Talanta 86 (2011) 293–304

Contents lists available at SciVerse ScienceDirect

Talanta

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

Development of a fast and reliable method for long- and short-term wine age prediction

Ana C. Pereira [a,b,*], Marco S. Reis [a], Pedro M. Saraiva [a], José C. Marques [b]

[a] CIEPQPF, Department of Chemical Engineering, University of Coimbra, Pólo II – Rua Sílvio Lima, 3030-790 Coimbra, Portugal
[b] Exact Sciences and Engineering, University of Madeira, Campus Universitário da Penteada, 9000-390 Funchal, Portugal

ARTICLE INFO

Article history:
Received 13 April 2011
Received in revised form 1 September 2011
Accepted 12 September 2011
Available online 16 September 2011

Keywords:
Wine age
UV–vis data
Pre-processing
Variable selection
PLSR

ABSTRACT

Wine age prediction based on its intrinsic characteristics can provide significant assistance to oenologists' quality evaluations, concerning wine ageing process control and wine quality assurance. Simpler, faster, cheaper and affordable analytical procedures would be greatly welcome to establish such a practice. In this study, we present a new and reliable strategy to predict wine age, in the long and short-term, centered on the use of wine UV–vis absorbance data, coupled with proper chemometric techniques.

The strategy followed consists essentially in first pre-processing the UV–vis data, secondly to carry out variable selection over such pre-processed data sets, and finally to use the set of selected variables for developing a PLS model focused on wine age prediction. We tested different data pre-processing methodologies, namely first and second derivatives, multiplicative scatter correction, standard normal variate and orthogonal signal correction, as well as different variable selection approaches, specifically interval partial least squares, VIPS, genetic algorithms and the wavelet transformation combined with a genetic algorithm.

In both case studies, regarding long and short-term ageing periods, we have found out that it is indeed possible to predict wine ages, in our case Madeira wine ages, with an accuracy of 1.4 years for longer ageing periods, and of 3 months for wines of an age comprised in the first two years of ageing. The genetic algorithm revealed to be very useful for proper wavelet coefficients selection, leading to the most parsimonious model among all those analyzed, which also presents the best predictive performance found.

1. Introduction

The use of reliable methods to ensure compliance of wines is imperative in order to guarantee their quality, control unfair competition and to limit or detect fraudulent practices. Considerable research has been devoted to the application of analytical methods for wine authentication and quality assurance tasks. Recently, a review paper on this topic was published [1], in which different studies for wine characterization were presented and the analytical procedures applied were scrutinized and discussed in terms of their associated feasibility, advantages, practical applications and merit of the results provided. The main purpose of such procedures is usually essentially focused on one of the following topics: to classify wines according to their geographical area of production; discriminate wines according to grape variety; predict wine harvest year; distinguish different winemaking practices; predict oenological parameters; and evaluate the sensorial features of wine. Undoubtedly, the success of the results achieved in all these cases relies, to a great extent, on the proper integration of modern analytical techniques, including spectroscopy, separation techniques, mass spectrometry or sensor equipment, with data analysis methodologies, especially those falling under the heading of "chemometric" approaches [1–7]. However, despite the plethora of analytical methods currently available, they tend to present limitations that hinder their widespread adoption in industry. Such main limitations are basically due to the fact that they are time-consuming, require skilled personnel and involve expensive equipment.

In wineries, spectroscopic techniques have gained a wide acceptance, mainly due to their ability to obtain wine chemical information at a low cost, without requiring an extensive sample pre-treatment, through non-destructive and simple procedures. In particular, UV–vis spectroscopy has been exploited in a number of publications [8]. Skogerson et al. investigated the application of multivariate methods to wine's UV–vis spectra in order to develop predictive models for determining a range of phenolic compounds, throughout the stages of winemaking. The partial least squares regression technique (PLS) was then employed for establishing

* Corresponding author. Tel.: +351 239 798 793; fax: +351 239 798 703.
E-mail address: apereira@eq.uc.pt (A.C. Pereira).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.09.016

individual predictive models for the different classes of phenolic compounds, and a validation data set was used to independently test such models. The results obtained showed that the proposed models do present potential for rapid determination of color and phenol components in juices, must and wines, being already in use in some wineries. PLS has also been applied to predict some wine's physico-chemical parameters, such as total acidity, volatile acidity, pH, free sulfur dioxide, tannins and total anthocyanins, from UV–vis spectral data [9]. In this case, the method proposed was tested on different types of samples regarding young wines, aged wines and commercial wines. Despite the small size of the data set studied, the estimated models illustrate the potential of using UV–vis spectra for predicting wine parameters that are relevant for routine quality control tasks, in a simple way. Urbano et al. [10] reported the use of UV–vis spectroscopy along with pattern recognition methods, for differentiating and classifying Spanish wines. Principal components analysis (PCA) and Soft Independent Modeling of Class Analogy (SIMCA) were used for exploratory data analysis and development of classification models, respectively. Acevedo et al. [11] also proposed a tool based upon UV–vis absorbance data and support vector machines (SVM) for classifying Spanish red and white wines according to their denomination of origin. Analyzing all these references, we can however notice, to the best of our present knowledge, a lack of applications dedicated to the important issue of predicting wine ages using UV–vis spectrum information. The ageing and typification of vintage Ports has been studied by Cruz Ortiz et al. using PLS and SIMCA modeling techniques, based on analytical quality control data, collected in a routine way. Rudnitskaya et al. [12] proposed a strategy for predicting Port wine age based on a potentiometric e-tongue. Recently, a similar study has been carried out by the same author, for Madeira wine age prediction [13]. In reference [14] a classification strategy to determine the Madeira wine harvest year, based on flavour composition, has also been proposed. Age prediction studies also have been carried out for spirit beverages. For example, Watts and co-authors [15] used SPME-GC–MS, along with PLS, to analyze cognac samples and establish a model to predict its age. The above mentioned examples have proved a reasonable success in addressing the problem of age prediction. However, notwithstanding their higher specificities, due to the employment of dedicated separation analytical procedures, simpler, faster and cheaper methods would be greatly welcome, in order to enable for an easier and affordable implementation of the wine age prediction task. In this paper, such a method for wine age prediction will be proposed, as a valid alternative to those based on chromatography techniques and e-tongue sensors.

The present study is contextualized in our recent work on the development and comparison of different procedures for predicting Madeira wine age. In reference [16] we have already illustrated the potential of using UV–vis data for developing wine age prediction models. In this study, twenty five aged Madeira wine samples (MW), ranging from one to nineteen years of ageing time, were used for training and validating a prediction model (Monte Carlo validation). The results obtained show that PLS applied to UV–vis pre-processed spectra provide indeed good prediction scores for wine age. The root mean square error (RMSE) achieved was of 1.4 years, using a 5 latent variables in our PLS method. In the sequence of this study [16], two important research problems do naturally arise. The first one involves the external validation of such a model, with a new, independent data set. The second issue is of a different nature, and regards determining whether a similar strategy can also be applied for characterizing the first months of ageing, and to predict wine ageing in such shorter time scales. The present study endeavors to address both of these topics, centered on the use of wine UV–vis absorbance data, coupled with proper chemometric techniques for adequately dealing with their structure. In particular, the nature of spectroscopy data requires the use of adequate pre-processing and variable selection approaches, in order to: (i) reduce the effect of systematic variations introduced by light scattering and variable path lengths, which are not related to the sample specific information [17], and (ii) to remove non-informative variables that might interfere, in a negative way, with the model's predictive ability [18], respectively. These two issues are considered in our study, in order to optimize prediction capability.

The remaining parts of our article are organized as follows. Section 2 provides a brief and general presentation of pre-processing and variable selection approaches adopted, together with a short description of PLS regression. In Section 3 we briefly describe the data set used in this work, as well as the analytical and statistical procedures employed. Section 4 is organized in two parts: in the first part, the results of long-term ageing prediction are presented and discussed, while in the second part the short-term ageing prediction results are analyzed. Finally, in Section 5, the main achievements of this work are summarized.

2. Background and theory

Partial least squares regression (PLSR) is a modeling technique that can be employed for predicting one or several responses, especially when the regressors or input variables are more and highly collinear. Even though PLS is able to cope quite well with high dimensionalities of the input space, it turns out that its prediction performance can often be optimized by eliminating those variables that do not bear any predictive power [19,20]. Furthermore, prior to the estimation of the regression model from spectroscopy data, it is important to correct it for features that can limit their utility, such as non-linearities introduced by light scattering. In this study, we address both of these issues, looking for the best pre-processing combinations for deriving our predictive model. The sequence of steps followed consists essentially in first pre-processing the UV–vis data, secondly to carry out variable selection on such pre-processed data sets, and finally to use the set of selected variables for developing a PLS model focused on wine age prediction. Different pre-processing and variable selection methodologies were tested, and the final selection was based upon their performance during the validation phase, using a Monte Carlo approach, as described in reference [16]. However, in the particular case of variable selection procedure selection, some modifications were introduced, which are described in Section 2.3.

In the remaining parts of this section, we make a brief reference to pre-processing and variable selection methodologies employed, focusing essentially on their methodological aspects. For more details on the PLS methodology, readers are referred to [21–24].

2.1. Notation

Multivariate data are usually organized in two-way matrices. In this text, such matrices are represented by bold capital letters (e.g. $\mathbf{X}$), vectors by bold lowercase letters (e.g. $\mathbf{x}$), and scalars by italicized characters, (e.g. w). Elements of a vector are denoted by italic characters with a subscript index (e.g. $\mathbf{X}_{w-1}$). The hat symbol ($\hat{}$) indicates a predicted value. The $\mathbf{X}$ and $\mathbf{y}$ symbols will be used to represent the predictor matrix (whose columns are relative to different wavelengths, while lines regard different samples) and response vector, respectively.

2.2. Spectral pre-processing

There are currently a rich variety of pre-processing methods that can be employed with spectroscopic data. In order to remove undesirable systematic variation in the data, three types of pre-processing methods are commonly reported in the analytical chemistry literature, namely: (i) differentiation – first and second

derivatives on smoothed data, (ii) signal correction – multiplicative scatter correction (MSC), standard normal variate (SNV), and (iii) filtering-based methods – such as orthogonal signal correction (OSC) [16].

A handy and frequently used method to determine the derivative of a spectral signal is the difference between the numerical values at adjacent points, $\mathbf{X}_{w-1}$, $\mathbf{X}_w$, $\mathbf{X}_{w+1}$, that is, the absorbance measured spectrum at wavelength indexes $w-1$, w and $w+1$. Occasionally, it is necessary to incorporate some preliminary smoothing in the derivative calculations, namely using the Savitzky–Golay approach. The idea behind this approach is to use the first or second derivative of the fitted polynomial, at each wavelength, to estimate the first and second derivative of the underlying spectrum at that wavelength, respectively [25]. In our analysis, we computed derivatives using first and second differences, with and without the adjusted polynomial from the Savitzky–Golay approach.

The Multiplicative Scatter Correction (MSC) method is essentially motivated by the empirical fact that when one looks at a plot of the sample-spectral values against the corresponding mean spectral values (taken from all the samples), a fairly linear relationship is obtained (a regression line). This empirical finding indicates that the "scatter effect" has multiplicative and additive components [25]. Consequently, the differences between such lines were interpreted as differences due to "scatter effects", while the deviation from the regression line was interpreted as corresponding to chemical information in the spectra. The regression line coefficients are unknown, and must be estimated individually for each sample. More details about the algorithm can be found in reference [26]. The practical difference between MSC and the Standard Normal Variate (SNV) approach is that SNV standardizes each spectrum using only data from that spectrum (it does not use the mean spectrum for all the samples, as is the case for MSC). It is a form of normalization that gives more weight to mid-range values than to outlying large absorbance values. Mathematically, it is identical to performing autoscaling of the rows instead of the columns, i.e. doing autoscaling of the transpose of the **X**-block.

Orthogonal Signal Correction, on the other hand, is a filtering method which attempts to reduce light scattering effects and other types of interferences. The variation in the **X**-block that is unrelated to **y** may disturb the multivariate modeling and cause imprecise predictions for new samples. Therefore, such information is removed by applying this filter, while all information in the spectrum related to the reference value (**y**) is maintained (the removed part of the **X**-block is mathematically orthogonal to **y**). The OSC algorithm is identical to the ordinary PLS algorithm (for example, it employs the NIPALS algorithm), except for the step where the weights of the **X**-block, **W**, are computed. Normally, these are calculated in order to maximize the covariance between the **X**-block and **Y**-block, but in OSC filtering they are computed so as to minimize such a covariance [27,28].

2.3. Selection of variables (wavelengths)

The problem of variable selection can be carried out through different approaches. For instance, one can try to exhaustively analyze all possible combinations of sets of variables with different dimensionalities, and select the best one. However, these methods (so called "best set" selection) do quickly become impracticable when handling a large number of variables, such as those produced in spectroscopy. Therefore, a variety of alternative variable selection methodologies have been proposed in order to find an adequate set of variables, namely one that is able to produce a small error when used in prediction tasks. The selection of the set of spectral wavelengths should be based upon a suitable and robust criterion, which assures its predictive ability in future applications, while maintaining its complexity at a minimum (i.e. using a small number of predictors).

The methods for variable selection essentially fall under two major categories: methods based on individual variable selection and approaches aimed at finding the most informative sequential groups of variables (intervals). The former ranks the importance of the individual variables according to one or several metrics, in order to evaluate the importance of each variable in predicting **Y** responses, and then uses a cutoff criterion to segment the relevant/irrelevant variables. The interval methods, on the other hand, try to find out the most informative group of variables (i.e. spectral bands). The intervals are essentially obtained by splitting the spectra into a certain number of intervals, or through an iterative construction procedure. The most informative intervals are also assessed using various metrics and a cutoff criterion. In our study, four variable selection methodologies were tested, namely: the VIPS approach (VIPs), interval partial least squares (iPLS), genetic algorithm (GA) and the wavelet transformation combined with a genetic algorithm (WT-GA). With the exception of the VIPs approach, the remaining were computed selecting windows of wavelengths instead of choosing wavelengths independently. A brief discussion of the underlying assumptions of each method and the important aspects concerning the associated procedures is outlined in the next subsection.

In order to further compare the results for all methods, a Monte Carlo study was carried out, where about 100 resampled data sets, each one comprising a training and a testing set, were randomly formed. We have confined the variable selection validation computations to such a number, as GA is computationally a time-consuming method. The predictive ability of all models, using the selected variables, is measured by the average of the root mean square error of prediction obtained for all the resampling groups, which will be denoted as *RMSE-MC*. The entire procedure is repeated with an increasing number of latent variables. The number of latent variables in the final model is chosen based on the minimum prediction error obtained.

2.3.1. VIPS variable selection approach (VIPs)

Among the four variable selection methodologies tested in this work, the so-called VIPs approach is the most simple in terms of the underlying algorithms computations. The variable importance on a projection metric, VIP, reflects how well the prediction variable is explained and how important it is for the prediction ability of the model [29]. The average value of the squared VIP over all wavelengths is one. Hence, a VIP value larger than one is usually considered to correspond to an informative variable. Here, the procedure adopted to determine the best subset of variables, according to the VIPs approach, consists in the following steps: (i) compute a regression model for the resampled training data set; (ii) then calculate the corresponding VIPs for all wavelengths, and (iii) finally, select the most relevant wavelengths. A wavelength was considered to be relevant when the average VIP value for all the resample trials is greater than one. The wavelengths thus selected are used to compute a PLS model, from which the wine age of testing data set will be predicted. The average of prediction errors regarding all trials is used as an estimate of the model performance.

2.3.2. Interval partial least squares (iPLS)

One of the first interval approaches proposed in the literature was developed by Nørgaard et al. [30]. It consists in dividing spectra into a certain number of equal width intervals, and computing independent PLS models for each interval. The most important intervals, i.e. those minimizing the prediction error, are kept in the data set. In order to optimize the selection of intervals, other procedures have also been developed, expanding the original iPLS concept [18]. In our study, we have used the algorithm developed by Brás

et al. [31], which combines the use of bootstrap methods with the concept of wavenumber distance for interval spectra selection. Its main guidelines are as follows: (i) the confidence intervals for the PLS regression coefficients are computed from a procedure called PLS-bootstrap; (ii) the first variable selection is then performed by excluding those variables whose confidence intervals, of the corresponding regression coefficients, do not contain the zero value; (iii) the indexes of selected variables are saved and the distances between consecutive values are calculated; (iv) the interval width is evaluated, so that if it exceeds a pre-established threshold, the spectral interval ends at that wavenumber, otherwise the interval width is enlarged; (v) the comparison process, described in the previous step, is conducted across the entire spectrum, and finally all the selected intervals are assembled. In order to avoid having intervals with very few variables, according to the specified interval width used in step (iv), and also to ensure the effective contribution of the selected variables, two additional criteria are considered for assessing the intervals previously obtained, as described in the next step; (vi) the first criterion regards the interval width, while the second one is based on the magnitude of the regression coefficients in those intervals. Intervals whose widths are lower than that arising from step (iv) are excluded. Then, for each of the remaining intervals, the regression coefficients are calculated separately, according to PLS-bootstrap, and compared with the overall average regression coefficients of all the variables selected in step (ii), excluding the variables in which the regression coefficients, computed in this last phase, is smaller than the one obtained in step (ii).

The procedure described above is included in our validation strategy. In particular, for each Monte Carlo resample trial, this variable selection methodology is carried out. Next, the selected variables are used to build a new PLS model from the training data set, which is then used to predict the testing data set. Once more, the average of prediction errors, obtained from all trials, is used as an estimate of model performance.

2.3.3. Genetic algorithm (GA)

The Genetic Algorithm tries to employ the principles of natural genetic selection to the variable selection problem. The basic idea underlying evolution theory is that individuals with a greater "fitness to the environment" have a greater probability of surviving and spreading their genetic material to the following generations. In this way, the genetic content of the "best" individuals (i.e. the ones showing higher fitness scores) will be dominant in the following generations. In the problem of variable selection, a similar situation is addressed, now involving variables whose combination may lead to the best characteristics, from the standpoint of a given criterion, linked to the explanation of the response variability [32]. Combined with PLS, the GA has been the variable selection method for spectral data which have been mostly used in applications [33–37]. In our study, we have used the algorithm proposed by Leardi [38], available at http://www.models.kvl.dk/GAPLS, that consists of the following steps: (i) first, a vector (commonly called chromosome) is computed, consisting of zeros and ones, with the same size as the number of variables. The randomly defined zeros and ones indicate which variables should be excluded and included in the model, respectively; (ii) next, the chromosome is perturbed randomly to make a number of other chromosomes that define the initial population. For each chromosome, (iii) a PLS model, using the selected original variables, is built, and evaluated by cross-validation. After creating and evaluating the initial population, (iv) the so-called reproduction step takes place. Here, a new population is made by recombining the initial chromosomes, following the idea that the chromosomes of the previous population, which give the best predictions, have a large chance of being copied, while the remaining chromosomes tend to disappear. Different GA methods propose different ways for shuffling the parent chromosomes, i.e. the variables that the offspring will receive from each parent. Leardi suggests drawing a random number for each variable (ranging from 0 up to 1) and does apply the following criterion: if the value is higher than 0.5, then the variable from parent #1 will be given, otherwise the variable will be given by parent #2. Finally, (v) the step simulating the mutations that occur in nature is also considered. A mutation is simply a change of a variable in a chromosome, and ensures that all variables can be selected in the coming generations, even if with a reduced probability (weak performance in previous generations). The mutation rate used was set to 1%. The whole process is repeated until a stopping criterion is met, which in our case corresponds to the consideration of 100 runs (generations). The difference between different runs relies on the probability that each variable has to be selected, which changes according to the frequency of selection in the previous runs. Another important feature of the algorithm proposed by Leardi relies on the fact that it takes into account the autocorrelation among adjacent spectral variables. This is done by smoothing the probability of selection with a moving average filter (window size 3). In this way, if a variable is thought to be relevant, the variables adjacent to it tend also to be relevant, and therefore their probability is also increased. The final model (vi) is obtained following a stepwise approach, in which the variables are selected according to a compromise between the value of the frequency of selections and the explained variance of the model established.

Since GA is mainly a stochastic algorithm, it is obvious that the final solutions of different GA's runs will not be exactly the same. Therefore, Leardi [38] suggests comparing the solutions of at last five solutions from different GA's, and getting an idea of the robustness of the method. In our study, we have followed this suggestion, and the entire GA was carried out 5 times in each Monte Carlo trial, and the five solutions were compared in terms of their prediction ability. The best result was then used for establishing a prediction model from the training data set and then evaluated in terms of performance to predict the respective testing data set in such a MC trial.

2.3.4. 2.3.4. Wavelet transformation combined with genetic algorithm (WT-GA)

Wavelet transforms (WT) aim to transform data from their original domain into another where some operations can be carried out in an easier way or data analysis can be carried out more effectively. The wavelet transform has been successfully applied in a wide range of application scenarios, such as signal processing and denoising [39–41], image processing [42], instrument standardization [43] and data compression [44]. In this study, we explore the wavelet transform ability to effectively compress signals with features exhibiting different locations and localizations in the time-frequency plane (such as UV–vis spectra), as a pre-processing step for the multivariate prediction tasks. Trygg and Wold [45] use the wavelet transform as a pre-processing method in PLS regression with NIR spectra, obtaining a good data compression ratio (of about 3%) with almost no loss of predictive information (predictive ability was basically the same as for the original uncompressed regression model). It should be noted that the wavelet transform, by itself, does not lead immediately to the compression of the original data set. Compression is only achieved by eliminating the low magnitude or low informative wavelet coefficients, by applying some sort of criterion. For accomplishing this step, we decided to use a genetic algorithm for selecting the more important wavelet coefficients for prediction purposes. The wavelet used is the simplest orthogonal wavelet with compact support, the Haar wavelet. Briefly, a wavelet transform involves the decomposition of a signal (e.g. UV–vis spectrum) into a low resolution representation of it, plus all the details that were lost

Table 1
Results obtained with the four variable selection methods applied to the test samples of Data Set #1. $\hat{y}$ and PI, the model's prediction estimate and their 95% coverage prediction interval, respectively; LV, the number of latent variables selected through Monte Carlo cross-validation; "% of reduction", the percentage of the original number of variables that are not used in the final version of the models; *RMSE-P*, root mean square error of prediction.

Sample	Full spectrum		VIPs[a]		IPLS[b]		GA[c]		WT-GA[d]	
	PI	$\hat{y}$	PI	$\hat{y}$	PI	$\hat{y}$	PI	$\hat{y}$	PI	$\hat{y}$
6yA	[4.5–9.2]	7.1	[4.12–8.93]	6.4	[3.23–8.49]	5.7	[3.73–8.97]	5.7	[4.11–8.54]	6.4
6yB	[4.8–8.7]	6.6	[3.03–7.92]	5.8	[1.92–7.87]	4.6	[2.13–7.99]	4.6	[2.97–7.38]	5.2
10yA	[9.7–13.9]	11.9	[6.35–11.03]	8.4	[5.52–10.94]	8.0	[5.82–11.09]	8.0	[5.83–10.21]	8.1
14yA	[11.6–15.4]	13.3	[9.59–14.61]	12.1	[9.75–15.88]	13.2	[9.08–15.07]	12.3	[9.81–14.51]	12.3
14yB	[14.5–18.6]	16.2	[10.40–15.40]	12.9	[10.12–16.28]	13.5	[9.56–15.58]	12.7	[10.97–15.44]	13.3
15yA	[13.6–17.9]	15.6	[15.56–21.19]	17.7	[15.02–22.14]	18.1	[15.32–22.61]	18.6	[14.04–19.20]	16.7
LV	5		5		4		4		3	
% of reduction	0%		60.6%		87.6%		87.6%		94.4%	
RMSEP (years)	1.51		1.52		1.71		1.92		1.36	

[a] Only VIPs larger than one are included in the model.
[b] The maximum distance between consecutive spectral variables (selected as relevant by the PLS-Bootstrap method) that can still be considered as belonging to the same interval is set equal to 15 nm. The minimum number of variables allowed in an interval is 3.
[c] The number of chromosomes used is 30. On average, 5 variables are "mutated" per chromosome and 30 is the maximum number of mutations. The probability of mutation used is 1%. The probability of cross-over is 50% and the number of generations or runs is 120.
[d] A decomposition depth of 5 was implemented in the Haar Wavelet decomposition. The GA parameters used in the analysis of the wavelet coefficients are the same as those indicated in footnote c.

across the successive decompositions considered. This is done by successively applying quadrature mirror filters to the original signal, and to the coefficients obtained after each decomposition stage, that provide increasingly coarser representations of the original signal. Each decomposition stage produces two signals of the same length: one for the coarser approximation of the signal (approximation coefficients), and another for the details lost in the decomposition (detail coefficients). Only the approximation coefficients obtained are further decomposed, in order to create two new coarser approximations and detail coefficients. This procedure is repeated until the desired decomposition depth is achieved, or only one approximation coefficient is obtained. For a signal of length 2^n the filtering procedure can be performed a maximum of n times, creating n different scales for signal representation [42,45]. The complete set of detail coefficients, at different scales and approximation coefficients at the coarsest scale, are then concatenated into a single vector of data, with the same number of entries as the original signal. In this study, for each Monte Carlo simulation carried out, the wavelet decomposition of the spectra of the training data set was performed, and a genetic algorithm applied in order to select the most important wavelet coefficients (approximation and detail) for each scale analyzed. Then, PLS regression is carried out using only the selected coefficients. The testing data set undergoes the same wavelet transformation and the same coefficients selected in the training are also picked up, after which the estimated PLS model is ran in order to predict the response values.

3. Experimental

3.1. Samples

A total of fifty-four samples of Madeira wines were analyzed in this study, thirty-one for addressing the long-term study (Data Set #1), and twenty-three for the short-term analysis of the ageing process (Data Set #2). Data Set #1 is formed by a training data set, which includes twenty five samples, covering the first nineteen years of ageing: 1 y (n = 3), 3 y (n = 3), 5 y (n = 3), 7 y (n = 3), 9 y (n = 2), 11 y (n = 3), 13 y (n = 2), 15 y (n = 1), 17 y (n = 2), 19 y (n = 2), and a testing data set, collected afterwards, formed by six samples with 6 y (n = 2), 10 y (n = 1), 14 y (n = 2) and 15 y (n = 1) of ageing in casks.

Table 2
RMSE-MC results, and their associated 95% confidence intervals, expressed in months, for the PLS models built from different variable selection methods, for the Data Set 2. The % of reduction indicates the percentage of wavelengths that have been discarded by the selection procedures.

Sample	Full spectrum		VIPs[a]		IPLS[b]		GA[c]		WT-GA[d]	
	RMSE-MC	IC	*RMSE-MC*	IC	*RMSE-MC*	IC	*RMSE−MC*	IC	*RMSE-MC*	IC
1	4.8	[4.67–4.84]	5.2	[5.11–5.28]	5.8	[5.76–5.93]	3.1	[2.98–3.15]	3.2	[3.14–3.61]
2	3.4	[3.27–3.43]	3.4	[3.33–3.49]	4.4	[4.31–4.54]	2.8	[2.69–2.84]	3.0	[2.94–3.14]
3	3.1	[3.02–3.21]	3.2	[3.10–3.27]	3.8	[3.74–3.91]	2.6	[2.54–2.72]	2.7	[2.59–2.76]
4	2.8	[2.69–2.85]	2.8	[2.73–2.92]	3.7	[3.59–3.80]	2.6	[2.50–2.65]	2.6	[2.52–2.70]
5	2.7	[2.64–2.82]	2.6	[2.54–2.75]	3.6	[3.47–3.78]	2.5	[2.42–2.59]	2.6	[2.46–2.67]
6	2.5	[2.36–2.55]	2.5	[2.37–2.56]	2.7	[2.57–2.82]	2.4	[2.35–2.54]	2.5	[2.37–2.58]
7	2.1	[2.05–2.24]	2.4	[2.32–2.52]	3.4	[3.19–3.70]	2.4	[2.53–2.55]	2.4	[2.28–2.47]
8	2.0	[1.96–2.13]	2.3	[2.17–2.40]	3.6	[3.41–3.86]	2.4	[2.35–2.52]	2.4	[2.27–2.47]
9	2.0	[1.19–2.06]	2.2	[2.13–2.31]	3.7	[3.53–3.93]	2.5	[2.40–2.57]	2.4	[2.25–2.48]
10	2.0	[1.91–2.08]	2.2	[2.08–2.92]	3.9	[3.71–4.14]	2.5	[2.42–2.59]	2.3	[2.22–2.37]
% of reduction										
3VL	0%		47%		43%		59%		97%	
7VL	0%		46%		63%		–		89%	

[a] VIPs larger than one are included in the model.
[b] The maximum distance between consecutive spectral variables (selected as relevant by the PLS-Bootstrap method) that can still be considered as belonging to the same interval is set equal to 12 nm. The minimum number of variables allowed in an interval is 4.
[c] The number of chromosomes used is 30. On average, 5 variables are "mutated" per chromosome and 30 is the maximum number of mutations. The probability of mutation used is 1%. The probability of cross-over is 50% and the number of generations or runs is 120.
[d] A decomposition depth of 5 was implemented in the Haar Wavelet decomposition. The GA parameters used in the analysis of the wavelet coefficients are the same as those indicated in footnote c.

Table 3
Results obtained by the four variable selection methods, when applied to all samples of Data Set #2. Samples are denoted with the grape variety indication, using the first letter of the grape variety name, followed by an indication of their ageing time in months (m). $\hat{y}$ and PI, the model's prediction estimate and its associated prediction interval with 95% coverage. #LV, the number of latent variables selected by Monte Carlo cross-validation; *RMSE-MC*, the root mean square error of cross-validation.

Sample	Full spectrum		VIPs[a]		IPLS[b]		GA[c]		WT-GA[d]	
	PI	$\hat{y}$	PI	$\hat{y}$	PI	$\hat{y}$	PI	$\hat{y}$	PI	$\hat{y}$
M0 m	[−3.54 to −1.96]	−1.4	[−7.31to −2.28]	−2.0	[−0.52 to −2.82]	0.9	[−6.05 to −5.93]	1.3	[−2.71−2.18]	−0.2
M2 m	[−2.78 to −2.34]	−0.6	[−7.30 to −2.18]	0.0	[0.45 to 3.79]	2.3	[−4.16 to −7.71]	3.5	[−1.88 to −2.84]	0.0
M8 m	[10.29 to 14.89]	13.3	[11.81 to 18.73]	16.4	[12.56 to 15.07]	13.9	[12.20 to 19.76]	16.4	[9.07 to 13.16]	11.4
M14 m	[9.23 to 14.70]	12.4	[9.03 to 18.32]	13.8	[11.98 to 14.85]	13.6	[6.70 to 18.18]	14.5	[10.62 to 14.62]	12.6
M17 m	[14.04 to 19.14]	16.7	[11.68 to 21.29]	15.8	[14.70 to 17.78]	15.8	[8.46 to 20.42]	16.2	[14.26 to 18.55]	16.8
M24 m	[22.42 to 27.48]	23.3	[13.38 to 23.11]	18.2	[20.31 to 23.68]	21.7	[12.62 to 22.43]	18.9	[22.07 to 26.42]	23.9
B0 m	[−6.55 to −1.22]	−3.2	[−7.18 to −2.28]	−1.7	[−1.4 to −2.08]	0.1	[−11.0 to −1.39]	−3.5	[−3.72 to −1.78]	−1.8
B6 m	[2.98 to 8.27]	7.2	[2.45 to 12.45]	8.6	[2.11 to 5.02]	3.5	[−0.33 to −11.22]	6.8	[3.66 to 7.84]	6.1
B12 m	[13.10 to 17.88]	16.0	[11.55 to 20.90]	17.3	[12.36 to 15.93]	14.2	[9.89 to 20.10]	15.8	[12.86 to 16.40]	14.7
B14 m	[15.57 to 20.04]	17.5	[10.82 to 21.41]	18.2	[13.85 to 17.37]	15.0	[10.55 to 22.13]	17.5	[13.96 to 18.21]	16.0
B19 m	[15.83 to 21.23]	16.9	[9.18 to 20.45]	14.8	[14.40 to 18.31]	16.2	[9.18 to 21.80]	16.5	[16.12 to 20.71]	18.1
B24 m	[17.09 to 22.55]	20.6	[14.04 to 25.98]	20.7	[18.97 to 23.01]	21.5	[14.14 to 25.94]	20.0	[19.49 to 23.90]	21.8
V0 m	[−1.18 to −4.06]	−0.2	[−7.81 to −2.55]	−1.5	[−2.83 to −0.28]	−1.2	[−8.59 to −3.25]	−1.9	[−3.47 to −1.20]	−0.7
V2 m	[1.64 to 6.49]	4.4	[0.04 to 8.83]	5.4	[5.73 to 8.35]	6.4	[−1.10 to −10.65]	6.3	[1.13 to 5.53]	3.8
V8 m	[6.27 to 11.10]	8.1	[4.47 to 13.95]	10.4	[6.35 to 9.44]	8.0	[4.53 to 15.01]	11.0	[6.59 to 10.36]	8.3
V14 m	[12.31 to 16.83]	14.2	[9.05 to 18.58]	15.7	[11.58 to 14.70]	13.9	[8.50 to 19.79]	15.7	[12.34 to 16.26]	14.4
V17 m	[11.28 to 15.40]	13.3	[8.42 to 18.08]	14.9	[13.83 to 16.89]	15.2	[7.78 to 19.10]	14.9	[13.20 to 16.92]	15.0
V24 m	[20.60 to 25.68]	22.1	[13.94 to 23.87]	20.2	[23.49 to 26.65]	24.9	[12.92 to 24.24]	20.0	[21.41 to 26.79]	22.7
S0 m	[−1.28 to −4.48]	1.0	[−8.3 to −1.87]	−3.0	[−2.9 to −0.04]	−1.3	[−11.5 to −0.49]	−3.9	[−0.5 to −3.83]	1.6
S10 m	[6.98 to 12.38]	9.8	[4.83 to 14.35]	9.5	[7.13 to 10.23]	8.9	[1.55 to 13.19]	8.9	[6.82 to 11.93]	8.9
S15 m	[10.38 to 15.84]	12.4	[7.32 to 16.75]	13.0	[10.95 to 13.93]	12.3	[6.01 to 17.41]	12.9	[12.09 to 17.09]	13.4
S17 m	[18.79 to 23.61]	22.0	[13.74 to 23.44]	19.5	[19.18 to 22.25]	21.0	[12.85 to 24.75]	19.9	[17.83 to 22.13]	20.5
S25 m	19.01 to 24.53]	22.8	[14.55 to 23.82]	20.7	[23.56 to 26.81]	25.6	[15.58 to 27.19]	22.4	[20.43 to 25.21]	22.7
LV	7		8		6		3		7	
RMSE-MC (years)	2.6		3.5		2.3		3.3		1.8	

[a] VIPs larger than one are included in the model.
[b] The maximum distance between consecutive spectral variables (selected as relevant by PLS-Bootstrap) that can still be considered as belonging to the same interval is set equal to 16 nm. The minimum number of variables allowed in an interval is 4.
[c] The number of chromosomes used is 30. On average, 5 variables are "mutated" per chromosome and 30 is the maximum number of mutations. The probability of mutation used is 1%. The probability of cross-over is 50% and the number of generations or runs is 120.
[d] A decomposition depth of 7 was implemented in the Haar Wavelet decomposition. The GA parameters used in the analysis of the wavelet coefficients are the same as those indicated in footnote c.

In order to distinguish samples with the same ageing period, we use letters A, B and C. All these wines are produced from the same grape variety (Malvasia) and were kept in casks under similar conditions. The samples were collected directly from the casks and stored at −20 °C, until they were analyzed. The two data sets were collected and analyzed at different times. Data Set #2 includes twenty-three samples collected during the first 24 months of ageing. In this case, wines are produced from four different grape varieties, specifically those recommended for Madeira wine production (Malvasia (M), Boal (B), Verdelho (V) and Sercial (S)). The sampling plan consisted in collecting samples approximately every six months, except for the first collection. However, this was not always possible. The effective sampling times are presented in Table 3, together with the model's age prediction.

3.2. UV–vis spectrophotometer analysis

The spectrum of each sample was recorded by means of a Lambda 2 Perkin-Elmer UV–vis spectrophotometer using 10 mm path-length quartz cells (Eppendorf). All samples were filtered through 0.45 μm PTFE syringe filters and diluted (1:20 or 1:50). Water was used for the reference scan. The spectral data were corrected for dilution and multivariate statistical analysis was performed over the absorbance measures.

3.3. Modeling and validation

As previously mentioned, PLS regression was used in order to establish an age prediction model. For Data Set #1, the response vector is the wine age, expressed in years, while for Data Set #2 the wine age is modeled in a month scale. For each pre-processed matrix of predictors, a prediction model was developed. Then, for the best model (in the sense of the established criterion), variable selection was performed, according to the four methods described in Section 2.3. In order to ensure that the more predictive variables will be selected, and simultaneously reduce the risk of overfitting, the proposed validation strategy included a step of variable selection in all of its resampling trials. The comparison between different models obtained was carried out according to their *RMSE-MC* values and the number of latent variables. All data analysis were implemented in Matlab (version 7.6, The Mathworks, Inc.) using home made code, the PLS-Toolbox package for data pre-processing, the algorithm sent by L. Brás for IPLs analysis, and the algorithm proposed by Leardi for GA, available at http://www.models.kvl.dk/GAPLS, and WaveLab 8.02 (available at http://www-stat.stanford.edu/∼wavelab/) for performing wavelet transformation computations.

4. Results and discussion

In this section, we present the results obtained for the analysis of the long-term and short-term wine ageing processes. We begin by addressing the long-term age prediction problem (Data Set #1), where the results obtained from using the full spectrum available, as well as those arising from the several variable selection methods, are presented and discussed. Then, we move on to the analysis of the short-term age prediction values, by employing analogous procedures, and present the results thereby obtained.

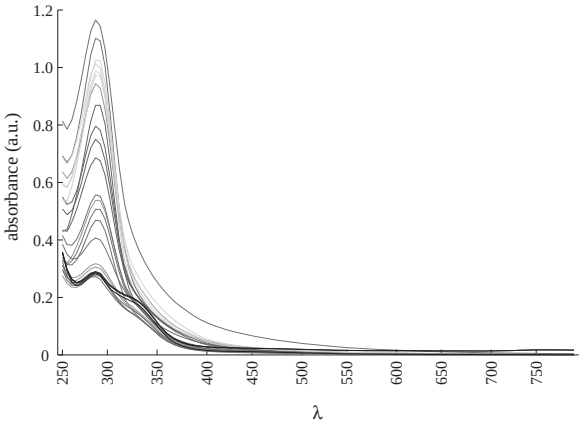

Fig. 1. UV–vis spectrum of different aged Madeira wines. Darker lines correspond to young wines, while pale lines match aged wines.

4.1. Long-term age prediction

Fig. 1 shows the absorbance spectra of samples of Madeira wines with different ageing periods, scanned from 245 to 785 nm, at 5 nm intervals. We have used a grey-scale color code, in order to facilitate the discrimination between young and older wines: the darker tones indicate young wines spectra, while the lighter ones represent aged wine spectra. From the analysis of Fig. 1, it is possible to verify that the UV–vis electromagnetic spectrum does indeed contain distinguishing features for wines with different ageing periods. The main differences typically rely on two absorbance regions, around 280 nm and 320 nm. These bands are closely connected with the wine phenolic and furan compositions, which are responsible, to some extent, for wine color and astringency. However, in order to effectively model the relationship between spectral data and wine age, the analysis will encompass, in a first stage, the entire spectrum, after which alternative approaches, that select and explore the most explicative spectral regions, will be employed.

The samples of Data Set #1 were already employed for establishing an age prediction model, in a previous work [16], where the authors proposed a PLS model with 5 latent variables. The prediction accuracy achieved lies in the range of 1.4 years. We will now test such a model, for the first time, with an independent data set collected subsequently, in order to evaluate, in a more consistent and unbiased way, its predictive performance, as well as those for the other age prediction models, derived from the several variable selection approaches under scrutiny. All models (denoted here simply by their acronyms) are compared in terms of their root mean square error of prediction *RMSE-P*, and the effective number of wavelengths employed in prediction. These results are summarized in Table 1, together with the model prediction estimates for each test sample, and the corresponding prediction intervals. Furthermore, the wavelength regions selected by each method are presented in Fig. 2, where mean spectra are also represented. All models presented were developed for the best pre-processing method found in the analysis of the full spectrum, namely the Savitzky–Golay approach using a second derivative, based on a

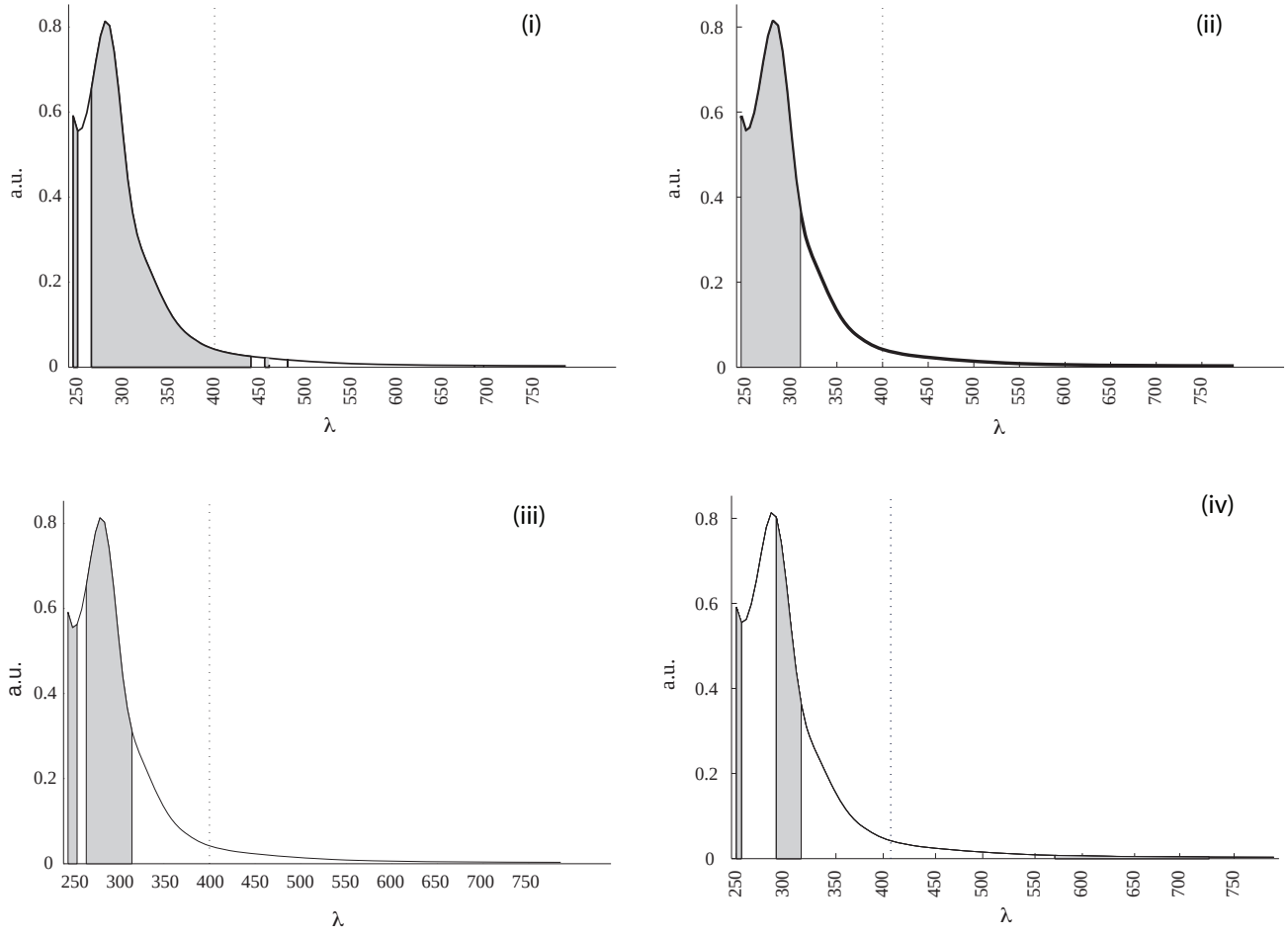

Fig. 2. Selected wavelengths over the mean of all aged training wines samples spectrum, for the different variable selection approaches tested: (i) VIPs, (ii) IPLs, (iii) GA, and (iv) WT-GA.

5-point window and a second-order polynomial for smoothing. The Bootstrap residuals methodology was also applied, in order to estimate the associated sample-specific prediction intervals [16].

The results obtained show that the full spectrum model presents a predictive performance on the test set which is very close to the one obtained during the validation phase, where only training samples were used in implementation of the cross-validation methodology, namely 1.5 and 1.4 years, respectively. This result confirms our previous conclusions, showing that our model is indeed suitable for MW long-term wine age prediction, as well as validates the correctness of its estimated accuracy. Even if there were some doubts regarding this issue, arising from the application of a Monte Carlo (MC) validation approach, which may still lead, sometimes, to slightly optimistic predictions, the analysis of an independent data set clarifies this issue, and allowed us to establish the suitability of such a model in the present case.

However, we are also aware of the fact that this model may contain a large number of uninformative wavelengths, which may be deteriorating its predictive performance to some extent. As presented in Section 2.3, there are systematic strategies available for performing variable selection. Among all methods tested, WT-GA has positively stood out: it presented the lowest *RMSE-P* score, used the smallest number of latent variables and employed the highest wavelength compression rate. It should be noticed that the WT-GA model works over wavelet coefficients, namely using only six of them, which correspond to the highlighted regions in Fig. 2-iv. As expected, the relevant wavelengths fall within 280–320 nm, closely connected to the presence of furanic, hydroxycinammics and hydroxybenzoic acid compounds, whose maximum absorbance in UV–vis spectra are located in such a range. On the other hand, the VIPs and IPLs methods led to broad bands, also covering wavelengths located in the visible region of the spectrum, while the GA model also selected the left half of the broad band, which was not selected by WT-GA.

One important issue on the GA evaluation is the definition of the number of runs (generations) which should be analyzed during each Monte Carlo simulation, in order to get a reliable model, that does not overfit data. Reference [36] suggests the realization of randomization tests, in order to consolidate a final decision. The idea is to evaluate whether important information, or just noise, is being modeled when the number of evaluations increases. To have an idea about when GA should be stopped, the performance of the average of the GA runs with the original response were compared with a similar number of shuffled responses (randomization tests). When such a difference is maximum, the corresponding number of runs is considered to be the greatest to effectively model the predictor information. In our case, the number of runs was set to 120. Fig. 3 presents the results of the different GA runs using such a number of evaluations. Looking at the selected variables, one would notice that selections performed by different replicates are rather consistent on the selected wavelengths, and, as a final choice, the selected wavelengths are those that appear, at least, in four of the five replicates analyzed.

Besides the superior performance of the WT-GA model, all the prediction intervals derived for the test samples do include the true wine age, even though their mean amplitude is the smallest among all the methods tested. For the remaining variable selection methods, this was not achieved for one of the test samples. Globally, the worst results were obtained for the GA model. This leads one to believe that the GA methodology, when applied to the original spectrum, leads to instabilities in wavelength selection that limit the model's performance. On the other hand, GA revealed to be very useful for proper wavelet coefficients selection. In this case, the results achieved were very satisfactory, leading to the most parsimonious model among all those derived, which also shows the best predictive performance and presents

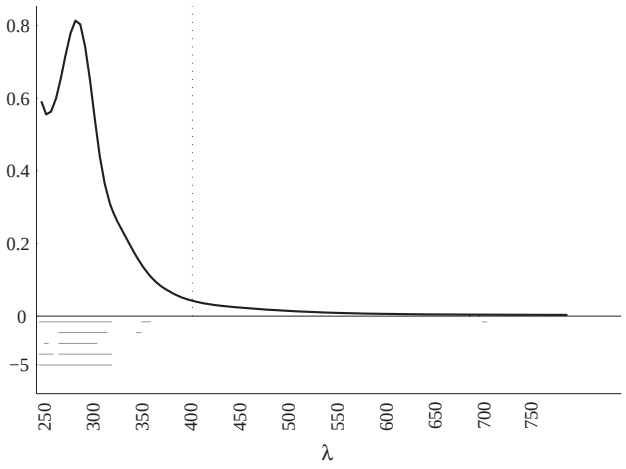

Fig. 3. Selected wavelengths from five GA models computed using the same parameters, from the top to the bottom lines: 19 variables, 11 variables, 10 variables, 14 variables and 15 variables.

full consistency between predictions and observations in the test set. Finally, we would like to underline the following important result: the MW age can be now predicted within an error of 1.36 y, from absorbance measurements. Therefore, we have developed a procedure based upon a fast, low-cost and easy-to-use analytical characterization technique, as a handy and reliable solution for addressing the problem of wine age prediction.

4.2. Short-term age prediction

Fig. 4 presents the spectra for all samples that constitute Data Set #2, scanned from 210 to 800 nm, at 2 nm intervals. Comparing the spectra of wines produced from different grape varieties, different absorbance characteristics are observed, during the first twenty-four months of ageing. We also find that differences across ageing time, for a particular type of wine, are not so clear. The set of Boal wine samples shows the lowest absorbance peaks, both around 280 and 320 nm, while the set of Malvasia wines does have the highest absorption wavelengths. The Sercial wine spectrum displays a similar trend, when compared with Malvasia wines, but does present lower absorbance values. Both show a well-defined absorbance band at 320 nm, disclosing the influence of hydroxycinnamic acids in wines produced from those grape varieties. These compounds are primarily located in the solid parts of the berry, and their contents decrease with maturation, thus being frequently used to differentiate grape cultivars [46]. The absorbance at 280 nm is commonly connected with the total phenolic content and the presence of furanic compounds (such as HMF, furfural, acetylfuran, 5-methylfurfural) [47]. In the particular case of Madeira wines, the latter compounds are considered to play a rather well-known role [14]. These are the main degradation products of carbohydrates, and their occurrence is related to the Madeira wine's typical non-enzymatic browning reactions, namely Maillard type of reactions, sugar degradation in an acid medium. For Malvasia and Boal samples (the sweetest wines), the evolution of the absorption peak around 280 nm is rather evident, while for Sercial (dry wines) slight variations in absorbance at 280 nm are observed, across the ageing period analyzed. This fact is closely connected to the wine's sugar content and to the above mentioned reactions, that take place in the wine ageing process. The analysis of these variations, together with other less noticeable, was done together, in order to find a parsimonious model for characterizing the wine ageing process, in particular for predicting wine age. In this regard, the effects of pre-processing and scaling (mean centering and autosclaing) on

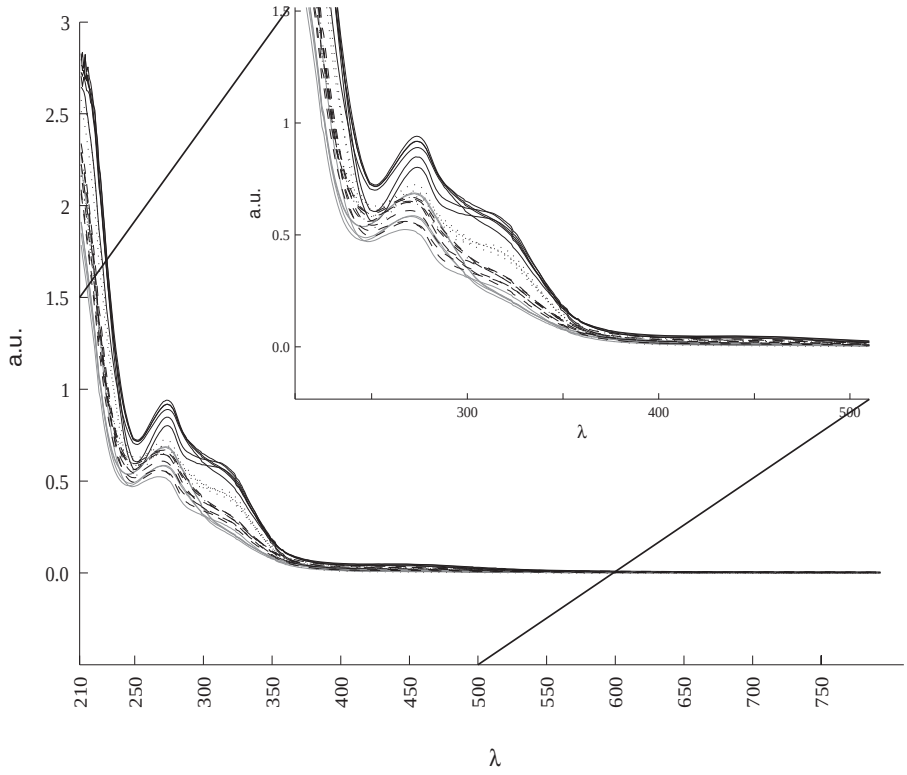

Fig. 4. UV–vis spectrum of different Madeira wines during the first 24 months of ageing: Malvasia wines – black line; Boal – grey line; Verdelho – dashed black line; Sercial – grey dotted line.

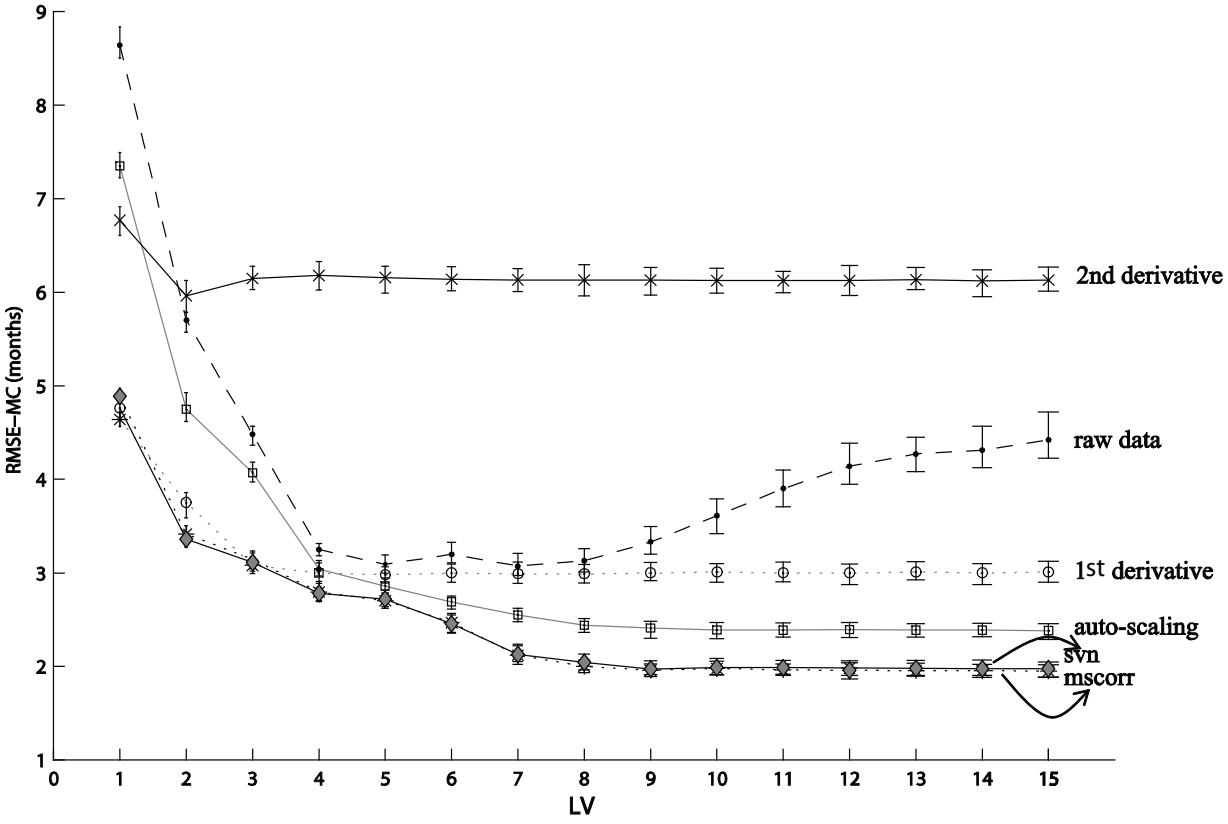

Fig. 5. Values for *RMSE-MC* obtained from the application of different pre-processing methods to the predictor matrix, as a function of the PLSR model dimensionality.

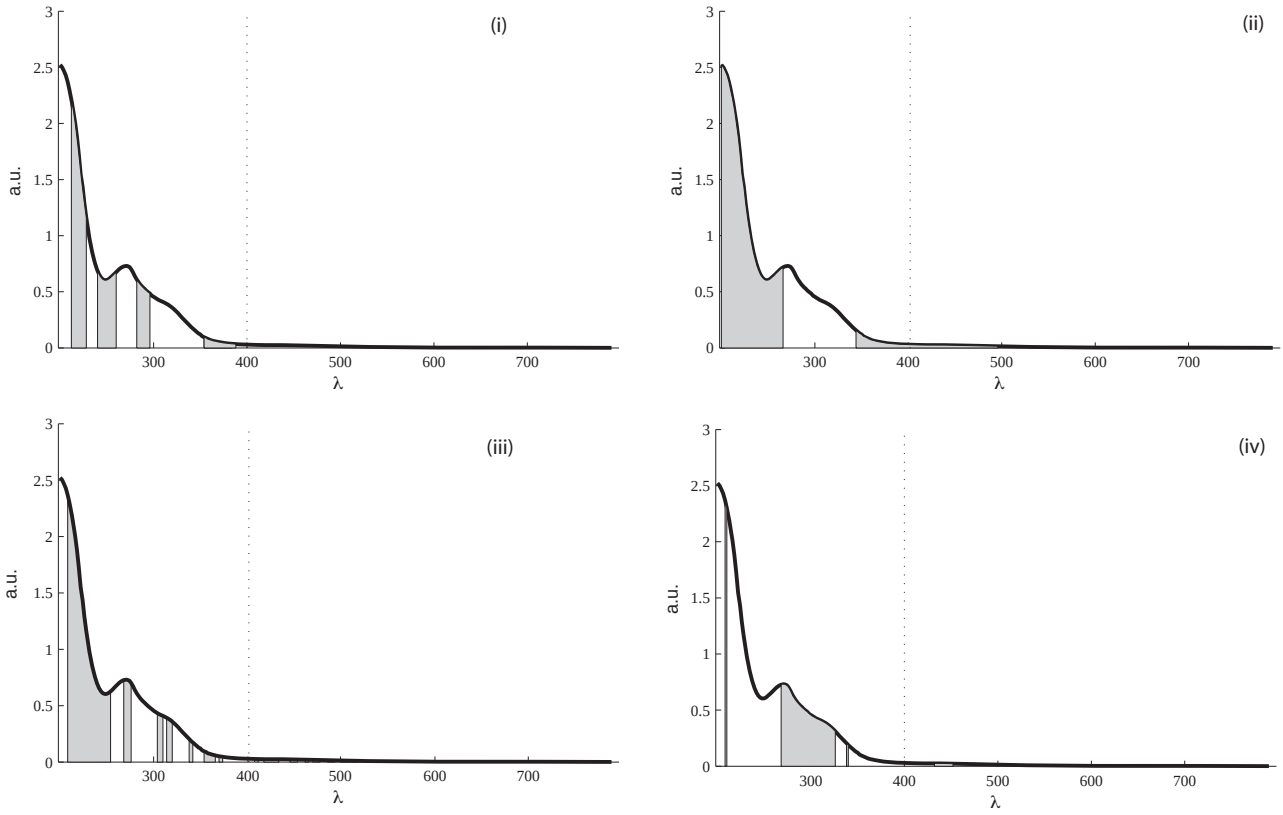

Fig. 6. Selected wavelengths over the mean of all training wines samples spectrum, during the first 24 months of ageing, for the different variable selection approaches tested: (i) VIPs, (ii) IPLs, (iii) GA, and (iv) WT-GA.

the performance of PLS age prediction models were preliminarily studied. First, all the spectrum information was considered in the development of the models, and age prediction models were developed for different pre-processing methods (involving only the predictors matrix). All models were tested and evaluated as a function of the number of latent variables, based on the Monte Carlo cross-validation approach. Fig. 5 shows the results obtained for the root mean square error of validation, *RMSE-MC*, as a function of the PLS model dimensionality.

Concerning scaling, autoscaled data on average perform better than no scaling (i.e. just performing mean centering). The best pre-processing approaches are SNV and MSC (no significant difference can be observed between them), improving the prediction ability achieved with raw data (not pre-processed) by about 31.5%. In this case, the derivative approaches do not lead to better results, probably due to the fact that they increase the level of noise to some extent, worsening the model's prediction ability. Different numbers of points were tested for the smoothing filter (i.e. for its width), as well as different polynomial degrees (first and second order). The results presented in Fig. 5 correspond to the best set of these parameters: a 10-point, quadratic Savitzky–Golay smoother. Also, derivatives were found not to be advisable in the pre-processing stage, especially when the variable selection approaches are very sensitive to noise, such as GA [38].

According to the results obtained, it is possible to predict wine age during the first two years of ageing, with a *RMSE-MC* of 2.1 months, using a PLS model with 7LV that encompasses all UV–vis absorbance data. As a large number of predictors is involved in such a regression model, it was decided to test whether prediction models built from variable selection methodologies can simplify model complexity and improve predictions. Therefore, the same variable selection approaches used for the long-term age prediction problem were also tested for Data Set #2 (short-term ageing period), after applying SNV pre-processing. In Table 2 we sum up the results obtained for the cross-validation analysis, regarding the best pre-processing method for the full spectrum model and for the variable selection methodologies. Comparing all models with 3LV, we can verify that GA and WT-GA provide the best age prediction results, without any statistical significant difference between them. In terms of the ability to eliminate irrelevant wavelengths from the models, the VIPs, IPLs, GA and WT-GA methods lead to 47%, 53%, 57% and 97% of "variable compression", respectively. Fig. 6 portrays the selected regions relative to each method. Undoubtedly, WT-GA arises, once again, as the best model for wine age prediction. When the number of LV is allowed to increase to more than 3, namely to the minimum of the *RMSE-MC versus* LV curve, improvements of 35%, 31%, 29%, and 11% in *RMSE-MC* were achieved for the full spectrum model (8LV), VIPs (9VL), IPLs (6VL) and WT-GA (7VL) models, respectively. For GA, the *RMSE-MC* obtained with 3LV is not statistically significantly different from the lowest *RMSE-MC*, according to the confidence limits computed. In this scenario, IPLs show the worst performance, while the remaining methods have closer scores. Our decision regarding the PLS model dimensionality is based upon the number of latent variables leading to the lowest *RMSE-MC*. The sample-specific prediction intervals were then computed. Table 3 presents the estimates obtained for the wine age of all the twenty-three samples, along with their associated 95% prediction intervals. These intervals were computed using the Bootstrap residual approach, as described in [16]. The strategy of combining wavelet decomposition with a genetic algorithm, in order to select informative wavelengths for wine age prediction, has revealed itself, once again, very promising. The WT-GA model is the one presenting a prediction error lower than 2 months.

However, it should be noted that, occasionally, there are some prediction intervals that do not contain the observed values. In fact, this leads us to consider that the methodology followed may be somewhat optimistic in its prediction inferences. Not excluding the hypothesis of a slightly optimistic scenario, we also consider that a more extensive and representative set of samples might be needed for further model validation regarding short-term ageing trends for these types of wines. Perhaps the rigid criteria of selecting the models dimensionality through the lowest *RMSE-MC* might also contribute to somewhat optimistic predictions, as only minor performance enhancements are sometimes verified during cross-validation, regarding more parsimonious solutions. Therefore, we have redone the whole analysis, using 4LV for the full spectrum model and VIPs, and 3VL for WT-GA models. For IPLs and GA, according to the results of validation, we decided to maintain the dimensionality initially defined. As expected, the *RMSE-P* increases a bit, namely to 3.2, 3.4 and 3.0 months for full spectrum, VIPs and WT-GA, respectively. However, the sample-specific prediction intervals seem now more consistent, regarding their ability to contain the true value of the wines age. We would like to point out that the latter models do provide wine age prediction estimates with good accuracy. We also consider that it is relevant to proceed with an independent validation study, as carried out for the long-term ageing prediction analysis, in order to further test and provide additional evidence on the potential of UV–vis absorbance for wine age prediction in short-term ageing periods.

5. Discussion

The results described in the previous sections show that it is indeed possible to reach good prediction accuracies for wine age from UV–vis absorbance data. Our strategy dedicated a special attention to the selection of the predictor variables to be used in the models, as well as to the high collinearity existing between spectrum variables and the undesired scatter effects on recorded spectra. Thus, different pre-processing and variable selection methodologies were tested. The results of such an analysis pointed to a trade-off between the performance of models hence obtained and the number of parameters that they use. We also verified that sometimes better prediction errors for more complex models (with more parameters) came at the expense of data overfitting, leading to worse prediction ability for the test data set. Therefore, our final decision was made based on a compromise between model complexity and improvement in prediction accuracy.

In both of our case studies, regarding long and short-term ageing periods, the results obtained reveal the potential of using an wavelet transform over pre-processed signals, along with a variable selection strategy (GA) and PLS regression. It was shown that a compression rate of 95% of the original matrix size is indeed possible, without any significant loss of predictive power. The performances achieved were indeed very satisfactory, leading us to believe that our methodology can be implemented as a practical tool for wine age prediction, providing further assistance to the oenologists' quality evaluations as well as for process control and improvement purposes, or quality assurance purposes.

6. Conclusions

The current study sought to investigate the application of chemometric methods to UV–vis spectra of wine samples, in order to develop predictive models for determining wine age in the long-term and short-term. The rationale behind the proposed approaches, is essentially to develop and configure a prediction tool that, once trained, will requires only a few operations to provide reliable results for wine age prediction. The matrix used in this study is a fortified wine, because of the importance to properly accomplish wine ageing monitoring in this type of high-value products.

The results obtained in the present study reveal and reinforce the potential of the proposed methodology built around UV–vis spectroscopy. We have found out that it is possible to predict MW age with an accuracy of 1.4 y for longer ageing periods, and with 3 months for wines of age comprised in the first two years of ageing in casks. Together with such performances, the main advantages of the proposed strategy are its handiness and affordability for wine makers. This method can be performed on site, and the results can be obtained in minutes, and will provide reliable information on how well the product is doing, in terms of the expected ageing trend.

Acknowledgments

A.C. Pereira acknowledges the Portuguese Fundação para a Ciência e Tecnologia for its financial support through PhD grant SFRH/BD/28660/2006. The authors also acknowledge Madeira Wine Company for kindly supplying all the wine samples used in this study. The authors thank Dr. Lígia Brás for kindly sharing the code for the iPLS algorithm.

References

[1] J. Saurina, Trends Anal. Chem. 29 (2010) 234–245.
[2] C.J. Bevin, R.G. Dambergs, A.J. Fergusson, D. Cozzolino, Anal. Chim. Acta 621 (2008) 19–23.
[3] D. Ballabio, T. Skov, R. Leardi, R. Bro, J. Chemom. 22 (2008) 457–463.
[4] P. Paneque, M.T. Álvarez-Sotomayor, A. Clavijo, I.A. Gómez, Microchem. J. 94 (2010) 175–179.
[5] D. Cozzolino, H.E. Smyth, W. Cynkar, R.G. Dambergs, M. Gishen, Talanta 68 (2005) 382–387.
[6] S. Pérez-Magariño, M. Ortega-Heras, M.L.G.-S. José, Z. Boger, Talanta 62 (2004) 983–990.
[7] L. Liu, D. Cozzolino, W.U. Cynkar, R.G. Dambergs, L. Janik, B.K. O'Neill, C.B. Colby, M. Gishen, Food Chem. 106 (2008) 781–786.
[8] S.A. Bellomarino, X.A. Conlan, R.M. Parker, N.W. Barnett, M.J. Adams, Talanta 80 (2009) 833–838.
[9] C. García-Jares, B. Médina, Analyst 120 (1955) 1891–1893.
[10] M.M.D.L. Urbano, d.C.P.M. Pérez, J. García-Olmo, M.A. Gómez-Nieto, Food Chem. 97 (2006) 166–175.
[11] F.J. Acevedo, J. Jiménez, S. Maldonado, E. Domínguez, A. Narváez, J. Agric. Food Chem. 55 (2007) 6842–6849.
[12] A. Rudnitskaya, I. Delgadillo, A. Legin, S.M. Rocha, A.M. Costa, T. Simões, Chemom. Intell. Lab. Syst. 88 (2007) 125–131.
[13] A. Rudnitskaya, S.M. Rocha, A. Legin, V. Pereira, J.C. Marques, Anal. Chim. Acta 662 (2010) 82–89.
[14] A.C. Pereira, M.S. Reis, P.M. Saraiva, J.C. Marques, Anal. Chim. Acta 660 (2010) 8–21.
[15] A.W. Vivian, C.E. Butzke, R.B. Boulton, J. Agric. Food Chem. 51 (2003) 7738–7742.
[16] A.C. Pereira, M.S. Reis, P.M. Saraiva, J.C. Marques, Chemom. Intell. Lab. Syst. 105 (2011) 43–55.
[17] Å. Rinnan, d.F.v. Berg, S.B. Engelsen, Trends Anal. Chem. 28 (2009) 1201–1222.
[18] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14–32.
[19] C.M. Andersen, R. Bro, J. Chemom. 72 (2010) 8–737.
[20] J.-P. Gauchi, P. Chagnon, Chemom. Intell. Lab. Syst. 58 (2001) 171–193.
[21] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[22] S. Wold, J. Trygg, A. Berglund, H. Antti, Chemom. Intell. Lab. Syst. 58 (2001) 131–150.
[23] M. Andersson, J. Chemom. 23 (2009).
[24] M.S. Reis, P.M. Saraiva, J. Chemom. 18 (2004) 526–536.
[25] T. Næs, T. Isaksson, T. Fearn, T. Davies, A User-friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK, 2002.
[26] I.S. Helland, T. Naes, T. Isaksson, Chemom. Intell. Lab. Syst. 29 (1995) 233–241.
[27] S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[28] S.N. Thennadil, E.B. Martin, J. Chemom. 19 (2005) 77–89.
[29] I.-G. Chong, C.-H. Jun, Chemom. Intell. Lab. Syst. 78 (2005) 103–112.
[30] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[31] L.P. Brás, M. Lopes, A.P. Ferreira, J.C. Menezes, J. Chemom. 22 (2008) 695–700.
[32] R. Leardi, J. Chromatogr. A 1158 (2007) 226–233.
[33] R. Leardi, M.B. Seasholtz, R.J. Pell, Anal. Chim. Acta 461 (2002) 189–200.
[34] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311–317.

[35] A.P. Ferreira, T.P. Alves, J.C. Menezes, Biotechnol. Bioeng. 91 (2005) 474–481.
[36] R. Leardi, A.L. González, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[37] P. Wiegand, R. Pell, E. Comas, Chemom. Intell. Lab. Syst. 98 (2009) 108–114.
[38] R. Leardi, J. Chemom. 14 (2000) 643–655.
[39] L. Pasti, B. Walczak, D.L. Massart, P. Reschiglian, Chemom. Intell. Lab. Syst. 48 (1999) 21–34.
[40] F.T. Chau, T.M. Smith, J.B. Gao, C.K. Chan, Appl. Spectrosc. 50 (1996) 339–347.
[41] M.S. Reis, P.M. Saraiva, B.R. Bakshi, Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier, Oxford, 2009, 25–55.
[42] M.S. Reis, A. Bauer, Chemom. Intell. Lab. Syst. 95 (2009) 129–137.
[43] B. Walczak, E. Bouvresse, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 41–51.
[44] U. Depczynski, K. Jetter, K. Molt, A. Niemöller, Chemom. Intell. Lab. Syst. 47 (1999) 179–187.
[45] J. Trygg, S. Wold, Chemom. Intell. Lab. Syst. 42 (1998) 209–220.
[46] M.V. Moreno-Arribas, M.C. Polo, Wine Chemistry and Biochemistry, Springer-Verlag, New York, 2008.
[47] A. Antonelli, F. Chinnici, F. Masino, Food Chem. 88 (2004) 63–68.