



HAL
open science

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

► To cite this version:

Aldrig Courand, Maxime Metz, Daphné Héran, Carole Feilhes, Fanny Prezman, et al.. Evaluation of a robust regression method (RoBoost-PLSR) to predict biochemical variables for agronomic applications: Case study of grape berry maturity monitoring. *Chemometrics and Intelligent Laboratory Systems*, 2022, 221, <10.1016/j.chemolab.2021.104485>. <hal-03538442>

HAL Id: hal-03538442

<https://hal.inrae.fr/hal-03538442v1>

Submitted on 21 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

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

5 Aldrig Courand^c, Maxime Metz^{a,b}, Daphné Héran^a, Carole Feilhes^d, Fanny
6 Prezman^d, Eric Serrano^d, Ryad Bendoula^a, Maxime Ryckewaert^{a,b}

7 ^a*ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*

8 ^b*ChemHouse Research Group, Montpellier, France*

9 ^c*UBO, Brest, France*

10 ^d*IFV, 1920 Route de Lisle-sur-Tarn, 81310 Peyrole, France*

11 **Abstract**

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

25 using Partial Least Square Regression (PLSR). Reference prediction criteria
26 using PLSR were obtained for all varieties with these following values: Syrah
27 ($R_p^2 = 0.971$; $RMSE_p = 5.36$ g/L), Fer-servadou ($R_p^2 = 0.788$; $RMSE_p = 11.69$
28 g/L) and Mauzac ($R_p^2 = 0.690$; $RMSE_p = 15.61$ g/L). Prediction qualities
29 are improved with RoBoost-PLSR: Syrah ($R_p^2 = 0.990$; $RMSE_p = 3.14$ g/L),
30 Fer-Servadou ($R_p^2 = 0.848$; $RMSE_p = 10.20$ g/L) and Mauzac ($R_p^2 = 0.927$;
31 $RMSE_p = 7.58$ g/L). Results confirm that Roboost-PLSR method allows a
32 better consideration of outliers within the calibration set.

33 *Keywords:* Robust regression, Chemometrics, Spectroscopy, Grapes,
34 maturity

35 1. Introduction

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

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

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

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

66 Robust methods have been developed to address this issue ([Serneels et al.,](#)
67 [2005b](#); [Hubert and Branden, 2003](#); [Filzmoser et al., 2008, 2020](#); [Griep et al.,](#)
68 [1995](#); [Metz et al., 2021](#)). Indeed, this type of method aims at reducing the
69 outlier impact automatically on PLSR model calibration. Recently, a method
70 called Roboost-PLSR has been developed ([Metz et al., 2021](#)) and has shown
71 its effectiveness to manage PLSR model calibration in the presence of outlier
72 data.

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

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

78 **2. Materials and methods**

79 *2.1. Biological material and reference measurements*

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



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

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

89 maturity degree using sodium chloride (NaCl) baths to achieve a densimetric
90 sorting (see fig. 1). Indeed, the increase in berry density during ripening is
91 mainly due to sugar accumulation in berries (Lanier and Morris, 1978a,b).
92 To this end, twelve NaCl baths with increasing concentrations from 70 to
93 190 g/L were used to classify berry density corresponding to sugar concen-
94 trations from 110 to 279 g/L (Bigard, 2018). First, berries were immersed
95 in the highest NaCl concentration solution. Then, floating fruits were re-
96 moved and immersed in a solution of lower concentration, whereas sinking
97 fruits were removed and sorted into the density level corresponding to the
98 NaCl solution. The procedure was repeated for all baths in order to obtain
99 twelve classes of homogeneous maturity. Sugar content measurements were
100 performed on berry musts (one must corresponds to one hundred berries)
101 with a refractometer (HI-96816, Hanna Instruments).

102 *2.2. Spectral acquisition*

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

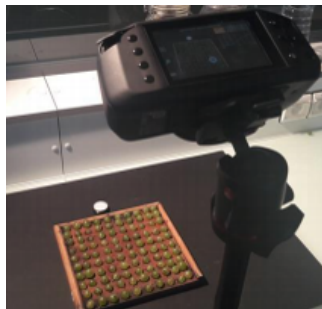


Figure 2: Hyperspectral acquisition of grape berries.

107 For each sample, reflected light intensity ($I_s(\lambda)$) was measured at each
 108 wavelength λ . The camera was positioned 1.5 m from the scene. Dark current
 109 image ($I_b(\lambda)$) was also recorded for each measure. A certified reflectance
 110 standard (Labsphere, SRS-40-010) was used as a reference reflected intensity
 111 ($I_o(\lambda)$) to standardise images from non-uniformities of instrumentation (light
 112 source, lens, detector). Illumination was provided using a halogen lamp
 113 (Arrilite 750 Plus ARRI, Munich, Germany). Constant angles of -50° and
 114 50° were maintained between the halogen lamp axes and the hyperspectral
 115 camera axis. From these measurements, a reflectance image ($R_s(\lambda)$) was
 116 obtained for each sample where each pixel of this image is a reflectance
 117 spectrum:

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

118 *2.3. Image preprocessing*

119 A segmentation process was implemented to retrieve berry reflectance
 120 spectra from images. First, three reference spectra were defined, correspond-
 121 ing to each grape variety, by calculating an averaged spectrum from a manual
 122 selection of an area of a berry. Then, the segmentation was performed by
 123 comparing each image pixel with these previously defined spectra (see fig.
 124 [3](#)).

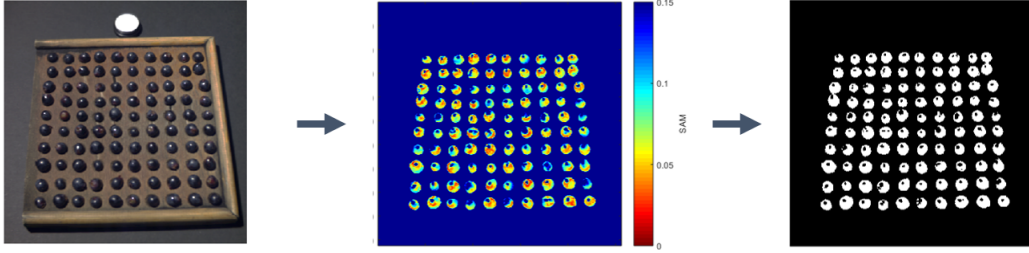


Figure 3: Segmentation by using spectral similarity threshold.

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

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

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

135 2.4. Data analysis

136 2.4.1. Regression methods used

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

140 a weight which is integrated in the calibration of the RoBoost-PLSR model.
141 This methods reduces outlier effect on model calibration by weighting them.
142 A particularity of this method is that outliers are defined latent variable
143 by latent variable. For each model with one latent variable, observation
144 weights are calculated according to three criteria: \mathbf{X} residuals, \mathbf{Y} residuals
145 and leverage points with the hyperparameters α , β and γ respectively. In this
146 study, sixty-four combinations of values for α , β and γ were tested to optimise
147 the model with these following possible values: 2, 4, 6, and infinite. RoBoost-
148 PLSR was compared to the reference regression method PLSR (Wold et al.,
149 2001).

150 Calculations were performed with the R software (version 3.6.1 (Core Team,
151 2013)), `rnirs` package for PLSR (<https://github.com/mlesnoff/rnirs>) and
152 `roboost` package for RoBoost-PLSR (<https://github.com/maxmetz/RoBoost-PLSR>).

153 *2.4.2. Calibration and test set definition*

154 To compare PLSR method with RoBoost-PLSR method, models were
155 established from three data sets corresponding to the three different grape
156 varieties. For each grape variety, data were split into two sets, one calibra-
157 tion set and one test set. The calibration set was formed with 75% of the
158 whole data set whereas the test set was formed with the remaining 25%.
159 This partitioning was chosen in order to have a sufficient amount of data to
160 evaluate the criteria on the test set. As showed in table 1, the total number
161 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

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

164 2.4.3. Assessment criteria

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

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

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

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

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

173 with \hat{y}_i the predicted value, y_i the observed value, y_m the average of all re-
 174 sponse values and N the total number of observations. RMSE_{cv} , MAD_{cv} and
 175 R_{cv}^2 denoted criteria obtained in the cross-validation step whereas RMSE_p ,
 176 MAD_p and R_p^2 denoted those obtained with the independent test set.

177 Likewise PLSR, RoBoost-PLSR models were calibrated by performing a
 178 k-fold cross-validation procedure with five blocks. However, so-called robust
 179 evaluation criteria were calculated by using a procedure of trimming ([Filz-
 180 moser and Nordhausen, 2021](#)) Trimming consisted in sorting out observations
 181 according to their weights before removing a percentage of observations hav-
 182 ing the weaker weights. Moreover, this percentage was adapted to each of
 183 the three grape varieties: 5% for Syrah, 15% for Fer and 20% for Mauzac.
 184 Among these new criteria, r- RMSE_{cv} and r- R_{cv}^2 were defined, corresponding
 185 respectively to the trimmed RMSE and the trimmed coefficient of determina-
 186 tion. The MAD calculated previously (eq. 4) was retained as it is considered
 187 a criterion for evaluating robustness.

188 So-called robust evaluation criteria were chosen according to Filzmoser
 189 and al work ([Filzmoser and Nordhausen, 2021](#)). MAD, considered as a ro-
 190 bustness evaluation criterion, was computed. r- RMSE_{cv} and r- R_{cv}^2 were com-
 191 puted as follows:

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

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

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

195 **3. Results and discussion**

196 *3.1. Data visualization*

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

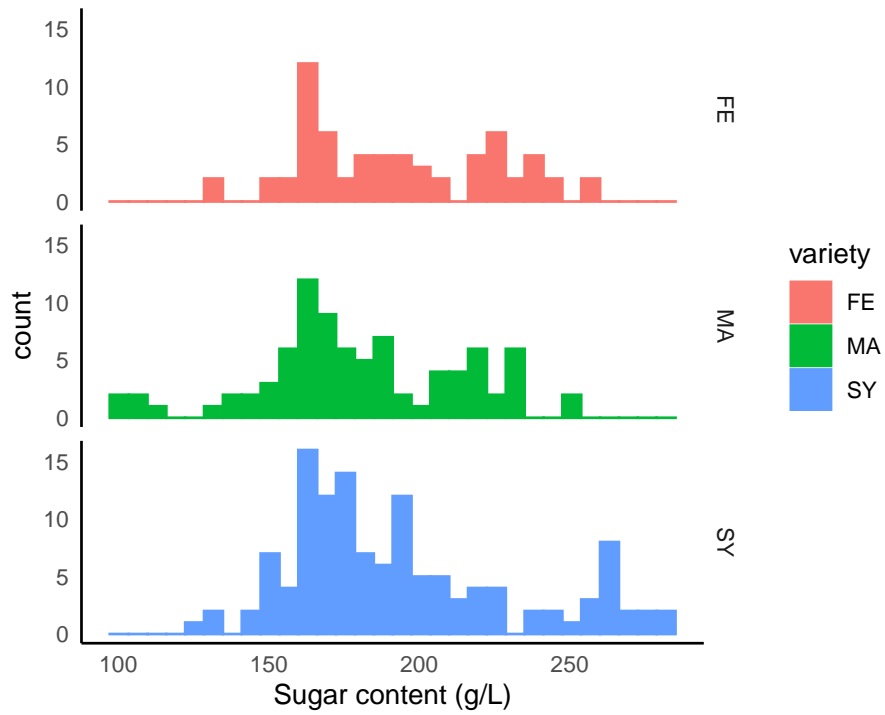


Figure 4: Sugar content (g/L) histograms for the three grape varieties: Fer Servadou (FE), Mauzac (MA) and Syrah (SY)

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

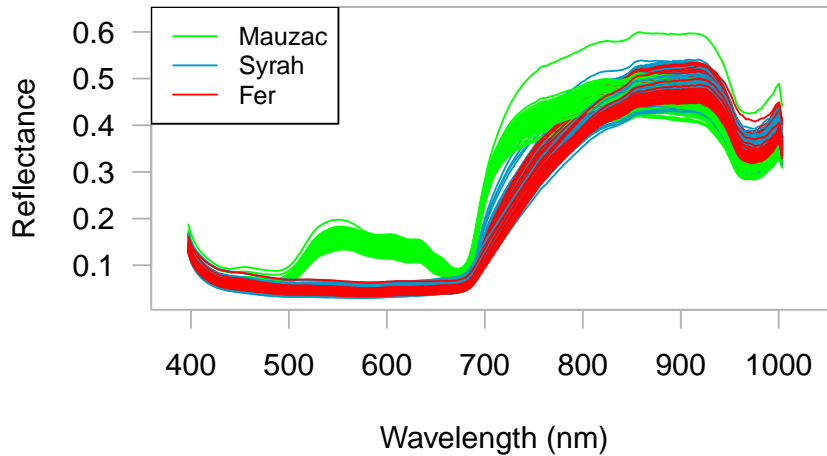


Figure 5: Reflectance spectra of the whole data set

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

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

215 *3.2. Prediction models*

216 *3.2.1. PLSR models*

217 Table 2 presents the values of the four criteria, latent variable number
 218 (nLV), prediction error (RMSE_{cv}), median (MAD_{cv}) and determination co-
 219 efficient (R_{cv}^2), based on the cross-validation of the three grape variety PLSR
 220 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_{cv}^2)

Model	Variety	nLV	RMSE_{cv} (g/L)	MAD_{cv} (g/L)	R_{cv}^2
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

221 Results show large disparities between grape varieties. Indeed, Syrah has
 222 the best results with a higher R_{cv}^2 of 0.937 and lower RMSE_{cv} and MAD_{cv} , of
 223 respectively 9.31 g/L and 8.09 g/L. For Fer Servadou variety, RMSE_{cv} and
 224 MAD_{cv} have values equal to 19.45 g/L and 15.84 g/L, which are nearly twice
 225 as large as the Syrah values. For Mauzac variety, RMSE_{cv} value is equal to
 226 28.78 g/L and MAD_{cv} value is 18.40 g/L. These values are two to three times
 227 higher than the ones obtained for Syrah.

228 Likewise, determination coefficient values differ between the three grape
 229 varieties. R_{cv}^2 obtained for Fer Servadou and Mauzac varieties are equal to
 230 0.623 and 0.298 respectively, much lower than Syrah result, especially for
 231 Mauzac. High discrepancies can be seen among the three grape varieties.

232 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_{cv}²)

Model	Variety	Trimming	Hyperparameters (α ; β ; γ)	nLV	r-RMSE _{cv} (g/L)	MAD _{cv} (g/L)	r-R _{cv} ²
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

233 The table 3 shows parameters from cross validation of RoBoost-PLSR
 234 method. These parameters are trimming percentage, hyperparameters (α , β ,
 235 γ), latent variable number, r-RMSE_{cv}, MAD_{cv} and r-R_{cv}². Hyperparameter
 236 values α , β and γ are respectively equal to infinite, 4, 6 for Syrah; infinite,
 237 4, infinite for Fer Servadou; and infinite, 4, 6 for Mauzac. Hyperparameters
 238 α , β and γ are selective criteria for outlier detection respectively on \mathbf{X} , \mathbf{Y}
 239 and leverage points. The lower the hyperparameter, the higher the outlier
 240 number identified by the model. Conversely, an infinite value means no
 241 outlier identified. This implies that there is no outlier detected by cross-
 242 validation on \mathbf{X} for the three grape varieties ($\alpha = \text{Inf}$). However, this is not
 243 the case for \mathbf{Y} (i.e. measures of sugar content), where $\beta = 4$ for the three
 244 grape varieties and means that several outliers are detected. Indeed, outliers
 245 could be introduced during sugar content measurements by densimetric bath.
 246 Finally, based on hyperparameter γ values, no leverage point is identified for
 247 Fer Servadou variety whereas some are detected for Mauzac and Syrah.

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

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

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

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

265 3.2.3. *Observed vs. predicted values of calibration models*

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

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

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

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

294 3.3. Model prediction on independent test sets

295 For each grape variety, PLSR and RoBoost-MLSR models previously pa-
296 rameterized during cross-validation steps and calibrated with calibration data
297 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_p^2)

Model	Variety	nLV	RMSE_p (g/L)	MAD_p (g/L)	R_p^2
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

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

302 First of all, a higher heterogeneity among results can be noticed for PLSR
 303 models than for RoBoost-PLSR ones. Regarding PLSR models, Syrah has
 304 the best performances, with the lowest RMSE_p and MAD_p values, equal to
 305 5.36 g/L and 4.99 g/L respectively, and the highest R_p^2 value, equals to 0.971.
 306 Fer Servadou and Mauzac have RMSE_p and MAD_p values, two to three times
 307 higher than Syrah ones. RMSE_p are equal to 11.69 g/L and 15.61 g/L for Fer
 308 and Mauzac respectively, whereas MAD_p values are 12.04 g/L and 10.97 g/L
 309 respectively. Moreover, R_p^2 are lower than for Syrah, with respective values of
 310 0.788 and 0.690. As said before during cross-validation step (section 3.2.1),
 311 discrepancies among varieties arise with PLSR models.

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

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

326 4. Conclusion

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

336 outliers number in the calibration set.

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

348 **Acknowledgement**

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

351 **References**

- 352 Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a ba-
353 sic tool of chemometrics. *Chemometrics and intelligent laboratory systems*,
354 58(2):109–130, 2001.
- 355 Sven Serneels, Christophe Croux, Peter Filzmoser, and Pierre J. Van Es-
356 pen. Partial robust m-regression. *Chemometrics and Intelligent Laboratory*
357 *Systems*, 79(1):55–64, 2005a. ISSN 0169-7439. doi: <https://doi.org/10.1002/cem.1000>

358 1016/j.chemolab.2005.04.007. URL [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/S0169743905000638)
359 [science/article/pii/S0169743905000638](https://www.sciencedirect.com/science/article/pii/S0169743905000638).

360 Maxime Ryckewaert, Maxime Metz, Daphné Héran, Pierre George, Bruno
361 Grèzes-Besset, Reza Akbarinia, Jean-Michel Roger, and Ryad Bendoula.
362 Massive spectral data analysis for plant breeding using parSketch-PLSDA
363 method: Discrimination of sunflower genotypes. *Biosystems Engineering*,
364 210:69–77, October 2021. doi: 10.1016/j.biosystemseng.2021.08.005.

365 Sven Serneels, Christophe Croux, Peter Filzmoser, and Pierre J. Van Es-
366 pen. Partial robust M-regression. *Chemometrics and Intelligent Labo-*
367 *ratory Systems*, 79(1):55–64, October 2005b. ISSN 0169-7439. doi: 10.
368 1016/j.chemolab.2005.04.007. URL [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/S0169743905000638)
369 [science/article/pii/S0169743905000638](https://www.sciencedirect.com/science/article/pii/S0169743905000638).

370 M. Hubert and K. Vanden Branden. Robust methods for par-
371 tial least squares regression. *Journal of Chemometrics*, 17
372 (10):537–549, 2003. ISSN 1099-128X. doi: [https://doi.org/](https://doi.org/10.1002/cem.822)
373 [10.1002/cem.822](https://doi.org/10.1002/cem.822). URL [https://analyticalsciencejournals.](https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.822)
374 [onlinelibrary.wiley.com/doi/abs/10.1002/cem.822](https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.822). eprint:
375 <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.822>.

376 Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identifica-
377 tion in high dimensions. *Computational Statistics & Data Analysis*, 52
378 (3):1694–1711, January 2008. ISSN 0167-9473. doi: 10.1016/j.csda.2007.
379 05.018. URL [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0167947307002204)
380 [S0167947307002204](https://www.sciencedirect.com/science/article/pii/S0167947307002204).

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

386 M. I. Griep, I. N. Wakeling, P. Vankeerberghen, and D. L. Massart. Com-
387 parison of semirobust and robust partial least squares procedures. *Chemo-*
388 *metrics and Intelligent Laboratory Systems*, 29(1):37–50, July 1995. ISSN
389 0169-7439. doi: 10.1016/0169-7439(95)80078-N. URL [https://www.
390 sciencedirect.com/science/article/pii/016974399580078N](https://www.sciencedirect.com/science/article/pii/016974399580078N).

391 Maxime Metz, Florent Abdelghafour, Jean-Michel Roger, and Matthieu
392 Lesnoff. A novel robust PLS regression method inspired from boosting
393 principles: RoBoost-PLSR. *Analytica Chimica Acta*, page 338823, July
394 2021. ISSN 00032670. doi: 10.1016/j.aca.2021.338823. URL [https://
395 //linkinghub.elsevier.com/retrieve/pii/S0003267021006498](https://linkinghub.elsevier.com/retrieve/pii/S0003267021006498).

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

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

- 404 Antoine Bigard. *Varietal differences in solute accumulation and grape devel-*
405 *opment*. phdthesis, Montpellier SupAgro, December 2018. URL [https:](https://tel.archives-ouvertes.fr/tel-02542686)
406 [//tel.archives-ouvertes.fr/tel-02542686](https://tel.archives-ouvertes.fr/tel-02542686).
- 407 F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T.
408 Shapiro, P. J. Barloon, and A. F. H. Goetz. The spectral image pro-
409 cessing system (SIPS)—interactive visualization and analysis of imag-
410 ing spectrometer data. *Remote sensing of environment*, 44(2-3):145–
411 163, 1993. URL [http://www.sciencedirect.com/science/article/](http://www.sciencedirect.com/science/article/pii/003442579390013N)
412 [pii/003442579390013N](http://www.sciencedirect.com/science/article/pii/003442579390013N).
- 413 Roberta H. Yuhas, Alexander FH Goetz, and Joe W. Boardman. Discrimi-
414 nation among semi-arid landscape endmembers using the spectral angle
415 mapper (SAM) algorithm. *Geography*, 1992. URL [https://ntrs.nasa.](https://ntrs.nasa.gov/search.jsp?R=19940012238)
416 [gov/search.jsp?R=19940012238](https://ntrs.nasa.gov/search.jsp?R=19940012238).
- 417 R. Core Team. R: A language and environment for statistical computing.
418 Vienna, Austria: R Foundation for Statistical Computing. *Available*, 2013.
- 419 Michael W Browne. Cross-Validation Methods. *Journal of Mathematical Psy-*
420 *chology*, 44(1):108–132, March 2000. ISSN 0022-2496. doi: 10.1006/jmps.
421 1999.1279. URL [https://www.sciencedirect.com/science/article/](https://www.sciencedirect.com/science/article/pii/S0022249699912798)
422 [pii/S0022249699912798](https://www.sciencedirect.com/science/article/pii/S0022249699912798).
- 423 Peter Filzmoser and Klaus Nordhausen. Robust linear regression for high-
424 dimensional data: An overview. *Wiley Interdisciplinary Reviews: Compu-*
425 *tational Statistics*, 13(4):e1524, 2021. Publisher: Wiley Online Library.