TA, other learning algorithms could be effectively applied as an effort to better exploit the available information.

In this work, we will evaluate the use of regression trees and of two ways of combining them in order to achieve greater performances in predictions: Random Forest, RF (Breiman, 2001), and gradient boosting machine, GBM (Friedman, 2001). RF has shown to be a state-of-the art method, allowing the highest accuracy, but it is still not widespread to date. According to a recent review by Scott et al., 2013 for chemometric classification problems (286 reviewed papers), RF is used in only 4.5% of the articles where machine-learning algorithms are applied. The same source evidences that boosting algorithm is even less used (1%).

The aim of the work was to evaluate different chemometric approaches in the evaluation of data obtained from parameters influenced by the grape ripening process, such as berry mechanical properties data and flavonoid content in berry skins. For this, the performances of RF and GBM algorithms were compared on a large dataset composed of approx. 800 berries belonging to 22 grapevine cultivars, their suitability for flavonoid content prediction in grape berries was evaluated on the basis of mechanical properties, and an informal explanation of the underlying algorithms was suggested. Furthermore, a predictive model was also developed. This approach could be used as an example for other compounds and fruits.

2. Materials and methods

2.1. Grape sampling

Grapes from 22 red grapevine cultivars (*Vitis vinifera* L.) were sampled in the CRA-VIT experimental collection (1.2 ha) located in Susegana (TV), Veneto Region (North-East Italy), in 2010 and 2011. Vines were 15 years old, grafted on SO4 rootstock (interspecific cross between *Vitis riparia* Michx. and *Vitis berlandieri* Planch.), and planted at 3.0 m between rows and 1.5 m between vines. They were Sylvoz pruned and trained with a vertical shoot position system. For each cultivar, samples were composed of approx. 3 kg of grape berries, which were picked up randomly from ten vines. In order to successfully compare berries at ripeness with adequate sugar content, the berries were calibrated using a densimetric method by berry flotation in different saline solutions (Rolle et al., 2011). This study was carried out only on the berries with sugar contents comprised between 183 ± 8 g L^{-1} and 217 ± 8 g L^{-1} corresponding to 11.0 ± 0.5% (v/v) and 13.0 ± 0.5% (v/v) potential alcohol, respectively.

The sorted berries were visually inspected before analysis; those with damaged skins were discarded. For each variety studied, a sub-sample of 36 sorted berries (therefore a total of 792 berries for all cultivars together) was randomly selected for the determination of the physical–mechanical properties and then for the flavonoid content. As described in the successive section, single berries measurements were then averaged by three to compose a single sample for predictive modeling.

2.2. Physical and mechanical properties

Grape berries were singularly weighed, with an analytical laboratory balance Radwag AS 220/X (Radwag, Radom, Poland), and then a Texture Profile Analysis (TPA) non destructive mechanical test was performed for each of them as described by Letaief et al., 2008. It allowed the measurement of berry hardness (N, as *H*), cohesiveness (adimensional, as Co), gumminess (N, as *G*), springiness (mm, as *S*), chewiness (mJ, as Ch) and resilience (adimensional, as *R*). A puncture test (Letaief et al., 2008) was then carried out on the same berries taken singularly to measure skin break force (N, as F_{sk}), skin break energy (mJ, as W_{sk}) and skin resistance to axial deformation (N mm^{-1}, as E_{sk}). All these measurements were performed on the equatorial position of whole berry, while skin thickness (μm, as Sp_{sk}) was measured in the skin after manual removal from the pulp with a razor blade (Letaief et al., 2008; Río Segade et al., 2011a). Analyses were made with a Universal Testing Machine (UTM) TAxT2i texture analyzer (SMS-Stable Micro Systems, Godalming, Surrey, UK) equipped with a 5 kg load cell and a HDP/90 platform. A SMS P/35 flat probe under 25% deformation, with a waiting period of 2s between the two compressions and a speed of 1 mm s^{-1}, was used for the TPA test. A SMS P/2N needle probe, with a test speed of 1 mm s^{-1} and a penetration depth of 3 mm, was used for the puncture test. A SMS P/2 flat probe, with a test speed of 0.2 mm s^{-1} was used to measure Sp_{sk}. All data were acquired at 400 Hz and evaluated using the Texture Expert Exceed software, version 2.54.

2.3. Skin flavonoid content

After the skin thickness test, each berry skin was individually immersed for 4 h in 5 mL of a buffer solution containing 12% v/v ethanol, 2 g L^{-1} of $Na_2S_2O_5$, 5 g L^{-1} of tartaric acid and adjusted to pH 3.20 with NaOH (Di Stefano and Cravero, 1991). Each skin was then homogenized at 8000 rpm for 1 min with an Ultraturrax T18 (IKA Labortechnik, Staufen, Germany), and the extract was centrifuged for 10 min at 3500 × g and 20 °C. The supernatant was then used for analysis after dilution with an ethanolic solution of HCl (70:30:1, ethanol:water:HCl, v/v) (Di Stefano and Cravero, 1991). Total flavonoid index (TF) was determined by a spectrophotometric method, reading the absorbance at 280 nm, using an Uvmini-1240 PC spectrophotometer (Shimadzu Scientific

Instruments, Columbia, MD, USA) and expressed as mg g^{-1} berry of (+)-catechin (Rolle et al., 2011; Di Stefano and Cravero, 1991).

2.4. Predictive modeling

2.4.1. Description of the used machine-learning techniques

The relationship between predictors and the outcome was modeled using Regression Trees, RT (Breiman et al., 1984) and two derived techniques: RF (Breiman, 2001) and stochastic gradient boosting with trees as base learners; the latter will be here called Gradient Boosting Machine (GBM) in reference to the work where this technique first appeared (Friedman, 2001). A comprehensive description of these techniques cannot be given in few words, nevertheless the following paragraphs will try to briefly and lightly introduce the subject. Readers interested in more technical details can find worthwhile information in (Hastie et al., 2009) and in the help and vignettes of the cited R packages.

Regression trees are rule based models that split the whole dataset in groups where data tend to be homogeneous with respect to the response. In the technique used in this work, which is known as Classification And Regression Tree, CART (Breiman et al., 1984), data in the terminal nodes (the final groups that are no further partitioned) are simply averaged to predict the outcome. At the beginning, the entire dataset is split in two groups to minimize the overall sum of squares, by searching every value of every predictor. The technique is then recursive, these two groups are split again in two parts each to further reduce the prediction error, according to the available predictor values. This technique is also known as recursive partitioning because of its iterative nature. The number of groups duplicates at each split until the terminal nodes are so small that they cannot be further partitioned. However, these “full grown” trees generally overfit, in the sense that they tend to fit the noise other than the structure in the training data. Therefore, they achieve poor performances on the validation data despite having great performances on the training data. Their growth must therefore be controlled, and this can be obtained by cross-validation procedures. Cross-validation is a form of internal validation, which is based on the use of a fraction only of the whole training data to develop the model, while using the remaining part for the validation. In k-fold cross validation, the training dataset is divided in k parts; $k-1$ parts are used to fit the model and the kth part is used to evaluate the structure of the model on simulated new data. The procedure is then iterative, and all k parts serve as validation once at a time. This allows determining the size of the trees enabling the best results on future unseen data.

A characteristic of regression trees is their instability, their structure can greatly vary with the data available for modeling. This property can appear at a first sight a deficiency of the method, instead it has become to be really useful and extremely well exploited by two state-of-the-art techniques in statistical learning such as bagging and boosting. These techniques are based on the “perturb and combine” strategies (Breiman, 1996a) and on the idea that combined learners can outperform single ones. For this combination to be effective, single learners must be able to capture a part of the structure in the data that is not modeled by other learners. The plasticity of trees can be exploited for this purpose: by artificially varying the available data through re-sampling techniques, they can be induced to learn different aspects of the dataset. Partial predictions from ensemble of trees are then combined to obtain the final predictions. Bagging (Breiman, 1996b) and boosting (Freund and Schapire, 1997) are two ways of combining learners. In bagging, trees are grown in parallel on a part of the available data, and predictions are then averaged across all trees. However, in boosting, trees are grown sequentially, and each successive tree models the residuals of the previous tree predictions. Bagging and boosting are the two techniques that, further optimized by increasing randomization, are respectively used in RF and GBM. In RF, trees are grown on re-sampled subsets of the training data by using only some of the available predictors, randomly chosen at each split. Final predictions for each tree are then averaged. This parameter is called *mtry* and has to be set by the user, as well as the number and depth of trees in the forest. In GBM, trees are sequentially built to reduce the errors of the previous trees, but residuals are resampled and just a fraction is available for modeling at each iteration. Furthermore, learning is regularized through shrinkage, i.e. learning rate is slowed by allowing the use of just a fraction of the whole value for each residual. As occurred in RF, even in GBM, the number and depth of trees have to be selected by the user. Parameter selection, also called parameter tuning, is generally made using cross-validation or bootstrap techniques in order to minimize the performances of the algorithm on simulated new data, being not the error on the training set a robust choice because of overfitting.

2.4.2. Details about the used procedure

A predictive model was built to predict the flavonoid content in berry skin using physical and mechanical properties of the whole berry and skin as inputs. The 792 berries data (36 berries × 22 cultivars) were averaged by three, randomly selected inside the same cultivar, to obtain 264 averaged samples (12 samples composed by 3 berries for 22 cultivars). Prior to model fitting, data were partitioned and a random approx. 20% of data (53 samples) were left out from the training set for later use as test set. Data were then centered and scaled. In this work, models were tuned using 10 repetitions of 5 folds cross-validation in order to optimize the Root Mean Square Error (RMSE) on the resampled data (more than 10,000 possible combinations were evaluated). Performances of the models were then compared on the same set of 100 bootstrap re-samples always using RMSE as a metric.

Data were analyzed with the R statistical software 3.1.2 (R Core Team, 2014) using the packages rpart (Therneau et al., 2015), randomForest (Liaw and Wiener, 2002), gbm (Ridgeway, 2013), caret (Kuhn et al., 2014).

3. Results and discussion

3.1. Descriptive analysis

Fig. 1 shows the TF content for the 22 cultivars used in the experiment. Raboso, Ancellotta and Teroldego had the highest amount of TF in the skin of fresh berries (with a median of 8.72, 6.52 and 5.35 mg g^{-1} berry, respectively) but also had the highest variance (1.52, 4.20 and 0.71 mg g^{-1} berry, respectively). Gamay, Schiava gentile and Aleatico had the lowest concentration in these compounds (with a median of 0.97, 0.91 and 0.98 mg g^{-1} berry, respectively) and the lowest variance (0.01, 0.01 and 0.01, respectively). In the global dataset (Table 1), TF had a mean of 2.60 mg g^{-1} berry, a median of 2.07 mg g^{-1} berry and a variance of 4.16 mg g^{-1} berry. The registered minimum was 0.69 mg g^{-1} berry (Aleatico), while the maximum was 12.87 mg g^{-1} berry (Ancellotta). Descriptive statistics for the physical–mechanical characteristics in the global dataset is shown in Table 1. Data were in agreement with those reported in scientific literature in several works (Zouid et al., 2013; Letaief et al., 2008; Río Segade et al., 2011a).

Table 2 shows Pearson correlations in the global dataset. Being TF the outcome of the developing model, good correlations with the available predictors would be welcomed, but r values for this variable were moderate. BW and S showed the highest correlations with TF (r values of -0.59 and -0.53, respectively, p-value < 0.001), followed by E_{sk} and Sp_{sk} (r values of 0.34 and 0.34, respectively,

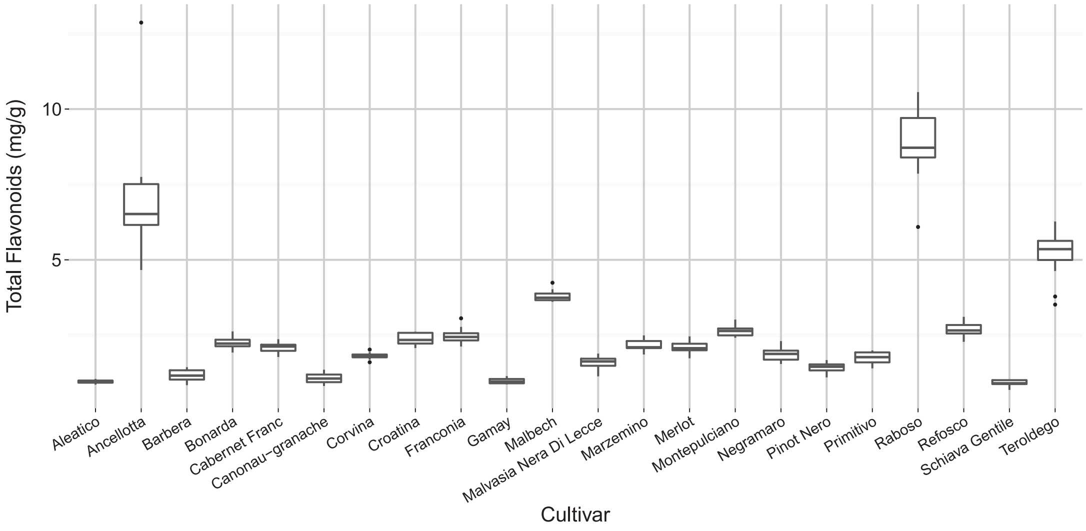


Fig. 1. Flavonoid content for all cultivars in the experiment. Flavonoid content (mg g^{-1} berry) for the 22 red wine grape cultivars used in the experiment.

Table 1
Descriptive statistics for the global dataset of physical–mechanical properties and total flavonoid content composed of 792 berries from 22 red wine grape cultivars.

	Min.	Max.	Mean	Median	Var.
BW (g)	1.31	4.13	2.57	2.54	0.331
F_{sk} (N)	0.18	1.01	0.65	0.66	0.024
W_{sk} (mJ)	0.11	1.47	0.64	0.64	0.040
E_{sk} (N mm^{-1})	0.15	0.47	0.29	0.29	0.004
Sp_{sk} (μm)	122.33	315.00	209.69	201.00	1599.699
SW (g)	0.12	0.45	0.26	0.25	0.005
H (N)	1.36	5.68	3.19	3.05	0.655
Co (adimens.)	0.60	0.89	0.79	0.80	0.003
G (N)	1.13	4.45	2.51	2.43	0.377
S (mm)	1.62	3.00	2.43	2.48	0.069
Ch (mJ)	1.86	11.89	6.21	6.01	3.193
R (adimens.)	0.31	0.51	0.45	0.41	0.001
TF (mg g^{-1})	0.69	12.87	2.60	2.07	4.164

p-value < 0.001). The less related predictors were W_{sk}, H and G, which did not show significant correlations. It should be noticed that the last two variables were well related to the anthocyanin extractability in the study published by (Zouid et al., 2013). In the present work, the number of cultivars taken in account is greatly higher compared to the cited work, where only Cabernet-Franc was measured. Given the lack of significance, it could be hypothesized that the relation is not uniform but depends on the cultivar. The Spearman's method, used to highlight possible monotonic, but non-linear relations with TF, did not give association values higher than those already observed (data not shown).

The strongest correlations observed in the dataset (r values higher than 0.75, p-value < 0.001) were those among the physical–mechanical predictors (H with G and Ch; F_{sk} with W_{sk} and E_{sk}; BW with S and SW), which is a consequence of the way they were measured or calculated (Letaief et al., 2008). H, G and Ch were strongly positive-related because G and Ch were calculated from H. Therefore, harder berries were also more gummy and chewy. Considering skin related mechanical properties, F_{sk} corresponds to the skin resistance to the needle probe penetration, while W_{sk} is represented by the area under the force/time curve. E_{sk} is defined as the slope of the stress–strain curve in the linear section. W_{sk} and E_{sk} were strongly positive-related to F_{sk}, so stiffer skins were also more resistant to the penetration and therefore harder. Furthermore, heavier berries, which are also bigger ones, had higher value of S and had, obviously, higher amount of skin. BW was retained instead of S because the relationship of BW with TF is well known in the literature.

It is important to highlight that the sugar content of berry showed some significant relations (alpha risk < 0.1) with other variables in the dataset, Table 2, but the correlation was strong with none of them. The effect of stage of ripening on mechanical properties values was in general less determining in comparison to variety effect (Río Segade et al., 2008). Cultivar variability of these properties across the 22 cultivars studied clearly dominated. Before continuing, it should be cleared that even if the correlations in Table 2 let to make an idea of the main relationships in the dataset, possible multidimensional relations were not taken into account. As an example, it seems logical that total skin weight could be the result of a linear combination between the berry size and the skin thickness for each berry. SW was highly related to a combination of BW and Sp_{sk} with a r value of 0.87. This value was clearly higher than those of the single relations, being probably redundant the information of SW if used in a model also containing BW and Sp_{sk}.

3.2. Predictor filtering

Sensibility to correlated predictors depends on the used statistical learning technique, but it is generally not welcomed because redundant and non informative inputs reduce model performances. When inference is the objective, the negative effect of correlated variables is even worse than for predictions alone. Furthermore, the measurement of a greater number of variables in order to apply a model would increase costs and time, and therefore a justification is required. Any of the three used learning techniques (RT, RF, GBM) completely fail when correlated predictors are present, the less sensitive technique probably being GBM because it shrinks effect estimates (Maloney et al., 2012), and the most sensitive one being RF (Strobl et al., 2007). In general, tree based techniques implicitly run feature selection because, if a predictor does not permit to reduce the residual sum of squares at any tree split, its contribution to the model is zero. However, if highly correlated predictors are present, the choice between them is somewhat random because they similarly reduce the sum of squares, and have a similar probability to be chosen for a given split. In RF, where an $mtry$ number of predictors is sampled at each split (see Section 2.4), the presence of correlated predictors increases the chance to sample similar information. It reduces

Table 2
Pearson correlations for the global dataset of physical–mechanical properties, total flavonoid and total soluble solids content obtained from 792 berries from 22 red wine grape cultivars. Numbers stay for the Pearson correlation coefficient *r*, while stars are for *p*-value.

	BW	F_{sk}	W_{sk}	E_{sk}	Sp_{sk}	SW	*H*	Co	*G*	*S*	Ch	*R*	TSS
BW													
F_{sk}	-0.13^{*}												
W_{sk}	0.04^{***}	0.92^{***}											
E_{sk}	-0.29^{***}	0.81^{***}	0.54^{***}										
Sp_{sk}	0.06	−0.07	−0.12	0.03									
SW	0.76^{***}	-0.21^{***}	−0.11.	-0.25^{***}	0.46^{***}								
H	0.17^{**}	0.54^{***}	0.30^{***}	0.75^{***}	0.06	0.14^{*}							
Co	0.03	−0.11	−0.06	-0.17^{**}	−0.01	0.08	-0.22^{***}						
G	0.19^{**}	0.53^{***}	0.31^{***}	0.73^{***}	0.05	0.16^{**}	0.97^{***}	0.01					
S	0.88^{***}	−0.06	0.02	−0.10	−0.06	0.62^{***}	0.31^{***}	0.15^{*}	0.36^{***}				
Ch	0.44^{***}	0.42^{***}	0.26^{***}	0.57^{***}	0.04	0.35^{***}	0.91^{***}	0.05	0.95^{***}	0.61^{***}			
R	0.06	-0.23^{***}	-0.15^{*}	-0.29^{***}	-0.20^{**}	0.01	-0.38^{***}	0.58^{***}	-0.24^{***}	0.25^{***}	-0.14^{*}		
TSS	$-0.11^{\cdot}$	−0.01	−0.02	0.01	−0.02	0.00	−0.05	0.04	−0.05	−0.11.	−0.07	0.04	
TF	-0.59^{***}	0.13^{*}	−0.03	0.34^{***}	0.34^{***}	-0.18^{**}	0.00	-0.13^{*}	−0.03	-0.53^{***}	-0.20^{**}	-0.19^{**}	0.03

*** *p*-value < 0.001.
** *p*-value < 0.01.
* *p*-value < 0.05.
· *p*-value < 0.1.

randomization and therefore independence across trees; important assumption to optimize performances. In addition, it dilutes the importance of key predictors and increases the importance of weak variables correlated to important ones (Strobl et al., 2007).

To account for these problems and to optimize model performances, predictors were first filtered to avoid correlation levels higher than 0.7 (according to Pearson correlation coefficients, Table 2) and therefore to reduce redundancy. The relationships of the predictors with the outcome were not considered for selecting predictors in this first phase. Feature selection was indeed performed successively using Recursive Feature Elimination (RFE). Four predictors were eliminated by this filtering step, which were F_{sk}, W_{sk}, *H*, and *G*. F_{sk} and W_{sk} were highly related ($r = 0.92$, $p < 0.001$). W_{sk} was not related to TF, contrarily to F_{sk}, but this last was also strongly related to E_{sk} ($r = 0.81$, $p < 0.001$). A previous study has reported that E_{sk} is related to cellular maturity index (EA%) as predictors of anthocyanin extractability (Río Segade et al., 2011b). Among *H*, *G* and Ch ($r = 0.91$–0.97, $p < 0.001$), this last was retained because it was also well related to anthocyanin extractability (Zouid et al., 2013). The information provided by SW in a model can be well approximated by a combined use of BW and Sp_{sk}, as previously explained. Furthermore, Sp_{sk} is considered as main texture parameter to predict anthocyanin extractability in winegrapes (Río Segade et al., 2011c). The final set of filtered predictors included BW, E_{sk}, Co, Sp_{sk}, Ch, W_{sk}, and *R*.

3.3. Recursive feature elimination and model tuning

Recursive Feature Elimination (RFE) (Guyon et al., 2002), a backward selection algorithm, was used in the way optimized by Ambroise and McLachlan, 2002, and therefore including feature selection in the model building process. Predictors elimination was evaluated on the basis of the performances achieved on resampled sets obtained by *k*-fold cross validation ($k = 5$). The process was run for RT, RF, and GBM, and models were also tuned to optimize performances during the process (see Section 2.4). RT, RF, and GBM were all tuned using the same set of re-samples therefore ensuring consistency in the evaluation and allowing comparison across model performances.

In all three cases (RT, RF, and GBM), RFE suggested the use of all seven available predictors (BW, E_{sk}, Co, Sp_{sk}, Ch, W_{sk}, and *R*), and therefore all of them had some influence on the techniques evaluated. The relative predictor importance in all models is shown in Table 3. In this table, the influence was scaled between 0 and 100 to allow an easier comparison between models, but as already stated some of the selected predictors had zero influence. The 0 and 100 are relative values obtained by subtracting the minimum registered influence (across all predictors) from the individual influence for each predictor, and then by dividing for the difference between the maximum and the minimum registered influence. The influence of each predictor in the model varied according to the model. Considering a single tree (RT), the overall relative influence of each predictor was higher, because all predictors were used once or few times, and this avoided the predominance of very strong predictors such as BW. In RT, Sp_{sk} was the predictor that allowed the greatest error reduction. In ensembles, and with the perturbation of data imposed in RF and GBM methods, the influence of some strong predictors popped up and seems to take advantage over the others. This was more evident in GBM than in RF, which had an intermediate behavior. These comments are valid only for this study.

Model tuning suggested the use of 7 splits for RT, 1000 trees and *mtry* = 4 for RF, 5000 trees having 4 splits each and a shrinkage of 0.005 for GBM. Fig. 2 shows the tuned RT with the aim of illustrating the basic element also composing RF and GBM ensembles. Single trees are very easy to interpret and to allow making an idea of the relationships in the dataset. It is important to remember that they are fairly unstable, and small perturbations in the dataset can completely change their structure. Therefore, trees just describe relationships relative to the data observed, and interpretations are difficultly generalizable. This is especially true for the lower splits. However, it is worthwhile to note that Sp_{sk}, which was the variable with the largest influence for RT (Table 3), acts in a controversial fashion. For the smallest berries, which were also the richest in flavonoids, a higher Sp_{sk} indicated a lower content of TF, while for the biggest berries, the inverse was true. It is also

Table 3
Relative influence of each predictor in all tested algorithms. Influence was scaled between 0 and 100 to allow an easier comparison.

	RT	GBM	RF
C	0.00	6.20	4.23
R	17.00	0.00	0.00
W_{sk}	51.25	1.31	0.98
Ch	35.41	2.79	15.80
Sp_{sk}	56.41	3.49	3.03
E_{sk}	100.00	4.77	26.77
BW	42.07	100.00	100.00

important to note the role of BW, which was negatively related to the amount in phenolic compounds as widely discussed in the literature (Barbagallo et al., 2011). Also E_{sk} seems to be an important parameter because elastic skins were associated to higher content in TF. BW, E_{sk}, and Sp_{sk} were the predictors with the highest influence in all models (Table 3).

Table 4
Results (R^2 and Root Mean Squared Error, RMSE) of the tested algorithm on the training set, the external validation (test set) and the internal validation (10 repetitions of 5-fold Cross-Validation, CV). For cross-validation estimations of the standard deviation for both metrics are also shown. RMSE results are expressed in mg g^{-1} berry, lower the error better the performances of the model.

	Train RMSE	Train R2	Test RMSE	Test R2	CV RMSE	CV R2	CV RMSE SD	CV R2 SD
RT	0.817	0.849	0.951	0.752	1.286	0.650	0.198	0.101
RF	0.419	0.965	0.729	0.836	1.071	0.754	0.148	0.063
GBM	0.364	0.971	0.745	0.836	1.074	0.753	0.147	0.058

3.4. Model comparison

Results of the tuned models are shown in Table 4 for both training and test data and for the cross-validated re-samples. It appears that all algorithms, starting from RT, tended to overfit the training set, which is probably a consequence of a training set too small when compared to the complexity of the relationships among predictors and between these and the outcome, as suggested by the weak correlations observed in Table 2. Despite this, the model accurately predicted the test set used as external validation. Predictions for the test set exceeded those obtained with 5-fold cross-validation, which can be considered an internal validation. Test sets are considered the ultimate proof of model performances, often neglecting cross-validation and bootstrap assessment methods. However, observations like the present one make us think about the method to prefer in model assessment. The performances observed over a test set could also be attributed to random select observations easier to predict than those contained in the training set used in cross-validation. Resampling methods are more robust from this point of view, but they can be upward (i.e. pessimistic) biased, especially the bootstrap, even if an alternative to avoid such bias is available, but only for classification problems (Efron and Tibshirani, 1997).

Fig. 3 shows the predictions on the train and test sets for all methods. Fig. 3a and b shows the blocky structure used in prediction by RT, where similar data were just predicted by the mean of the group they belong according to the rules in Fig. 2, and are therefore grouped also in predictions. In GBM and RF ensembles, the predictions are averaged from many trees and allow the methods to be more adaptable to the form of data and to also model non-linearity (and interactions). In Fig. 3c/d and e/f, predictions were no longer grouped for the same values of predictions. Comparing Fig. 3c/d with e/f, it appears that both RF and GBM methods well predict the test set, but predictions, as shown by the location of points in the scatterplots, although similar were not exactly identical. Predictions obtained on the same re-sampled data by GBM and RF were highly correlated (0.82), however looking at Table 3 it appears that they did not make use of the same predictors in the same way. It will be possible that combining both methods in a single ensemble will boost the overall accuracy a little bit. These algorithms were combined by weighted average of their predictions, using a greedy optimization method as described in (Caruana et al., 2006, 2011). Combining methods in a single ensemble of models has the highest efficacy when algorithms are

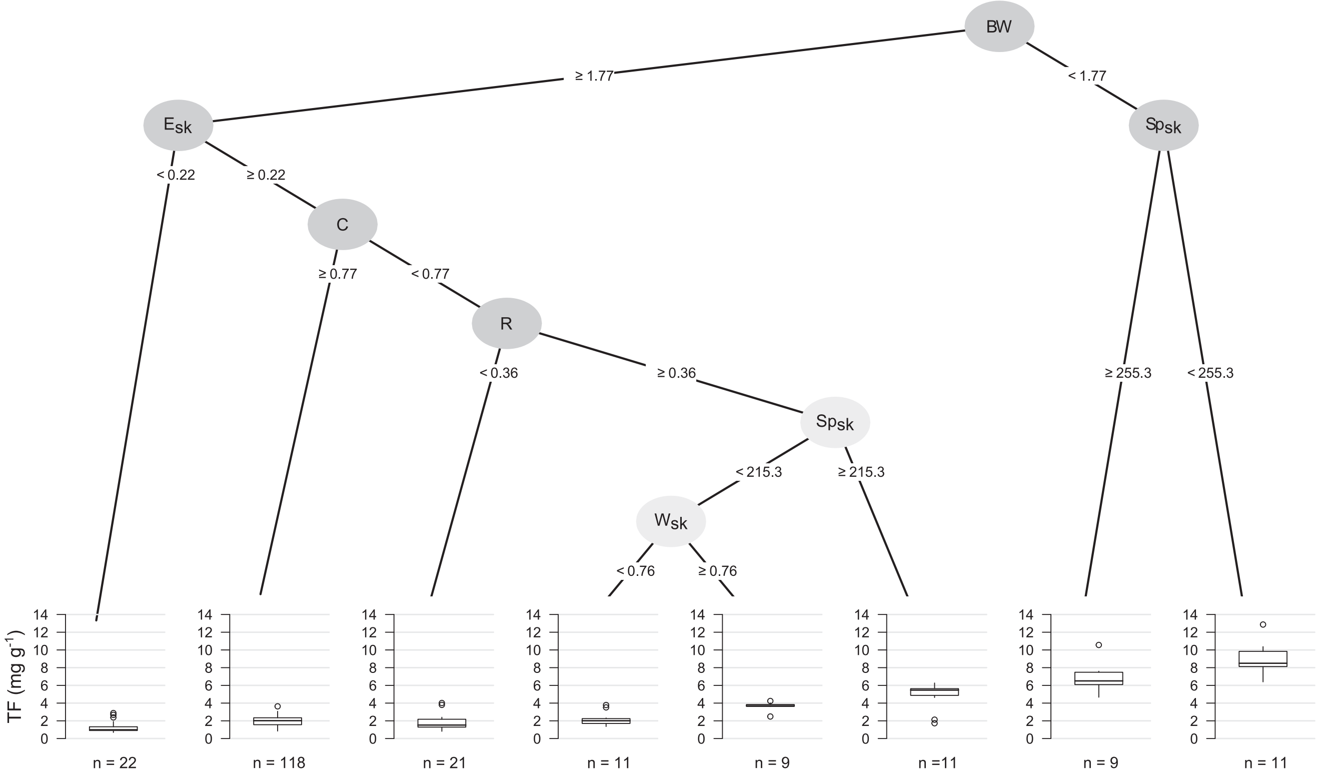


Fig. 2. Best regression tree model to predict total flavonoids from texture analysis data. Figure representing the tuned RT on the training dataset. The whole training dataset is recursively split in two parts, according to the predictor that allows the greatest reduction in the residual sum of squares. The selected predictor at each split is shown inside the ellipse, while just under there is the rule used for splitting which is a corresponding predictor value. Here numbers are expressed in the original measure unit of each predictor, readers are kindly referred to Table 1 for the complete list. The optimal number of splits according to the results of the cross-validation procedure was equal to 7.

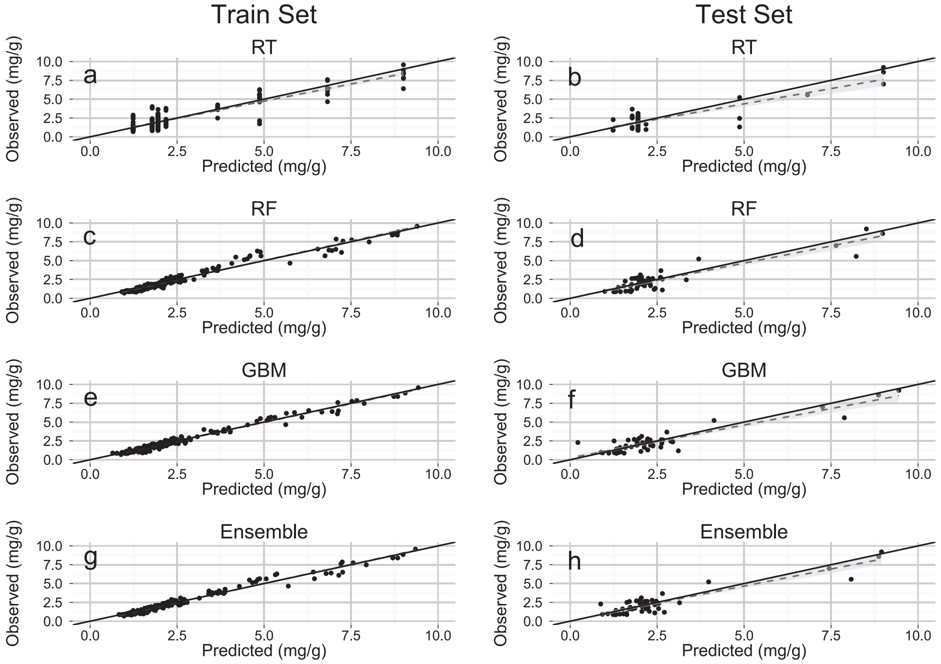


Fig. 3. Results of the machine-learning techniques on train and test data-sets. Relationships between observed and predicted TF data (mg g^{-1}) over the train test and the test set used as external validation for all algorithms. Solid black line is the identity line, while the dashed gray line is a linear regression (OLS) applied to the data and the filled gray region is its 95% confidence interval.

different, and therefore predictions uncorrelated. It is not the case here, where these assumptions are not really respected. However, the combination of RF and GBM brings to a nice improvement in model predictions. GBM and RF predictions were weighted 0.51 and 0.49, respectively, for averaging, and the resulting RMSE of their ensemble was 1.05 mg g^{-1}, which was slightly lower than that obtained by single methods (Table 4). The result on the test set for the RMSE was 0.701 mg g^{-1} and for R^2 was 0.85. The corresponding predictions on the train set are in Fig. 3g and in the test set are shown in Fig. 3h.

To compare the results of this work with others found in the literature is somewhat difficult, because the use of texture analysis to predict flavonoid content in grape berries is novelty, and also a so varied dataset, containing 22 cultivars, is rare to be found. Texture analysis was already used to develop rapid method for the evaluation of total phenolic content and phenol extractability in grape seeds with a good accuracy (Rolle et al., 2013), and in skins but limited to the anthocyanin content (Rolle et al., 2012b; Río Segade et al., 2011c). In grape berries, however, a rapid evaluation of the phenolic content has generally been made using NIR spectroscopy, and several works have reached very good performances (Ferrer-Gallego et al., 2011). This work was performed on a single cultivar (Graciano), and data were expressed in mg g^{-1} of berry skins. In the present work, data were expressed in mg g^{-1} of whole berry, which from an industrial point of view could be more practical. The results of the last two studies are therefore not directly comparable.

The results obtained showed that RF and GBM, and even their average can reach a very high accuracy for TF prediction from physical–mechanical data obtained for many different cultivars. RF is simpler to perform and accurately tune than GBM. Furthermore, its use of features in the dataset less overfitted BW influence.

However, even if the performances of those algorithms were very high, it is also true that for real world application, model performances were still too low to be practically used in a generalized way. Results obtained with prediction could be useful for the comparison of the phenolic maturity of different vineyards, but at this time they were hardly suitable for the monitoring of TF during ripening for cultivars with low amounts of flavonoids. Conversely, they could be used for those cultivars with very high amounts of flavonoids (such as Raboso, Ancellotta and Teroldego), because a reduced relative error in prediction.

It is possible that, being the physical–mechanical characteristics linked to total flavonoids in a way that depends on the cultivar and is not universal, cultivar-specific calibration will be necessary to improve model performances. This will probably allow the use of TA to monitor grape ripening even for cultivars with low amount in flavonoids. Cultivar-specific calibration or the inclusion of the cultivar as a categorical term in the developed models was not possible in this work because the number of observations by cultivar was too low.

To further increase model accuracy, it will also be interesting to test the average of more than three berries for a single sample in order to improve the accuracy in the TA predictors. It could also be interesting to normalize the results using other properties or evaluated parameters. Finally, it will be important to acquire more data and to develop cultivar-specific calibrations. Indeed, except BW, which had a homogeneous behavior for all 22 cultivars in the dataset, other physical–mechanical parameters greatly varied by the cultivar, and general patterns were weak.

4. Conclusions

This work collected and assessed a large and varied dataset of texture analysis data and flavonoid content from the analysis of every single berry. It tried to evaluate different machine-learning algorithms to assess their suitability to model the relationships between physical–mechanical characteristics of grape and the concentration of skin flavonoids. The reason for modeling such a relation is that grape berries show changes in their physical–mechanical properties during ripening, which are variety dependent. The approaches evaluated here (RF and GBM) are state-of-the art techniques, but have still rarely been used in chemometrics. This work brings an interesting case-study while also trying to simply and informally explain the way these methods work, starting from their basic element, RT. It will serve as an introduction and will offer some valuable insights for food scientists interested in learning more about these techniques or searching for domain-specific examples of application.

Presented models are able to capture a huge portion of the variability in the dataset, as shown by the reached R^2 and accuracy (given by the RMSE), and they can be useful for a fast screening of many cultivars, because it does not ask for sample preparation and extraction, but not yet for fine measurements. It should also be considered that, even if the number of cultivars in this study was high, universal considerations cannot be inferred because, as already reported in the discussion, the evolution of physical–mechanical parameters with ripening could be different across cultivars. Therefore it is probable that conclusions obtained from this study could be different when developing models for a single cultivar, especially in the role and importance of the used physical predictors. It is also highly probable that machine-learning techniques, once applied on single cultivars or on groups of cultivars presenting a similar evolution of physical–mechanical properties with the ripening, will reach outstanding performances and could allow a rapid and accurate estimation of ripening-influenced parameters like TF in grape berries.

References

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS 99, 6562–6566.

Barbagallo, M.G., Guidoni, S., Hunter, J.J., 2011. Berry size and qualitative characteristics of *Vitis vinifera* L. cv. Syrah. S. Afr. J. Enol. Vitic. 32, 129–136.

Battista, F., Tomasi, D., Porro, D., Caicci, F., Giacosa, S., Rolle, L., 2015. Winegrape berry skin thickness determination: comparison between histological observations and texture analysis determination. Ital. J. Food Sci. 27, 136–141.

Braidot, E., Zancani, M., Petrussa, E., Peresson, C., Bertolini, A., Patui, S., Macri, F., Vianello, A., 2008. Transport and accumulation of flavonoids in grapevine (*Vitis vinifera* L.). Plant Signal. Behav. 3, 626–632.

Breiman, L., 1996. Bias, variance, and arcing classifiers. In: Technical report 460 Statistics Department University of California.

Breiman, L., 1996b. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Statistics and probaility series, Wadsworth, Belmont, CA.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cozzolino, D., 2015. The role of visible and infrared spectroscopy combined with chemometrics to measure phenolic compounds in grape and wine samples. Molecules 20, 726–737.

Caruana, R., Munson, A., Alexandru, N.-M. Getting the most out of ensemble selection. Int. Conf. Data Min., 2006, pp. 1–12.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2011. Ensemble selection from libraries of models. Proc. ICML '04 2011, vol. 34, pp. 1–21.

Di Stefano, R., Cravero, M.C., 1991. Metodi per lo studio dei polifenoli dell'uva. Riv. di Vitic. ed Enol. 44, 37–45.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560.

Ferrer-Gallego, R., Hernàndez-Hierro, Rivas.-Gonzalo, J., Escribano-Bailón, M.T., 2011. Determination of phenolic compounds of grape skins during ripening by NIR spectroscopy. LWT-Food Sci. Technol. 44, 847–853.

Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Giordano, M., Zecca, O., Belviso, S., Reinotti, M., Gerbi, V., Rolle, L., 2013. Volatile fingerprint and physico-mechanical properties of 'Muscat blanc' grapes grown in mountain area: a first evidence of the influence of water regimes. Ital. J. Food Sci. 25, 329–338.

Gonzalez, J., Donoso, W., Sandoval, N., Reyes, M., Gonzalez, P., Gajardo, M., Morales, E., Neira, A., Razmilic, I., Yuri, J.A., Moore-Carrasco, R., 2015. Apple peel supplemented diet reduces parameters of metabolic syndrome and atherogenic progression in ApoE −/− Mice. Evidence-Based Complement. Altern. Med.

González-Neves, G., Charamelo, D., Balado, J., Barreiro, L., Bochicchio, R., Gatto, G., Gil, G., Tessore, A., Carbonneau, A., Moutonet, M., 2004. Phenolic potential of Tannat, Cabernet-Sauvignon and Merlot grapes and their correspondence with wine composition. Anal. Chim. Acta 513, 191–196.

Guyon, I., Weston, J., Barnhill, S., Vladimir, V., 2002. Gene selection for cancer classification using Support Vector Machines. Mach. Learn. 46, 389–422.

Harnly, J.M., Doherty, R.F., Beecher, G.R., Holden, J.M., Haytowitz, D.B., Bhagwat, Gebhardt, S., 2006. Flavonoid content of U.S. fruits, vegetables, and nuts. J. Agric. Food Chem. 54, 9966–9977.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction, second ed. Springer, Netherlands.

Ignat, I., Volf, I., Popa, V.I., 2011. A critical review of methods for characterisation of polyphenolic compounds in fruits and vegetables. Food Chem. 126, 1821–1835.

Kuhn, N., Guan, L., Dai, Z.W., Wu, B.H., Lauvergeat, V., Gomès, E., Li, S.H., Godoy, F., Arce-Johnson, P., Delrot, S., 2013. Berry ripening: recently heard through the grapevine. J. Exp. Bot. 65, 4543–4559.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., 2014. Caret: classification and regression training. R package version 6.0-37. <http://CRAN.R-project.org/package=caret>.

Koundouras, S., Marinos, V., Gkoulioti, A., Kotseridis, Y., Van Leeuwen, C., 2006. Influence of vineyard location and vine water status on fruit maturation of non-irrigated cv. Agiorgitiko (*Vitis vinifera* L.). Effects on wine phenolic and aroma components. J. Agric. Food Chem. 54, 5077–5086.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Letaief, H., Rolle, L., Gerbi, V., 2008. Mechanical behavior of wine grapes under compression tests. Am. J. Enol. Vitic. 59, 323–329.

Lorrain, B., Ky, I., Pechamat, L., Teissedre, P.L., 2013. Evolution of analysis of polyhenols from grapes, wines, and extracts. Molecules 18 (1), 1076–1100.

Lourenço, F., Gago, B., Barbosa, R.M., De Freitas, V., Laranjinha, J., 2008. LDL isolated from plasma-loaded red wine procyanidins resist lipid oxidation and tocopherol depletion. J. Agric. Food Chem. 56, 3798–3804.

Maloney, K.O., Schmid, M., Weller, D.E., 2012. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. Methods Ecol. Evol. 3, 116–128.

Noll, C., Hamelet, J., Matulewicz, E., Paul, J.L., Delabar, J.M., Janel, N., 2009. Effects of red wine polyphenolic compounds on paraoxonase-1 and lectin-like oxidized low-density lipoprotein receptor-1 in hyperhomocysteinemic mice. J. Nutr. Biochem. 20, 586–596.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>.

Ridgeway, G., 2013. Gbm: generalized boosted regression models. R package version 2.1. <http://CRAN.R-project.org/package=gbm>.

Ristic, R., Bindon, K., Francis, L.I., Herderich, M.J., Iland, P.G., 2010. Flavonoids and C13-norisoprenoids in *Vitis vinifera* L. cv. Shiraz: Relationships between grape and wine composition, wine colour and wine sensory properties. Aust. J. Grape Wine Res. 16, 369–388.

Río Segade, S., Rolle, L., Gerbi, V., Orriols, I., 2008. Phenolic ripeness assessment of grape skin by texture analysis. J. Food Compos. Anal. 21, 644–649.

Río Segade, S., Orriols, I., Giacosa, S., Rolle, L., 2011a. Instrumental texture analysis parameters as winegrapes varietal markers and ripeness predictors. Int. J. Food Prop. 14, 1318–1329.

Río Segade, S., Soto Vázquez, E., Orriols, I., Giacosa, S., Rolle, L., 2011b. Possible use of texture characteristics of winegrapes as markers for zoning and their relationship with anthocyanin extractability index. Int. J. Food Sci. Technol. 46, 386–394.

Río Segade, S., Giacosa, S., Gerbi, V., Rolle, L., 2011c. Berry skin thickness as main texture parameter to predict anthocyanin extractability in winegrapes. LWT-Food Sci. Technol. 44, 392–398.

Rolle, L., Rio Segade, S., Torchio, F., Giacosa, S., Cagnasso, E., Marengo, F., 2011. Influence of grape density at harvest date on changes in phenolic composition, phenol extractability indices, and instrumental texture properties during ripening. J. Agric. Food Chem. 59, 8796–8805.

Rolle, L., Torchio, F., Lorrain, B., Giacosa, S., Río Segade, S., Cagnasso, E., Gerbi, V., Teissedre, P.L., 2012a. Rapid methods for the evaluation of total phenol content and extractability in intact grape seeds of Cabernet-Sauvignon: Instrumental mechanical properties and FT-NIR spectrum. J. Int. Sci. Vigne Vin 46, 29–40.

Rolle, L., Torchio, F., Ferrandino, A., Guidoni, S., 2012b. Influence of wine-grape skin hardness on the kinetics of anthocyanin extraction. Int. J. Food Prop. 15, 249–261.

Rolle, L., Giacosa, S., Torchio, F., Perenzoni, D., Río Segade, S., Gerbi, V., Mattivi, F., 2013. Use of instrumental acoustic parameters of winegrape seeds as possible predictors of extractable phenolic compounds. J. Agric. Food Chem. 61, 8752–8764.

Scott, I.M., Lin, W., Liakata, M., Wood, J.E., Vermeer, C.P., Allaway, D., Ward, J.L., Draper, J., Beale, M.H., Corol, D.I., Baker, J.M., King, R.D., 2013. Merits of random forests emerge in evaluation of chemometric classifiers by external validation. Anal. Chim. Acta 801, 22–33.

Therneau, T., Atkinson, B., Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-9. <http://www.CRAN.R-project.org/package=rpart>.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8, 25.

Yao, L.H., Jiang, Y.M., Shi, J., Tomás-Barberán, F.A., Datta, N., Singanusong, R., 2004. Flavonoids in food and their health benefits. Plant Foods Hum. Nutr. 59, 113–122.

Zouid, I., Siret, R., Jourjon, F., Mehinagic, E., Rolle, L., 2013. Impact of grapes heterogeneity according to sugar level on both physical and mechanical berries properties and their anthocyanins extractability at harvest. J. Texture Stud. 44, 95–103.