Evaluation of a robust regression method (RoBoost-PLSR) to predict biochemical variables for agronomic applications: Case study of grape berry maturity monitoring

Aldrig Courand, Maxime Metz, Daphné Héran, Carole Feilhes, Fanny Prezman, Eric Serrano, Ryad Bendoula, Maxime Ryckewaert

Evaluation of a robust regression method (RoBoost-PLSR) to predict biochemical variables for agronomic applications: case study of grape berry maturity monitoring

Aldrig Courand[c], Maxime Metz[a,b], Daphné Héran[a], Carole Feilhes[d], Fanny Prezman[d], Eric Serrano[d], Ryad Bendoula[a], Maxime Ryckewaert[a,b]

[a] *ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*
[b] *ChemHouse Research Group, Montpellier, France*
[c] *UBO,Brest,France*
[d] *IFV, 1920 Route de Lisle-sur-Tarn, 81310 Peyrole, France*

Abstract

Visible and near infrared spectroscopy (VIS-NIR) is increasingly being transferred from laboratory to industry for in-line and portable applications in various domains. By intensively using VIS-NIR spectroscopy, some abnormal observations may certainly arise. It is then important to properly handle outliers to elaborate effective prediction models. The objective of this study is to investigate the potential of using a robust method called Roboost-PLSR to improve prediction model performances for a viticulture application. This work focuses on a case study to predict sugar content in grape berries of three different grape varieties of Vitis Vinifera in a maturity monitoring context. Hyperspectral images were acquired of grape berries of Syrah, Fer-Servadou and Mauzac varieties. Reference measurements of sugar levels were made in the laboratory by densimetric baths. Performances of RoBoost-PLSR models were compared to performances of reference models

using Partial Least Square Regression (PLSR). Reference prediction criteria using PLSR were obtained for all varieties with these following values: Syrah ($R^2_p = 0.971$; $RMSE_p$= 5.36 g/L), Fer-servadou ($R^2_p = 0.788$; $RMSE_p$= 11.69 g/L) and Mauzac ($R^2_p = 0.690$; $RMSE_p = 15.61$ g/L). Prediction qualities are improved with RoBoost-PLSR: Syrah ($R^2_p = 0.990$; $RMSE_p$=3.14 g/L), Fer-Servadou ($R^2_p = 0.848$; $RMSE_p$= 10.20 g/L) and Mauzac ($R^2_p = 0.927$; $RMSE_p$= 7.58 g/L). Results confirm that Roboost-PLSR method allows a better consideration of outliers within the calibration set.

Keywords: Robust regression, Chemometrics, Spectroscopy, Grapes, maturity

---

1. Introduction

It is increasingly common that visible and near-infrared (VIS-NIR) spectroscopy transfers from laboratory to industry for in-lign and portable applications in various domains. By intensively using VIS-NIR spectroscopy, some abnormal observations may certainly arise. Among these, observations are called leverage points when they have a strong impact on the construction of a prediction model. When they are detrimental to the prediction model, they are called outliers. It is then important to properly handle these outliers to elaborate effective prediction models. In chemometrics, Partial Least Square Regression (PLSR) (Wold et al., 2001) is a widely-used tool. Particularly, PLSR is effective when dealing with high-dimensional data such as spectral data, where the sample number is lower than variable number. Besides, the PLSR method performs admirably when the relationship between explanatory variables and response variable to be predicted is linear.

However, estimating this linear relationship may be disturbed in presence of outliers (Serneels et al., 2005a).

These outlier data are generally due to variations of measurement conditions (view angle, reference, sensor temperature), physico-chemical variations in measured samples or experimental errors (annotation, operator). All these variations require efforts to identify and remove outliers from the calibration set. In addition, inspecting each observation manually is complicated and time-consuming in the case of large databases.

These problems are also found in agronomy, where the use of VIS-NIR spectroscopy is tending to be more frequently used (Ryckewaert et al., 2021). Indeed, rich spectral information is an added value to predict biochemical variables to assess agronomic parameters for various agronomic applications. This technological trend operates at different scales depending on the objectives: prediction models can be used at fruit scale for quality control, at the leaf/canopy scale for plant health monitoring or at the plot scale for production monitoring. Multiple use cases of spectral data encourage a particular development on the management of outliers.

Robust methods have been developed to address this issue (Serneels et al., 2005b; Hubert and Branden, 2003; Filzmoser et al., 2008, 2020; Griep et al., 1995; Metz et al., 2021). Indeed, this type of method aims at reducing the outlier impact automatically on PLSR model calibration. Recently, a method called Roboost-PLSR has been developed (Metz et al., 2021) and has shown its effectiveness to manage PLSR model calibration in the presence of outlier data.

This article highlights the interest of RoBoost-PLSR method to improve

prediction models for agronomic applications and more particularly in the case of monitoring grape berry maturity of Vitis Vinifera. For this purpose, Roboost-PLSR method was compared to the reference method PLSR to predict sugar content in grape berries of three different grape varieties.

2. Materials and methods

2.1. Biological material and reference measurements

Grape berries were collected during a campaign carried out in Gaillac (France), in summer 2020. The sampling started one or two weeks after veraison and preharvest, on three plots corresponding to three different grape varieties of the experimental vineyard Domaine Expérimental Viticole Tarnais: with two red grape varieties (Syrah and Fer Servadou) and one white grape variety (Mauzac). Thirty bunches were randomly sampled in each plot about once a week.

Figure 1: Picture of densimetric baths used for maturity degree sorting of grape berries.

In the laboratory, grape berries were cut from bunches at the pedicel level to preserve entire fruits. Grape berries were then sorted in batches with same

maturity degree using sodium chloride (NaCl) baths to achieve a densimetric sorting (see fig. 1). Indeed, the increase in berry density during ripening is mainly due to sugar accumulation in berries (Lanier and Morris, 1978a,b). To this end, twelve NaCl baths with increasing concentrations from 70 to 190 g/L were used to classify berry density corresponding to sugar concentrations from 110 to 279 g/L (Bigard, 2018). First, berries were immersed in the highest NaCl concentration solution. Then, floating fruits were removed and immersed in a solution of lower concentration, whereas sinking fruits were removed and sorted into the density level corresponding to the NaCl solution. The procedure was repeated for all baths in order to obtain twelve classes of homogeneous maturity. Sugar content measurements were performed on berry musts (one must corresponds to one hundred berries) with a refractometer (HI-96816, Hanna Instruments).

2.2. Spectral acquisition

Before preparing a hundred berry must, these berries were placed on a tray for spectral acquisition. Reflectance spectra were acquired with a hyperspectral camera (Specim IQ, Specim, Finland) having a spectral range from 400 nm to 1000 nm and a spectral resolution equal to 7 nm (see Fig 2).

Figure 2: Hyperspectral acquisition of grape berries.

For each sample, reflected light intensity ($I_s(\lambda)$) was measured at each wavelength λ. The camera was positioned 1.5 m from the scene. Dark current image ($I_b(\lambda)$) was also recorded for each measure. A certified reflectance standard (Labsphere, SRS-40-010) was used as a reference reflected intensity ($I_o(\lambda)$) to standardise images from non-uniformities of instrumentation (light source, lens, detector). Illumination was provided using a halogen lamp (Arrilite 750 Plus ARRI, Munich, Germany). Constant angles of -50° and 50° were maintained between the halogen lamp axes and the hyperspectral camera axis. From these measurements, a reflectance image ($R_s(\lambda)$) was obtained for each sample where each pixel of this image is a reflectance spectrum:

$$R_s(\lambda) = \frac{I_s(\lambda) - I_b(\lambda)}{I_0(\lambda) - I_b(\lambda)} \quad (1)$$

2.3. Image preprocessing

A segmentation process was implemented to retrieve berry reflectance spectra from images. First, three reference spectra were defined, corresponding to each grape variety, by calculating an averaged spectrum from a manual selection of an area of a berry. Then, the segmentation was performed by comparing each image pixel with these previously defined spectra (see fig. 3).

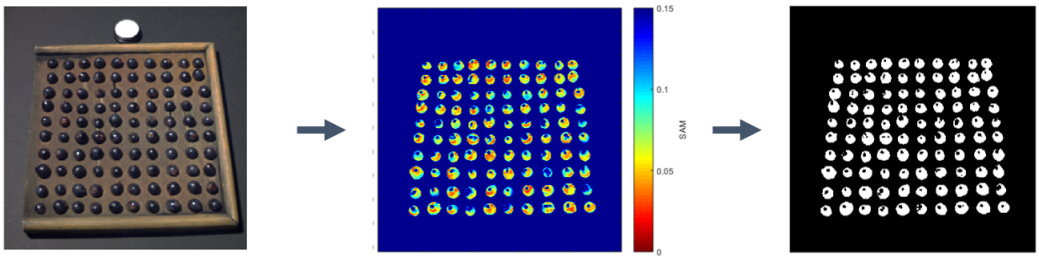


Figure 3: Segmentation by using spectral similarity threshold.

To this end, Spectral Angle Mapper (SAM) (Kruse et al., 1993; Yuhas et al., 1992) was selected to evaluate spectral similarity between the reference spectrum defined for a given grape variety and spectra contained in hyperspectral images. Indeed, this criterion corresponds to an angle between two spectra (assimilated to vectors) and is favourably independent to intensity levels. The angle α defined between the corresponding variety reference spectrum $\mathbf{y}$ and the spectrum of a given pixel $\mathbf{x}$, was calculated as follows:

$$\alpha = cos^{-1} \frac{\sum_{\lambda} \mathbf{xy}}{\sqrt{\sum(\mathbf{x})^2 \sum(\mathbf{y})^2}} \tag{2}$$

By defining a spectral similarity threshold, berry spectra were retrieved from the images (see fig. 3). Finally, for each image a berry average spectrum was computed, to consider a unique sugar content.

2.4. Data analysis

2.4.1. Regression methods used

RoBoost-PLSR (Metz et al., 2021) was used as a robust regression method to predict sugar content $\mathbf{Y}$ from spectral data $\mathbf{X}$. The purpose of this method is to define the outlyingness for each individual. This measure is expressed as

a weight which is integrated in the calibration of the RoBoost-PLSR model. This methods reduces outlier effect on model calibration by weighting them. A particularity of this method is that outliers are defined latent variable by latent variable. For each model with one latent variable, observation weights are calculated according to three criteria: $\mathbf{X}$ residuals, $\mathbf{Y}$ residuals and leverage points with the hyperparameters α, β and γ respectively. In this study, sixty-four combinations of values for α, β and γ were tested to optimise the model with these following possible values: 2, 4, 6, and infinite. RoBoost-PLSR was compared to the reference regression method PLSR (Wold et al., 2001).

Calculations were performed with the R software (version 3.6.1 (Core Team, 2013)), rnirs package for PLSR (`https://github.com/mlesnoff/rnirs`) and roboost package for RoBoost-PLSR (`https://github.com/maxmetz/RoBoost-PLSR`).

2.4.2. Calibration and test set definition

To compare PLSR method with RoBoost-PLSR method, models were established from three data sets corresponding to the three different grape varieties. For each grape variety, data were split into two sets, one calibration set and one test set. The calibration set was formed with 75% of the whole data set whereas the test set was formed with the remaining 25%. This partitioning was chosen in order to have a sufficient amount of data to evaluate the criteria on the test set. As showed in table 1, the total number of observations was different depending on the grape variety.

Table 1: Number of observations constituting the whole data set, the calibration set and the test set, for the three grape varieties, Syrah, Fer and Mauzac.

Number of observations	**Syrah**	**Fer**	**Mauzac**
Whole dataset	126	63	85
Calibration set	95	48	67
Test set	31	15	18

Besides, test sets were created avoiding abnormal observations according to (Metz et al., 2021).

2.4.3. Assessment criteria

PLSR models were calibrated by performing a cross-validation procedure (Browne, 2000). For each grape variety, a k-fold cross-validation with five blocks was defined on the corresponding calibration data set.

Model evaluation was performed using several criteria: root-mean-square error (RMSE), median absolute deviation (MAD) and determination coefficient R^2. Besides, the number of latent variables was optimised thanks to the RMSE parameter and was chosen to be lower than twenty. These criteria were calculated thanks to the following equations:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{3}$$

$$\text{MAD} = median(|y_i - \tilde{y}|) \tag{4}$$

$$R^2 = 1 - \frac{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}{\frac{\sum_{i=1}^{N}(y_i - y_m)^2}{N}} \quad (5)$$

with $\hat{y}_i$ the predicted value, y_i the observed value, y_m the average of all response values and N the total number of observations. $RMSE_{cv}$, MAD_{cv} and R^2_{cv} denoted criteria obtained in the cross-validation step whereas $RMSE_p$, MAD_p and R^2_p denoted those obtained with the independent test set.

Likewise PLSR, RoBoost-PLSR models were calibrated by performing a k-fold cross-validation procedure with five blocks. However, so-called robust evaluation criteria were calculated by using a procedure of trimming (Filzmoser and Nordhausen, 2021) Trimming consisted in sorting out observations according to their weights before removing a percentage of observations having the weaker weights. Moreover, this percentage was adapted to each of the three grape varieties: 5% for Syrah, 15% for Fer and 20% for Mauzac. Among these new criteria, r-$RMSE_{cv}$ and r-R^2_{cv} were defined, corresponding respectively to the trimmed RMSE and the trimmed coefficient of determination. The MAD calculated previously (eq. 4) was retained as it is considered a criterion for evaluating robustness.

So-called robust evaluation criteria were chosen according to Filzmoser and al work (Filzmoser and Nordhausen, 2021). MAD, considered as a robustness evaluation criterion, was computed. r-$RMSE_{cv}$ and r-R^2_{cv} were computed as follows:

$$\text{r-RMSE}_{cv} = \sqrt{\frac{\sum_{i=1}^{N_t}(\hat{y}_i - y_i)^2}{N_t}} \quad (6)$$

$$\text{r-R}^2_{cv} = 1 - \frac{\frac{\sum_{i=1}^{N_t}(\hat{y}_i - y_i)^2}{N_t}}{\frac{\sum_{i=1}^{N_t}(y_i - y_m)^2}{N_t}} \quad (7)$$

With $\hat{y}_i$ the predicted y, y_i the observed y, y_m the average y and N_t the number of retained observations. The r-RMSE was chosen as the criterion to minimise during cross-validation.

3. Results and discussion

3.1. Data visualization

Sugar content distributions measured on grape berries of the three varieties (Fer Servadou, Mauzac and Syrah) can be seen in Figure 4.

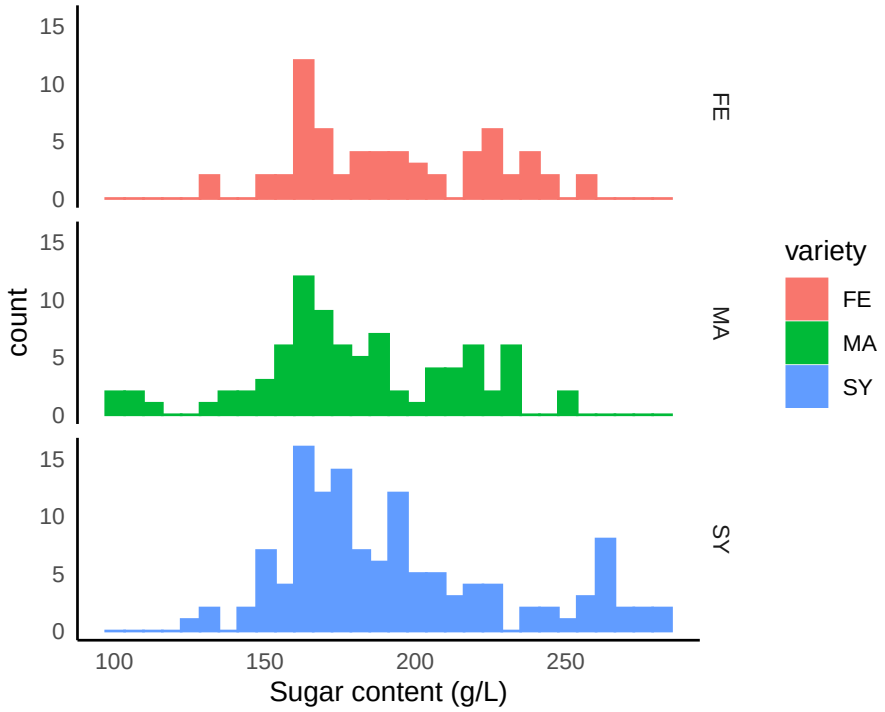


Figure 4: Sugar content (g/L) histograms for the three grape varieties: Fer Servadou (FE), Mauzac (MA) and Syrah (SY)

For the three varieties, sugar content values are similar and comprised between 100 and 300 g/L. Most values lie between 150 and 200 g/L which correspond to expected sugar contents for grape berries at different maturity stages. As sugar content values cover the same range for the three varieties, comparing results obtained for each grape variety is relevant.

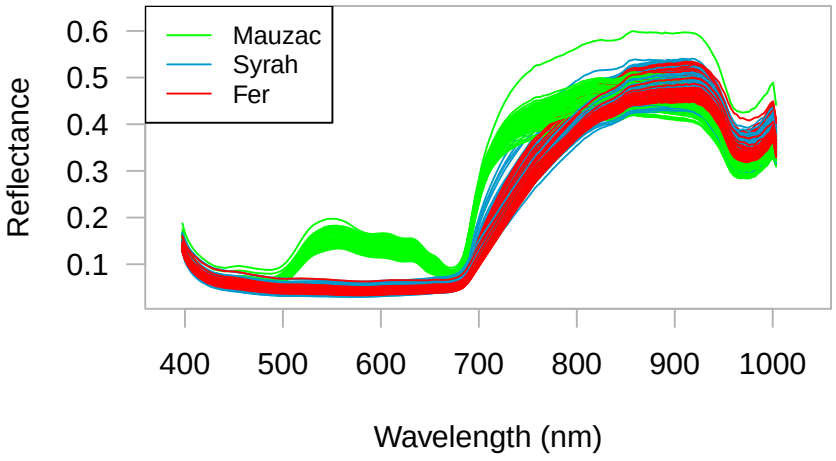


Figure 5: Reflectance spectra of the whole data set

Reflectance spectra comprised between 400 nm and 1000 nm of the whole data set are shown in figure 5. The two varieties Syrah and Fer Servadou are similar over the whole spectral range. However, Mauzac spectra differ from the two other varieties. Mainly, reflectance values are higher in the spectral range comprised between 500 nm and 680 nm. Moreover, the spectrum slope is steeper around 700 nm.

Syrah and Fer Servadou are red grape varieties and are known to possess high anthocyanin contents. Besides, visible light is largely absorbed by anthocyanins which causes low reflectance values between 500 nm and 700 nm, as can be seen on spectra for these two varieties. Spectra visualisation confirms the establishment of prediction models by variety.

3.2. Prediction models

3.2.1. PLSR models

Table 2 presents the values of the four criteria, latent variable number (nLV), prediction error ($RMSE_{cv}$), median (MAD_{cv}) and determination coefficient (R^2_{cv}), based on the cross-validation of the three grape variety PLSR models.

Table 2: Selected criteria obtained for cross-validation of PLSR prediction models on calibration data set: latent variable number (nLV), prediction error ($RMSE_{cv}$), median (MAD_{cv}) and determination coefficient (R^2_{cv})

Model	Variety	nLV	$RMSE_{cv}$ (g/L)	MAD_{cv} (g/L)	R^2_{cv}
PLSR	Syrah	6	9.31	8.09	0.937
	Fer Servadou	7	19.45	15.84	0.623
	Mauzac	5	28.78	18.40	0.298

Results show large disparities between grape varieties. Indeed, Syrah has the best results with a higher R^2_{cv} of 0.937 and lower $RMSE_{cv}$ and MAD_{cv} ,of respectively 9.31 g/L and 8.09 g/L. For Fer Servadou variety, $RMSE_{cv}$ and MAD_{cv} have values equal to 19.45 g/L and 15.84 g/L, which are nearly twice as large as the Syrah values. For Mauzac variety, $RMSE_{cv}$ value is equal to 28.78 g/L and MAD_{cv} value is 18.40 g/L. These values are two to three times higher than the ones obtained for Syrah.

Likewise, determination coefficient values differ between the three grape varieties. R^2_{cv} obtained for Fer Servadou and Mauzac varieties are equal to 0.623 and 0.298 respectively, much lower than Syrah result, especially for Mauzac. High discrepancies can be seen among the three grape varieties.

3.2.2. *RoBoost-PLSR models*

Table 3: Selected criteria obtained for cross-validation of RoBoost-PLSR prediction models on calibration data set: trimming, hyperparameters, latent variable number (nLV), prediction error (r-RMSE$_{cv}$), median (MAD$_{cv}$) and determination coefficient (r-R$^2_{cv}$)

Model	Variety	Trimming	Hyperparameters (α; β; γ)	nLV	r-RMSE$_{cv}$ (g/L)	MAD$_{cv}$ (g/L)	r-R$^2_{cv}$
RoBoost-PLSR	Syrah	5%	Inf; 4; 6	6	8.57	6.86	0.951
	Fer	15%	Inf; 4; Inf	7	12.5	14.3	0.844
	Mauzac	20%	Inf; 4; 6	6	12.1	15.50	0.794

The table 3 shows parameters from cross validation of RoBoost-PLSR method. These parameters are trimming percentage, hyperparameters (α, β, γ), latent variable number, r-RMSE$_{cv}$, MAD$_{cv}$ and r-R$^2_{cv}$. Hyperparameter values α, β and γ are respectively equal to infinite, 4, 6 for Syrah; infinite, 4, infinite for Fer Servadou; and infinite, 4, 6 for Mauzac. Hyperparameters α, β and γ are selective criteria for outlier detection respectively on $\mathbf{X}$, $\mathbf{Y}$ and leverage points. The lower the hyperparameter, the higher the outlier number identified by the model. Conversely, an infinite value means no outlier identified. This implies that there is no outlier detected by cross-validation on $\mathbf{X}$ for the three grape varieties ($\alpha = \text{Inf}$). However, this is not the case for $\mathbf{Y}$ (i.e. measures of sugar content), where $\beta = 4$ for the three grape varieties and means that several outliers are detected. Indeed, outliers could be introduced during sugar content measurements by densimetric bath. Finally, based on hyperparameter γ values, no leverage point is identified for Fer Servadou variety whereas some are detected for Mauzac and Syrah.

Among the three grape varieties, Syrah obtains the best results with a r-RMSE$_{cv}$ equals to 8.57 g/L which corresponds to the lowest value. Further-

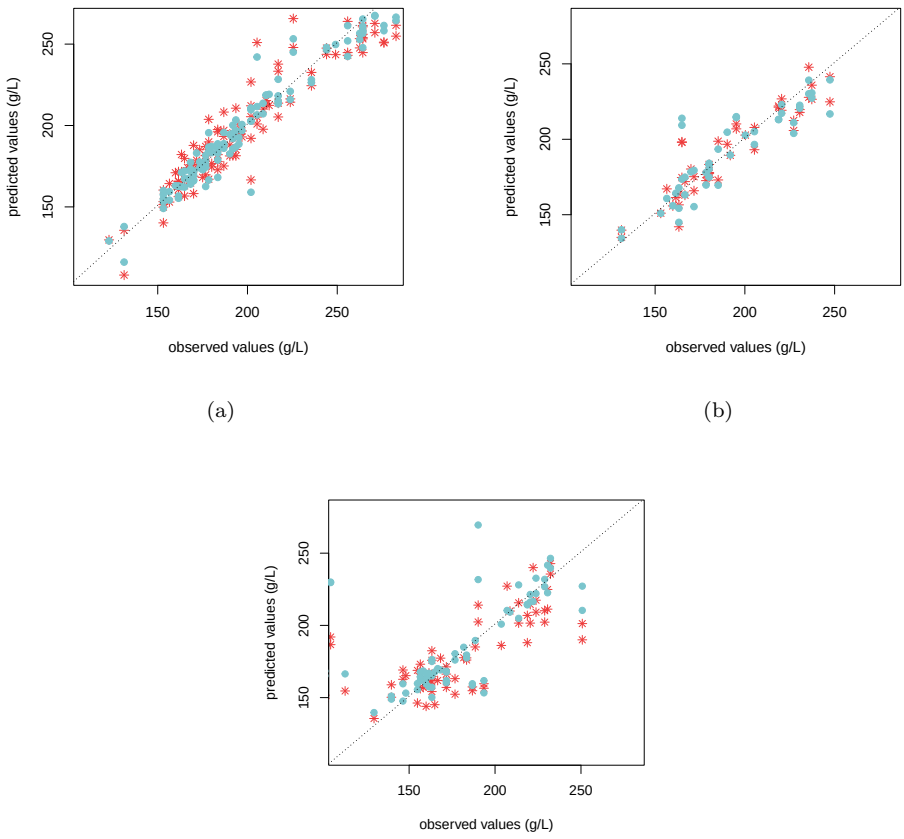
more, this value is slightly lower than the one obtained with the PLSR model (see table 2). Regarding Fer Servadou and Mauzac varieties, r-RMSE$_{cv}$ values are close to each other with values equal to 12.5 g/L and 12.1 g/L respectively. These results are improved compared to the values previously obtained with PLSR cross-validation (see table 2) and closer to Syrah value. Indeed, during the cross-validation procedure, RoBoost-PLSR deals with outliers by attributing weights to observations.

Besides, the same analysis can be done for r-R^2_{cv} values. Syrah obtains the best value with 0.951 whereas Fer Servadou and Mauzac obtain 0.844 and 0.764. Again, these values are lower and closer to each other than the ones previously obtained with PLSR cross-validation (see table 2).

Finally, the comparison of both cross-validation results, PLSR (table 2) and RoBoost-PLSR (table 3), indicates that RoBoost-PLSR decreases the prediction quality discrepancies between grape varieties. This result confirms the presence of outlier points among Fer Servadou and Mauzac data sets.

3.2.3. Observed vs. predicted values of calibration models

The visualisation of observed values by predicted values shown in figure 6 helps to better understand criteria values obtained in cross-validation (tab 2 and 3). It provides a means to assess model quality, observation by observation.



(a)

(b)

(c)

Figure 6: Sugar content observed values versus predicted values based on (*, red) PLSR and (•, blue) RoBoost-PLSR for the three grape varieties: (a) Syrah, (b) Fer, (c) Mauzac

Figures 6a, 6b and 6c compare predicted values of the calibration data set obtained with RoBoost-PLSR and PLSR for Syrah, Fer and Mauzac respectively.

Regarding Syrah variety (fig. 6a), relationship between predicted $\mathbf{Y}$ and observed $\mathbf{Y}$ is linear and point dispersion is the lowest obtained for the three

varieties and this with RoBoost and PLSR. The same holds true for Fer Servadou variety (fig. 6b), where a linear tendency between predicted $\mathbf{Y}$ and observed $\mathbf{Y}$ can be noticed. However, several points obtained with PLSR deviate from this tendency. These same points are further deviated from the linear trend with RoBoost-PLSR. The identified points deviating from the linear tendency are possibly outliers (also called vertical outliers) or leverage points.

As far as Mauzac is concerned (fig. 6c), the relationship between predicted $\mathbf{Y}$ and observed $\mathbf{Y}$ deviates from a linear tendency with several points strongly dispersed. Some points deviate more strongly from this trend than previously. These same points are even further apart with RoBoost-PLSR, while an improvement appears on the majority of the other points. These points are clearly identified by the RoBoost-PLSR method as vertical outliers or leverage points. These points are weighted when building the prediction model with RoBoost-PLSR. RoBoost-PLSR thus improves the linearity between predicted and observed values.

By comparing these three figures (6a, 6b and 6c), calibration data set which have the best predictions are Syrah first, then Fer Servadou and finally Mauzac. This confirms the results obtained in cross-validation (table 3).

3.3. Model prediction on independent test sets

For each grape variety, PLSR and RoBoost-MLSR models previously parameterized during cross-validation steps and calibrated with calibration data sets are now tested on the test data sets.

Table 4: Performance evaluation of PLSR and RoBoost-PLSR prediction models on test data sets: latent variable number (nLV), prediction error ($RMSE_p$), median (MAD_p) and determination coefficient (R^2_p)

Model	Variety	nLV	$RMSE_p$ (g/L)	MAD_p (g/L)	R^2_p
PLSR	Syrah	6	5.36	4.99	0.971
	Fer Servadou	7	11.69	12.04	0.788
	Mauzac	5	15.61	10.97	0.690
RoBoost PLSR	Syrah	6	3.14	3.38	0.990
	Fer Servadou	7	10.20	10.50	0.848
	Mauzac	6	7.58	9.36	0.927

Table 4 outlines the prediction quality of both PLSR and RoBoost-PLSR models, applied to the test data sets of each grape variety. To this end, the following criteria are presented: latent variable number (nLV), $RMSE_p$, MAD_p and R^2_p.

First of all, a higher heterogeneity among results can be noticed for PLSR models than for RoBoost-PLSR ones. Regarding PLSR models, Syrah has the best performances, with the lowest $RMSE_p$ and MAD_p values, equal to 5.36 g/L and 4.99 g/L respectively, and the highest R^2_p value, equals to 0.971. Fer Servadou and Mauzac have $RMSE_p$ and MAD_p values, two to three times higher than Syrah ones. $RMSE_p$ are equal to 11.69 g/L and 15.61 g/L for Fer and Mauzac respectively, whereas MAD_p values are 12.04 g/L and 10.97 g/L respectively. Moreover, R^2_p are lower than for Syrah, with respective values of 0.788 and 0.690. As said before during cross-validation step (section 3.2.1), discrepancies among varieties arise with PLSR models.

As far as RoBoost-PLSR models are concerned, all three varieties predictions are improved compared to PLSR models. This is all the more true in the case of Mauzac and Syrah. Indeed, Syrah obtains R^2_p, $RMSE_p$ and MAD_p values equal to 0.990, 3.14 g/L and 3.38 g/L respectively. Besides, Fer Servadou obtains R^2_p, $RMSE_p$ and MAD_p values equal to 0.848, 10.20 g/L and 10.50 g/L. Lastly, Mauzac obtains R^2_p, $RMSE_p$ and MAD_p values equal to 0.927, 7.58 g/L and 9.36 g/L. These last results outperform PLSR models and lead to performances close to Syrah ones.

It is worth noticing that PLSR model allows to predict sugar content for Syrah in an effective way. This implies that there is a limited number of outlier points in the data set. The same does not hold true for Fer Servadou and Mauzac, as noticed in figure 6. In all cases, RoBoost-PLSR method allows to build predictive models with higher performances than PLSR when dealing with outliers points among calibration data sets.

4. Conclusion

The potential of RoBoost-PLSR method to calibrate prediction models in the presence of outliers in an agronomic context was studied. The method was evaluated on a case of Vitis Vinifera grapes berry maturity context and especially to predict berry sugar content. RoBoost-PLSR method was compared to the reference method (PLSR) on spectral data from berries of three grape varieties (Syrah, Mauzac, Fer Servadou). For these three varieties, results obtained from RoBoost-PLSR method outperformed those from the PLSR method. The improvements in the prediction of sugar content for Fer Servadou and Mauzac are the most significant due to a potentially higher

outliers number in the calibration set.

This study validates the use of the RoBoost-PLSR method for monitoring grapes berries maturity in the laboratory. The advantage of this method is to provide good prediction models despite outliers presence. Despite optimal measurement conditions, outliers were identified as detrimental to the model calibration. This method could be challenged on data collecting directly in the field where measurement conditions most often lead to outliers. This would open up multiple possibilities for the use of VIS-NIR spectroscopy for agronomic applications. Other robust methods could be compared ti RoBoost-PLSR in such an application context. This method also contributes to perspectives in other disciplines where multivariate data is involved such as analytical chemistry.

Acknowledgement

This work has benefited from a financial support from the InterregSudoe under the reference SOE3/P2/E0911.

References

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

Sven Serneels, Christophe Croux, Peter Filzmoser, and Pierre J. Van Espen. Partial robust m-regression. *Chemometrics and Intelligent Laboratory Systems*, 79(1):55–64, 2005a. ISSN 0169-7439. doi: https://doi.org/10.

1016/j.chemolab.2005.04.007. URL https://www.sciencedirect.com/science/article/pii/S0169743905000638.

Maxime Ryckewaert, Maxime Metz, Daphné Héran, Pierre George, Bruno Grèzes-Besset, Reza Akbarinia, Jean-Michel Roger, and Ryad Bendoula. Massive spectral data analysis for plant breeding using parSketch-PLSDA method: Discrimination of sunflower genotypes. *Biosystems Engineering*, 210:69–77, October 2021. doi: 10.1016/j.biosystemseng.2021.08.005.

Sven Serneels, Christophe Croux, Peter Filzmoser, and Pierre J. Van Espen. Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 79(1):55–64, October 2005b. ISSN 0169-7439. doi: 10.1016/j.chemolab.2005.04.007. URL https://www.sciencedirect.com/science/article/pii/S0169743905000638.

M. Hubert and K. Vanden Branden. Robust methods for partial least squares regression. *Journal of Chemometrics*, 17 (10):537–549, 2003. ISSN 1099-128X. doi: https://doi.org/10.1002/cem.822. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.822. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.822.

Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52 (3):1694–1711, January 2008. ISSN 0167-9473. doi: 10.1016/j.csda.2007.05.018. URL https://www.sciencedirect.com/science/article/pii/S0167947307002204.

Peter Filzmoser, Sven Serneels, Ricardo Maronna, and Christophe Croux. Robust Multivariate Methods in Chemometrics. In *Comprehensive Chemometrics*, pages 393–430. Elsevier, 2020. ISBN 978-0-444-64166-3. doi: 10.1016/B978-0-12-409547-2.14642-6. URL https://linkinghub.elsevier.com/retrieve/pii/B9780124095472146426.

M. I. Griep, I. N. Wakeling, P. Vankeerberghen, and D. L. Massart. Comparison of semirobust and robust partial least squares procedures. *Chemometrics and Intelligent Laboratory Systems*, 29(1):37–50, July 1995. ISSN 0169-7439. doi: 10.1016/0169-7439(95)80078-N. URL https://www.sciencedirect.com/science/article/pii/016974399580078N.

Maxime Metz, Florent Abdelghafour, Jean-Michel Roger, and Matthieu Lesnoff. A novel robust PLS regression method inspired from boosting principles: RoBoost-PLSR. *Analytica Chimica Acta*, page 338823, July 2021. ISSN 00032670. doi: 10.1016/j.aca.2021.338823. URL https://linkinghub.elsevier.com/retrieve/pii/S0003267021006498.

Lanier and Morris. Density separation of muscadine grapes. *Arkansas Farm Research*, 1978a. URL https://scholar.google.com/scholar_lookup?title=Density+separation+of+muscadine+grapes.&author=Lanier+M.R.&publication_year=1978.

Lanier and Morris. Maturation rates of muscadine grapes. *Arkansas Farm Research*, 1978b. ISSN 0004-1785. URL https://scholar.google.com/scholar_lookup?title=Maturation+rates+of+muscadine+grapes.&author=Lanier+M.R.&publication_year=1978.

Antoine Bigard. *Varietal differences in solute accumulation and grape development*. phdthesis, Montpellier SupAgro, December 2018. URL https://tel.archives-ouvertes.fr/tel-02542686.

F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3):145–163, 1993. URL http://www.sciencedirect.com/science/article/pii/003442579390013N.

Roberta H. Yuhas, Alexander FH Goetz, and Joe W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. *Geography*, 1992. URL https://ntrs.nasa.gov/search.jsp?R=19940012238.

R. Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. *Available*, 2013.

Michael W Browne. Cross-Validation Methods. *Journal of Mathematical Psychology*, 44(1):108–132, March 2000. ISSN 0022-2496. doi: 10.1006/jmps.1999.1279. URL https://www.sciencedirect.com/science/article/pii/S0022249699912798.

Peter Filzmoser and Klaus Nordhausen. Robust linear regression for high-dimensional data: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(4):e1524, 2021. Publisher: Wiley Online Library.