New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR)

Z. Genisheva[a,1], C. Quintelas[a,1,*], D.P. Mesquita[a], E.C. Ferreira[a], J.M. Oliveira[a], A.L. Amaral[a,b]

[a] CEB – Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal
[b] Instituto Politécnico de Coimbra, ISEC, DEQB, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal

ARTICLE INFO

ABSTRACT

This work aims to explore the potential of near infrared (NIR) spectroscopy to quantify volatile compounds in *Vinho Verde* wines, commonly determined by gas chromatography. For this purpose, 105 *Vinho Verde* wine samples were analyzed using Fourier transform near infrared (FT-NIR) transmission spectroscopy in the range of $5435\,cm^{-1}$ to $6357\,cm^{-1}$. Boxplot and principal components analysis (PCA) were performed for clusters identification and outliers removal. A partial least square (PLS) regression was then applied to develop the calibration models, by a new iterative approach. The predictive ability of the models was confirmed by an external validation procedure with an independent sample set. The obtained results could be considered as quite good with coefficients of determination (R^2) varying from 0.94 to 0.97. The current methodology, using NIR spectroscopy and chemometrics, can be seen as a promising rapid tool to determine volatile compounds in *Vinho Verde* wines.

1. Introduction

The control of wines quality and organoleptic characteristics is of paramount importance for the wine industry. The main constituents of wine are water, varying between 70 and 90%, ethanol, between 8 and 20%, sugars and acids. However, wines also contain other minor compounds that might greatly influence the sensory properties of the final product (Rankine, 1989). The rapid quantification of these compounds in wine is crucial not only to monitoring the fermentation process and for the final product, but also to detect fraudulent practices. In fact, it is relatively common for wine to be adulterated with alcohol, sugar, chalk, colorants, water or other more noxious chemicals to increase profits (Zeldin, 1977). According to these authors, the fraudulent practices were widespread for many years, mainly because the tools for its detection were not readily available.

The typical method to analyze wine volatile compounds is gas chromatography (GC) coupled to at least one detector, such as a flame ionization detector (GC-FID) or mass spectrometer (GC–MS). The latter is particularly advantageous as it provides fragmented data useful for molecular identification (Rebière, Clark, Schmidtke, Prenzler, & Scollary, 2010). However, the analysis of these compounds by GC requires sample preparation, specific analytical equipment and is time-

consuming. For these reasons, it is necessary to develop rapid determination methods, in order to predict the wine content on chemical compounds, and able to provide an alternative to the traditional analysis methods.

The recent developments in both chemometrics and instrumentation have resulted in rapid methods for predicting the concentration of specific chemical constituents and helped to reduce the demand for traditional analysis. Near infrared spectroscopy (NIR) is such a rapid and non-destructive technique and generally requires minimal sample processing prior to analysis. NIR is part of the electromagnetic spectrum, laying between the visible and IR regions of the electromagnetic spectrum, usually defined by the wavelength range of 700–2500 nm (Ye, Gao, Li, Yuan, & Yue, 2016). Many studies have been conducted to apply NIR to quality and safety measurements for food and agricultural materials, including fruits, vegetables, food and beverages (Huang, Yu, Xu, & Ying, 2008). This analytical technique has also been applied to predict phenolic compounds in red wine fermentations (Cozzolino et al., 2004), for the rapid measurement of total non-structural carbohydrate concentrations in grapevine trunk (De Bei et al., 2017), to discriminate Riesling wines from different countries (Liu et al., 2008), and to determine the technological maturity of grapes and total phenolic compounds of grape skins in red and white cultivars during

---

ripening (Nogales-Bueno, Hernández-Hierro, Rodríguez-Pulido, & Heredia, 2014).

The FT-NIR technique has been already successfully used by several authors to quantify a number of different parameters from wine focusing, in general, in the quantification of alcohol, total reducing sugars, total acidity, volatile acidity, total and free SO_2, glycerol, organic acids and pH (Cocciardi, Ismail, & Sedman, 2005; Patz, Blieke, Ristow, & Dietrich, 2004; Urbano-Cuadrado, Luque de Castro, Pérez-Juan, García-Olmo, & Gómez-Nieto, 2004). However, in most of the published works so far, little attention has been given to the individual quantification of quite relevant wine volatile compounds, encompassing individual alcohol and esters concentrations. In fact, to the authors' knowledge, the quantification of volatile compounds by FT-NIR is not usual, probably due to the diversity and complexity of the chemical data obtained from the GC analysis.

Therefore, the aim of the present work was to examine the potential of NIR spectroscopy to determine the concentration of 10 of the most relevant volatile compounds in white wines from Portuguese *Vinhos Verdes* Appellation of Origin at 5 stages of conservation and using different preservation methods. For such purpose, a new approach was used. First, a box plot procedure was employed in order to determine (and remove) the Y variables (studied compounds) outliers. Next, a first principal component analysis (PCA) allowed to both identify sample clusters and locate X variables (NIR values) outliers. Following the X outliers' removal, a second PCA allowed to identify the pruned dataset sample clusters. Furthermore, the choice of the calibration and validation data was also performed in order to minimize bias in both sets, by means of an iterative procedure up to a maximum of 2000 possible random combinations. The determination of the models was also performed by a new approach, taking into account the wavelength weight similarity in a first PLS analysis in order to perform a X data reduction (by grouping similar weights wavelengths) for the final PLS regression.

2. Materials and methods

2.1. Sample preparation and GC analysis

Wines designated "Appellation of Origin Vinhos Verdes" are produced in Northern Portugal and there are seven recommended white grape varieties (Alvarinho, Arinto, Avesso, Azal, Batoca, Loureiro and Trajadura) (Oliveira, Oliveira, Baumes, & Maia, 2008). Must fermentation, using a mixture of grapes from the Vinho Verde appellation of origin, was carried out in 500L fermenters. Vinifications were performed according to the traditional technology of the Vinhos Verdes region: the must obtained by crushing, pressing and static sedimentation was inoculated with *Saccharomyces cerevisiae bayanus* QA23. Fermentations then took place at 18 °C. Then, five bottles of each assay were prepared for analysis, in triplicate, at 5 stages of conservation (t0, t2, t3, t4 and t5) and using 7 different preservation methods. The employed preservation methods were the addition of SO_2 (until a final concentration of 35 mg/L), glutathione (20 mg/L) and SO_2 (C_f = 35 mg/L), caffeic acid (60 mg/L) and SO_2 (C_f = 35 mg/L), gallic acid (60 mg/L) and SO_2 (C_f = 35 mg/L), ascorbic acid (100 mg/L) and SO_2 (C_f = 35 mg/L), one bottle without any treatment (only with the natural antioxidant of wine) and one with glutathione (20 mg/L), caffeic acid (60 mg/L) and gallic acid (20 mg/L). All the compounds added to the wine are antioxidants well known as responsible for wine preservation. Time 0 (t0) – day of antioxidant treatment, t2 – 4 months of storage, t3 – 12 months of storage, t4 – 25 months of storage, t5 – 42 months of storage. A total of 105 samples were analysed.

Major volatile compounds, ethyl acetate, methanol, 2-methyl-1-butanol, 3-methyl-1-butanol, 2-phenylethanol, were directly analysed after adding 410 µg of 4-nonanol (internal standard—IS) to 5 mL of wine. A Chrompack CP-9000 gas chromatograph equipped with a split/splitless injector, a flame ionization detector (FID) and a capillary column, coated with CP-Wax 57CB (50 m × 0.25 mm; 0.2 µm film

thickness, Chrompack), was used. The temperatures of the injector and the detector were both set to 250 °C. The oven temperature was initially held at 60 °C, for 5 min, then programmed to rise from 60 °C to 220 °C, at 3 °C/min, and finally maintained at 220 °C for 10 min. The carrier gas was helium 4 × (Praxair) at an initial flow rate of 1 mL min^{-1} (125 kPa at the head of the column). The analyses were performed by injecting 1 µL of sample in the split mode (15 mL/min). The quantification of major volatile compounds, after the determination of the detector response factor for each analyte, was performed with the software Star-Chromatography Workstation version 6.41 (Varian) by comparing test compounds retention times with those of pure standard compounds.

Minor volatile compounds, 3-methylbutyl acetate, ethyl lactate, ethyl octanoate, diethyl succinate and diethyl malate, were analysed by GC–MS after extraction of 8 mL of wine with 400 µL of dichloromethane, spiked with 3.28 µg of 4-nonanol (IS), according to the methodology proposed by Genisheva, Vilanova, Mussatto, Teixeira, and Oliveira (2014). A gas chromatograph Varian 3800, with a 1079 injector and an ion-trap mass spectrometer Varian Saturn 2000, was used. A 1 µL injection was made in splitless mode (30 s) in a Varian Factor Four VF-Wax ms (30 m × 0.15 mm; 0.15 µm film thickness) column. The carrier gas was helium 4 × (Praxair) at a constant flow rate of 1.3 mL/min. The detector was set to electronic impact mode with an ionization energy of 70 eV, a mass acquisition range from 35 m/z to 260 m/z and an acquisition interval of 610 ms. The oven temperature was initially 60 °C for 2 min and then raised from 60 °C to 234 °C at a rate of 3 °C/min, raised from 234 °C to 250 °C at 10 °C/min and finally maintained at 250 °C for 10 min. The temperature of the injector was maintained at 250 °C during the analysis time and the split flow was maintained at 30 mL/min. The identification of compounds was performed using the software MS WorkStation version 6.9 (Varian) by comparing their mass spectra and retention indices with those of pure standard compounds. The minor compounds were quantified in terms of 4-nonanol equivalents only.

2.2. Near infrared scanning

Near infrared (NIR) spectra were recorded on a Fourier-transform near infrared spectrometer (FTLA 2000, ABB, Thermo Electron Corporation) equipped with an indium–gallium–arsenide (InGaAs) detector, from 14000 to 600 cm^{-1}, in transmitance mode using a flow cell with a 0.7 mm pathlength. For each sample, 32 scans were made with a spectral resolution of 8 cm^{-1} and then averaged. Samples were temperature equilibrated at 23 °C (approximately 3 min) in the instrument before scanning. The integration time was adjusted until the peaks at 1100–1200 nm for NIR were close to 60,000 intensity units. Grams/AI software (Thermo Electron Corporation) was used for spectrometer configuration, control, and data acquisition. Distilled water was used as background. Although the entire obtained spectral range (14000–600 cm^{-1}) was stored for each sample, only a selected interval (5435–6357 cm^{-1}) was considered to avoid interferences.

2.3. Chemometrics and data analysis

The chemometric analysis was based on a principal component analysis (PCA) to identify the interrelationships between the samples and possible clusters and a partial least squares (PLS) regression to derive the models for each studied compound. Regarding PCA, this method extracts the most relevant information from an X dataset by projecting it into latent variables (LVs), by a linear combination of the original variables, thus reducing high-dimensional and strongly correlated original datasets. In PCA, the objective is to maximize the explained variance in the new orthogonal space and, thus, each new orthogonal space accounts for less explanatory power than the previous. In this process, this method allows determining the interrelationships between the original X variables and samples that will appear in the new LVs space as clusters in the loadings and scores map, respectively.

With respect to the PLS regression, this method constructs LVs out of the X data in new, and orthogonal, spaces to maximize the captured predictive power of the X-space with regard to the Y-space. Matlab 7.11 (The Mathworks, Inc. Natick, USA) was used to that effect. Furthermore, and in order to remove undesirable variations the X data matrix was pre-processed using the standard normal variate (SNV) method. Also cross-validation (CV) was used to test the predictive significance of the PLS regression and determine the optimal number of LVs. Further details for these techniques can be found in Einax, Zwanziger, and Geiss (1997).

The collected NIR spectra used in this analysis ranged between 5435–6357 cm^{-1} (1573–1840 nm). Initially a boxplot analysis, regarding the dependent variables (Y), was performed resulting in Y outliers' identification and removal. Next a PCA–X analysis was carried out, regarding the independent variables (X) for the identification of distinct clusters and possible X outliers. This led to four different datasets fed to the PLS analysis: [1] – ensemble dataset with no X outliers removed; [2] – ensemble dataset with X outliers removed; [3] – dataset divided in 3 groups with no X outliers removed; and [4] – dataset divided in 3 clusters with X outliers removed. Hence, in the performed analysis, the total number of samples varied between 89 and 99, with 2/3 used for modelling (calibration) purposes and 1/3 for validation. Regarding the PLS approach, and concerning the choice of the calibration and validation data, a procedure was implemented in order to select the most unbiased sets, regarding the overall (calibration + validation) model results, by screening at 2000 possible random combinations for the selection of the calibration and validation data. In this sense, the most unbiased combination, thus resulting in the best overall (calibration + validation) results, between the calibration and validation data was chosen. Next, an iterative method was applied, first determining the weights of each wavelength for the entire wavelength values PLS, next grouping the wavelength values together according to the weights similarity and, finally, recalculating the PLS with the averaged wavelength values. For all PLS analyses, the maximum number of PLS components allowed was set at half of the calibration data.

3. Results and discussion

3.1. Analytical data

The number of volatile compounds present in white and red wines is quite large and, in this study six of the most significant esters (ethyl acetate, 3-methylbutyl acetate, ethyl lactate, ethyl octanoate, diethyl succinate, and diethyl malate) and four alcohols (methanol, 2-methyl-1-butanol, 3-methyl-1-butanol, and 2-phenylethanol) were selected to perform the calibration (Oliveira et al., 2008; Vilanova, Escudero, Graña, & Cacho, 2013). Accordingly, their minimum, average and maximum values in the samples are presented in Table 1. Each of these

Table 1
Minimum (Min.), maximum (Max.) and average (Avg.) values of the studied wine compounds.

Compound	Values (mg L^{-1})		
	Min.	Avg.	Max.
Ethyl acetate	6.7	63.1	99.4
Methanol	8.6	31.7	75.6
2-Methyl-1-butanol	8.7	16.6	24.6
3-Methyl-1-butanol	65.6	127.7	191.7
2-Phenylethanol	0.0	18.6	34.3
3-Methylbutyl acetate	0.0	522.3	2001.3
Ethyl lactate	228.2	453.8	898.9
Ethyl octanoate	70.2	241.2	782.0
Diethyl succinate	323.0	2497.7	5758.0
Diethyl malate	579.7	6297.9	14657.5

samples was then divided into two groups, the calibration group (ranging from 61 in [2] and [4] to 66 samples in [1] and [3]) and the external validation group (ranging from 28 in [2] and [4] to 33 samples in [1] and [3]).

3.2. Boxplot analysis

In order to identify the Y data (volatile compounds concentrations) outliers, a boxplot analysis was performed and is presented in Fig. 1. This analysis allowed identifying five outliers subsequently removed, regarding methanol, 2-methyl-1-butanol, 3-methyl-1-butanol, and ethyl octanoate concentrations. For this boxplot analysis, the maximum whisker length allowed was of 3, that is, three times the interquartile distance between the 25th and 75th percentiles. This resulted in the identifications of the outliers that fell outside a 99.9% coverage of a normally distributed data.

3.3. PCA analysis

The performed PCA analysis to the X dataset (wavelengths values) with no outliers removed allowed identifying three different clusters, as depicted in Fig. 2a. The first, second and third principal components (or latent variables – LV) explained 98.2% of the X dataset variance, with respectively 69.1%, 27.0% and 2.1% for LV1, LV2 and LV3. Furthermore, this PCA analysis also allowed to identify X outliers (represented within the dotted circles), removed in the subsequent analysis ([3] and [4]). For this purpose, the 2nd and 3rd LVs were used, given that the outliers were present mainly in the LV2 *vs* LV3 axis. In that sense, the samples presenting a distance above the value of one standard deviation, regarding the LV2 *vs* LV3 centroid of the respective cluster, were identified as outliers and removed for the subsequent analysis.

A second PCA analysis was then performed to the X dataset with outliers removed that also allowed to identify three different clusters, as depicted in Fig. 2b. The 1 st, 2nd and 3rd principal components (or latent variables – LV) explained 98.1% of the X dataset variance, with respectively 89.0%, 7.5% and 1.6% for LV1, LV2 and LV3. As discussed in the PLS analysis, the identification of these three clusters was fundamental for the best model choice regarding the ethyl acetate, methanol and ethyl lactate compounds.

3.4. PLS regression

Calibration models based on the GC–MS analysis and FT-NIR spectra were developed using partial least squares regression (PLS) for all the studied compounds. In order to determine the best model results for each compound, the root mean square error (RMSE) values (in relative percentage of the mean concentration value for each compound) of the overall (calibration + validation) model results were taken into consideration. Thus, the best models were identified, and are presented in Table 2.

For all the studied compounds, the best coefficients of determination (R^2) were equal or above 0.94. Regarding the RMSE values, two (2-methyl-1-butanol and 3-methyl-1-butanol) were below 5% of the sample average values, four (ethyl acetate, methanol, 2-phenylethanol and ethyl lactate) were between 5% and 10%, three (ethyl octanoate, diethyl succinate and diethyl malate) were between 10% and 15%, and only one (3-methylbutyl acetate) was slightly above 20%. Although no significant differences were found, between the studied alcohol and ester groups, regarding the R^2 values (with both presenting an average R^2 of 0.96), regarding the RMSE, the alcohols presented better results with an average value of 5.4% (against 12.8% for the esters). The same conclusion can be withdrawn regarding the major *vs*. minor volatile compounds, with the major compounds presenting an average RMSE value of 5.8% against 13.9% for the minor compounds. Given the prior results, this comes as no surprise since 4 out of the 5 studied major volatile compounds were alcohols.
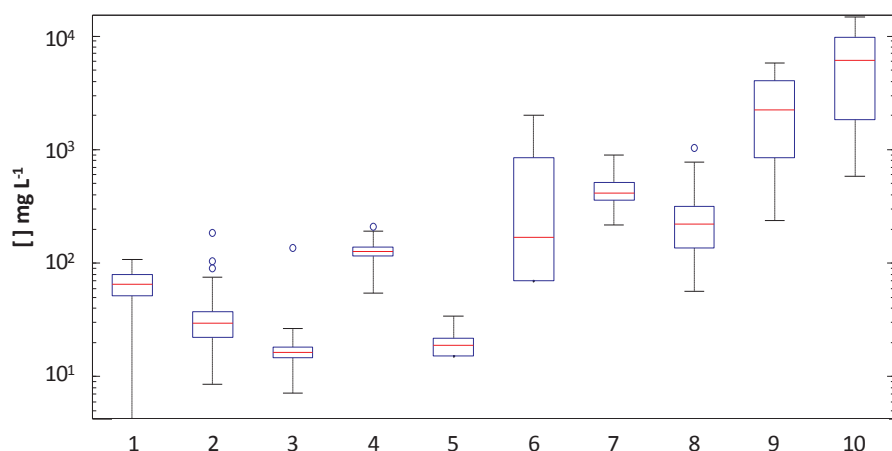
Fig. 1. Boxplot analysis of the Y dataset. 1 – ethyl acetate; 2 – methanol; 3 – 2-methyl-1-butanol; 4 – 3-methyl-1-butanol; 5 – 2-phenylethanol; 6 – 3-methylbutyl acetate; 7 – ethyl lactate: 8 – ethyl octanoate; 9 – diethyl succinate; 10 – diethyl malate.
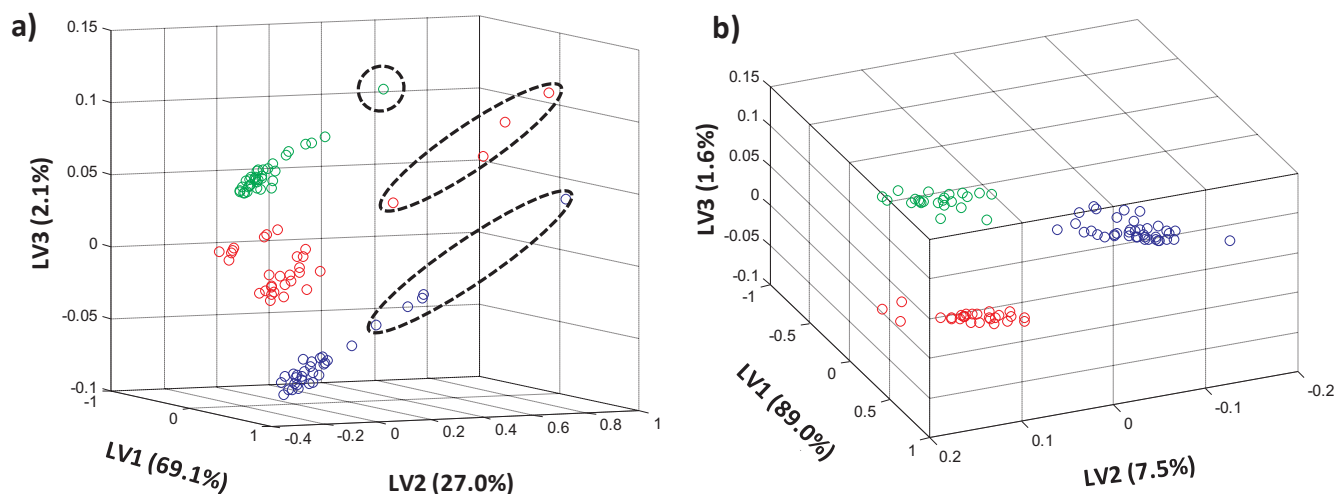

Fig. 2. PCA analysis of the X dataset, showing the first (LV1), second (LV2) and third (LV3) latent variables with, a) no outliers removed (samples in [1] and [3]; b) outliers removed (samples in [2] and [4]). The data points within the dotted circles in a) represent the samples removed (X outliers) in the subsequent analysis.

Table 2
Coefficient of determination (R^2), root mean square error (RMSE) (%), number of PLS components (n) and residual predictive deviation (RPD) of the PLS analysis for the overall (calibration + validation) model results.

Compound	[1]				[2]				[3]				[4]			
	R^2	RMSE(%)	n	RPD	R^2	RMSE(%)	n	RPD	R^2	RMSE(%)	n_1, n_2, n_3	RPD	R^2	RMSE(%)	n_1, n_2, n_3	RPD
Ethyl acetate	0.88	11,4	25	2,7	0.79	13,8	15	2,3	**0.95**	**7,6**	**10, 10, 9**	**3,9**	0.91	9,7	13, 9, 7	2,9
Methanol	0.90	13,6	18	2,9	0.92	11,7	24	3,2	**0.96**	**8,5**	**8, 10, 9**	**4,4**	0.92	11,7	11, 7, 7	3,2
2-Methyl-1-butanol	0.96	4,2	23	4,5	**0.96**	**3,6**	**19**	**4,7**	0.92	5,4	9, 9, 9	3,1	0.91	5,4	11, 8, 7	3
3-Methyl-1-butanol	**0.96**	**3,9**	**20**	**4,7**	0.95	4,5	18	3,9	0.92	5,6	9, 9, 8	3,1	0.91	5,7	9, 8, 7	3
2-Phenylethanol	**0.97**	**5,4**	**19**	**4,8**	0.94	6,5	17	3,8	0.92	7,5	10, 7, 8	3,2	0.91	8,1	11, 9, 7	2,8
3-Methylbutyl acetate	**0.96**	**23,0**	**21**	**4,4**	0.96	23,0	27	4,1	0.95	25,7	10, 9, 8	3,7	0.83	42,5	11, 9, 6	2,2
Ethyl lactate	0.90	9,8	18	3	0.94	7,7	21	3,6	0.95	7,1	9, 10, 7	4	**0.96**	**5,9**	**12, 8, 7**	**4,5**
Ethyl octanoate	0.91	15,6	18	3,2	**0.94**	**13,4**	**19**	**3,7**	0.92	15,4	8, 8, 9	3,1	0.93	14,2	11, 9, 7	3,3
Diethyl succinate	**0.97**	**12,4**	**20**	**4,9**	0.94	16,7	22	3,5	0.92	18,8	8, 8, 10	3,1	0.92	18,5	8, 9, 7	3,1
Diethyl malate	**0.95**	**14,5**	**16**	**4,1**	0.95	13,6	20	4,2	0.93	17,2	7, 8, 9	3,8	0.95	13,9	9, 7, 7	4,1

The residual predictive deviation (RPD), defined as the ratio between the standard deviation of the population (SD) and the standard error of cross validation (SECV) for the NIR predictions was also calculated and is presented in Table 2. According to Fearn (2002), an RPD value greater than 3 is considered fair and recommended for screening purposes, and an RPD value greater than 5 is considered good for quality control. All the RPD values calculated in the present report are higher than 3 confirming the good performance of the developed models.

In most cases (seven out of the ten studied compounds), no advantage was found by identifying first the different clusters, and next proceeding with individual calibrations for each cluster. Only in three cases (ethyl acetate, methanol, and ethyl lactate), this procedure gave better model results. Likewise, only for three initial datasets (2-methyl-1-butanol, ethyl lactate, and ethyl octanoate), the X outliers' removal procedure was found advantageous.

The best prediction model results for the studied compounds are next presented in Fig. 3. Analysing this figure it is apparent that both
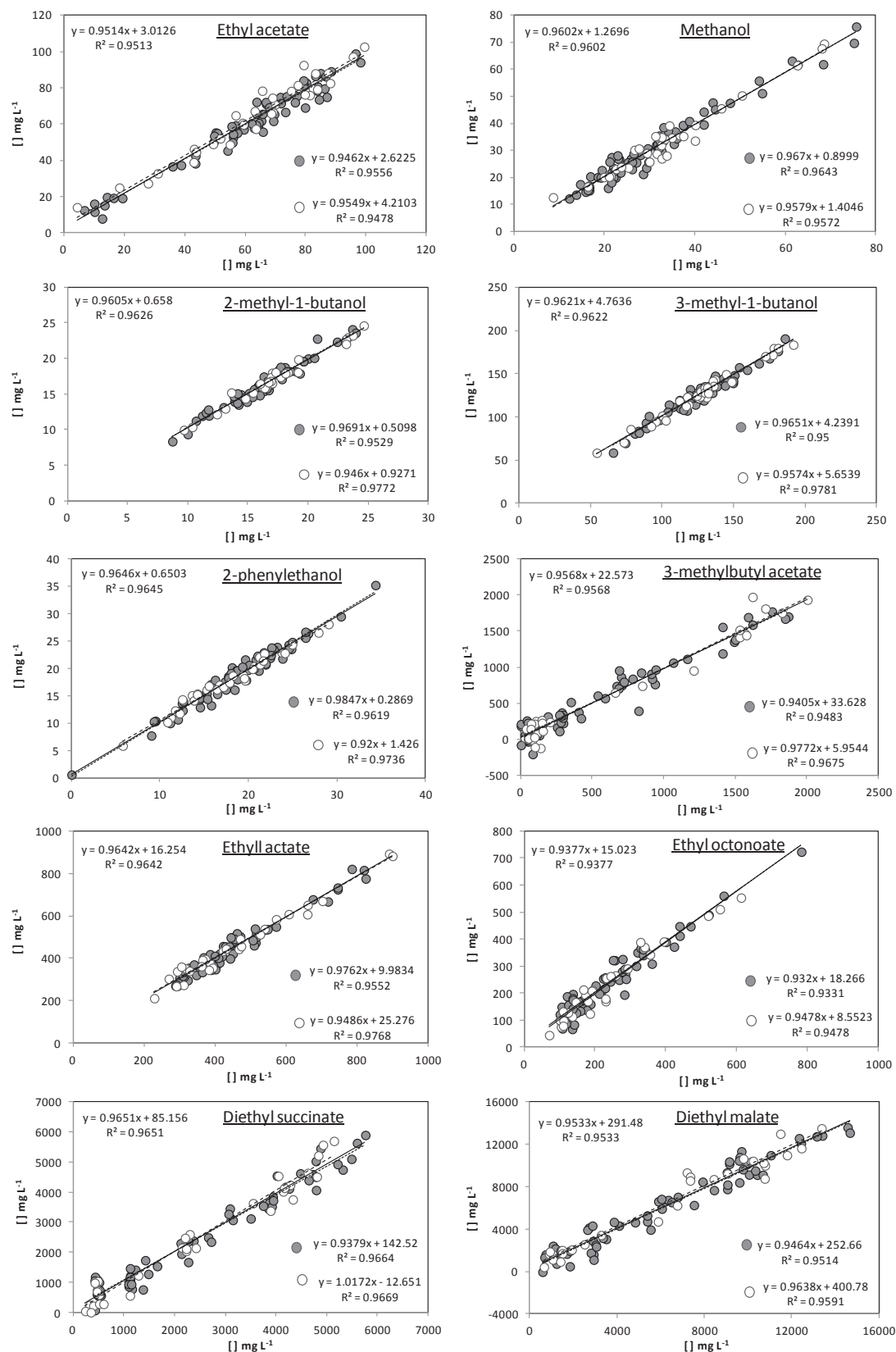
Fig. 3. Best model results for the studied compounds. The grey circles represent the calibration data and the white circles represent the validation data.

the calibration data (represented by the grey circles) and the validation data (represented by the white circles) are in good agreement with the resulting model. This conclusion is reinforced analysing the obtained R^2 values for the calibration data and validation data alone (also presented in Fig. 3).

3.5. Novelty and advantages of the presented methodology

As explained previously in the Introduction section, the FT-NIR technique is commonly used, by a number of authors, to quantify several wine compounds. In fact, taking into account the studied parameters, and the presented results by these authors, it can be concluded that fair to good calibration models were obtained for most of the components and parameters tested. Therefore, these results validate, at some extent, the feasibility of using FT-NIR spectra data for monitoring some of the most important wine compounds and parameters. However, as previously mentioned, the individual alcohol and ester concentrations have been given little attention, although a few works deserve a more detailed discussion. Such is the case of the work published by Dambergs, Kambouris, Francis, and Gishen (2002) which studied the quantification of methanol by NIR and was able to obtain a good correlation (R^2 value of 0.98) with the conventional GC analysis. Taking into account that the present work allowed for an R^2 value of 0.96, which is not too distant from the result obtained by Dambergs et al. (2002), it should also be stressed that the studied methanol range in the current work is quite narrower, and thus, with a wider set of concentrations the presented methodology can be expected to perform better.

Smyth et al. (2008) tested this technique to quantify eleven volatile compounds from Riesling wine and obtained R^2 values of 0.74 for the overall quantification of esters (ethyl butanoate, ethyl hexanoate, ethyl decanoate, ethyl octanoate, and ethyl dodecanoate), 0.90 for monoterpene alcohols (linalool, geraniol, and nerol) and 0.80 for short-chain fatty acids (hexanoic, octanoic, and decanoic acids). Comparing these authors' results with the ones obtained by the current methodology, a large improvement is obtained regarding the esters determination. In fact, an average R^2 value of 0.96 was obtained by the proposed methodology, which compares quite favorably to the 0.74 value of Smyth et al. (2008).

Also Lorenzo, Garde-Cerdán, Pedroza, Alonso, and Salinas (2009) have essayed to determine red wine volatile compounds by NIR analysis, studying mostly individual esters (ethyl octanoate, ethyl decanoate, ethyl 2-phenylacetate, diethyl succinate, diethyl glutarate, and 2-phenylethyl acetate), alcohols (2-phenylethanol) and organic acids (hexanoic and octanoic acids) composition. In addition, although the calibration models, against the GC analytical data, were near to perfect in all cases (with R^2 values around 0.99), when subjected to a full cross validation procedure, the prediction ability steeply decreased. In fact, even when dividing the data by zone of origin, the prediction R^2 value did not surpass 0.94 in all cases (and presented a lower of 0.42 for ethyl decanoate). In fact, the largest R^2 values obtained for ethyl octanoate (0.85), diethyl succinate (0.86), and 2-phenylethanol (0.83) were quite below the current results (respectively 0.94, 0.97, and 0.97).

In another study, Ye et al. (2016) tested the detection of volatile compounds in 42 apple wines samples using FT-NIR spectroscopy, regarding a number of esters, higher alcohols and fatty acids. The prediction ability found by these authors with respect to the studied esters group (ethyl acetate, ethyl caprylate, ethyl butyrate, ethyl hexanoate, 2-phenethyl acetate, ethyl laurate, ethyl decanoate, and hexyl acetate) presented an average R^2 value of 0.90 (and a maximum of 0.92 for ethyl acetate). Therefore, the esters results obtained by the current methodology can be considered better, ranging from 0.94 to 0.97, with an average value of 0.96. Regarding the alcohols group (isobutanol, hexanol, 2,3-butanediol, 2-phenethanol, and 3,4,5-trimethyl-4-heptanol) studied by Ye et al. (2016), presented an average R^2 value of 0.92 (and a maximum of 0.95 for hexanol). Again, the current methodology

allowed obtaining better alcohols prediction abilities ranging from 0.96 to 0.97, with an average value of 0.96.

It should also be emphasized that almost all the referred authors used a similar procedure to analyze the spectral data. This procedure aims at: a) pre-process the data to minimize the spectra baseline variation and b) enhance spectra differences. In order to do so, most methodologies apply a number of the following steps: multiplicative scatter correction; first and/or second derivative; vector and/or min/max normalization; constant offset elimination. Finally, the development of the calibration models is usually performed by PLS regression. Considering this, the chemometric approach used in the present work to correlate the FT-NIR spectra and GC data, explained in the materials and methods section, can be considered as novel and, furthermore, allowed to obtain good prediction abilities for the parameters tested.

4. Conclusions

This work addressed the potential of NIR transmission spectroscopy to quantify ten different volatile compounds in *Vinho Verde* wines, namely ethyl acetate, methanol, 2-methyl-1-butanol, 3-methyl-1-butanol, 2-phenylethanol, 3-methylbutyl acetate, ethyl lactate, ethyl octanoate, diethyl succinate, and diethyl malate. For this purpose, a chemometric approach was used employing first boxplot and PCA analysis for outliers' removal and clusters identification. Next a new iterative PLS procedure was used with a twofold objective: a) first reduce the NIR dataset and b) construct the calibration models based on standard GC–MS analysis. This procedure resulted in high coefficients of determination (R^2), and low root mean square errors (RMSE), for the prediction ability of the 10 studied volatile compounds. In conclusion, this study has shown that NIR spectroscopy can be used in *Vinho Verde* wines to determine volatile compounds in a simple and fast way, thus avoiding the traditional costly, complex and time-consuming analytical techniques. Furthermore, this procedure has also the advantage of being a non-destructive and non-contaminating method, avoiding the need for prior sample treatment.

Acknowledgments

Conflict of interest

There is no conflict of interest.

References

Cocciardi, R. A., Ismail, A. A., & Sedman, J. (2005). Investigation of the potential utility of single-bounce attenuated total reflectance Fourier transform infrared spectroscopy in the analysis of distilled liquors and wines. *Journal of Agricultural and Food Chemistry, 53*, 2803–2809.

Cozzolino, D., Kwiatkowski, M. J., Parker, M., Cynkar, W. U., Dambergs, R. G., Gishen, M., & Herderich, M. J. (2004). Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy. *Analytica Chimica Acta, 513*, 73–80.

Dambergs, R. G., Kambouris, A., Francis, J. L., & Gishen, M. (2002). Rapid analysis of methanol in grape-derived distillation products using near-infrared transmission spectroscopy. *Journal of Agricultural and Food Chemistry, 50*, 3079–3084.

De Bei, R., Fuentes, S., Sullivan, W., Edwards, E. J., Tyerman, S., & Cozzolino, D. (2017). Rapid measurement of total non-structural carbohydrate concentration in grapevine trunk and leaf tissues using near infrared spectroscopy. *Computers and Electronics in Agriculture, 136*, 176–183.

Einax, J. W., Zwanziger, H. W., & Geiss, S. (1997). *Chemometrics in environmental analysis.* Weinheim, Germany: Wiley-VCH.

Fearn, T. (2002). Assessing calibrations: SEP, RPD, RER and R^2. *NIR News, 13*, 12–14.

Genisheva, Z., Vilanova, M., Mussatto, S. I., Teixeira, J. A., & Oliveira, J. M. (2014). Consecutive alcoholic fermentations of white grape musts with yeasts immobilized on grape skins: Effect of biocatalyst storage and SO2 concentration on wine characteristics. *LWT – Food Science and Technology, 59*, 1114–1122.

Huang, H., Yu, H., Xu, H., & Ying, Y. (2008). Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review. *Journal of Food Engineering, 87*, 303–313.

Liu, L., Cozzolino, D., Cynkar, W. U., Dambergs, R. G., Janik, L., O'Neill, B. K., ... Gishen, M. (2008). Preliminary study on the application of visible–near infrared spectroscopy and chemometrics to classify Riesling wines from different countries. *Food Chemistry, 106*, 781–786.

Lorenzo, C., Garde-Cerdán, T., Pedroza, M. A., Alonso, G. L., & Salinas, M. R. (2009). Determination of fermentative volatile compounds in aged red wines by near infrared spectroscopy. *Food Research International, 42*, 1281–1286.

Nogales-Bueno, J., Hernández-Hierro, J. M., Rodríguez-Pulido, F. J., & Heredia, F. J. (2014). Determination of technological maturity of grapes and total phenolic compounds of grape skins in red and white cultivars during ripening by near infrared hyperspectral image: A preliminary approach. *Food Chemistry, 152*, 586–591.

Oliveira, J. M., Oliveira, P., Baumes, R. L., & Maia, O. (2008). Changes in aromatic characteristics of Loureiro and Alvarinho wines during maturation. *Journal of Food Composition and Analysis, 21*, 695–707.

Patz, C.-D., Blieke, A., Ristow, R., & Dietrich, H. (2004). Application of FT-MIR spectrometry in wine analysis. *Analytica Chimica Acta, 513*, 81–89.

Rankine, B. (1989). *A manual of winemaking practice for Australia and New Zealand.* Sydney, Australia: Pan Macmillan.

Rebière, L., Clark, A. C., Schmidtke, L. M., Prenzler, P. D., & Scollary, G. R. (2010). A robust method for quantification of volatile compounds within and between vintages using headspace-solid-phase micro-extraction coupled with GC–MS – Application on Semillon wines. *Analytica Chimica Acta, 660*, 149–157.

Smyth, H. E., Cozzolino, D., Cynkar, W. U., Dambergs, R. G., Sefton, M., & Gishen, M. (2008). Near infrared spectroscopy as a rapid tool to measure volatile aroma compounds in Riesling wine: possibilities and limits. *Analytical and Bioanalytical Chemistry, 390*, 1911–1916.

Urbano-Cuadrado, M., Luque de Castro, M. D., Pérez-Juan, P. M., García-Olmo, J., & Gómez-Nieto, M. A. (2004). Near infrared reflectance spectroscopy and multivariate analysis in enology. Determination or screening of fifteen parameters in different types of wines. *Analytica Chimica Acta, 527*, 81–88.

Vilanova, M., Escudero, A., Graña, M., & Cacho, J. (2013). Volatile composition and sensory properties of NorthWest Spain white wines. *Food Research International, 54*, 562–568.

Ye, M., Gao, Z., Li, Z., Yuan, Y., & Yue, T. (2016). Rapid detection of volatile compounds in apple wines using FT-NIR Spectroscopy. *Food Chemistry, 190*, 701–708.

Zeldin, T. (1977). *France 1848-1945: volume 2: Intellect, taste and anxiety.* Oxford, UK: Oxford University Press.