JOURNAL OF
AGRICULTURAL AND
FOOD CHEMISTRY

Use of Near-Infrared Spectroscopy and Feature Selection Techniques for Predicting the Caffeine Content and Roasting Color in Roasted Coffees

CONSUELO PIZARRO,*[,†] ISABEL ESTEBAN-DÍEZ,[†]
JOSÉ-MARÍA GONZÁLEZ-SÁIZ,[†] AND MICHELE FORINA[§]

Department of Chemistry, University of La Rioja, c/Madre de Dios 51,
26006 Logroño (La Rioja), Spain, and Dipartimento di Chimica e Tecnologie Farmaceutiche ed
Alimentari, Università di Genova, Via Brigata Salerno (Ponte), 16147 Genova, Italy

Near-infrared spectroscopy (NIRS), combined with diverse feature selection techniques and multivariate calibration methods, has been used to develop robust and reliable reduced-spectrum regression models based on a few NIR filter sensors for determining two key parameters for the characterization of roasted coffees, which are extremely relevant from a quality assurance standpoint: roasting color and caffeine content. The application of the stepwise orthogonalization of predictors (an “old” technique recently revisited, known by the acronym SELECT) provided notably improved regression models for the two response variables modeled, with root-mean-square errors of the residuals in external prediction (RMSEP) equal to 3.68 and 1.46% for roasting color and caffeine content of roasted coffee samples, respectively. The improvement achieved by the application of the SELECT-OLS method was particularly remarkable when the very low complexities associated with the final models obtained for predicting both roasting color (only 9 selected wavelengths) and caffeine content (17 significant wavelengths) were taken into account. The simple and reliable calibration models proposed in the present study encourage the possibility of implementing them in online and routine applications to predict quality parameters of unknown coffee samples via their NIR spectra, thanks to the use of a NIR instrument equipped with a proper filter system, which would imply a considerable simplification with regard to the recording and interpretation of the spectra, as well as an important economic saving.

KEYWORDS: Multivariate calibration; NIRS; feature selection; stepwise orthogonalization; roasted coffee; quality assurance; roasting color

INTRODUCTION

Coffee is one of the three most traded commodities and the second largest commodity industry worldwide. Nevertheless, although coffee is the world’s most heavily consumed beverage after water, with over 400 billion cups consumed annually, most consumers know little about its profound underlying chemical complexity, affected by a huge number of factors (including bean production, coffee type, roasting process, and cup preparation) that can determine the quality of a specific roasted coffee and, consequently, mark the difference between a great cup of coffee and a mediocre one (*1*−*7*).

In particular, the roasting process can be probably considered the most important step in the long quality chain from agriculture to industry influencing the final quality of coffee. During this process coffee beans undergo a deep physicochemical transformation as a result of many pyrolytic and complex reactions that take place and lead to the formation of a wide range of substances the combination of which will be responsible for their final sensory properties. The most influential parameter that determines the roasting process is the quantity of heat transferred to the beans, which can be controlled by the roasting temperature and time. In practice, the color of the beans can be used as an indicator of the degree of roasting to define the end of the operation, as the intensity of this color is directly correlated to the final roasting temperature: the higher the temperature, the darker the coffee. Thus, the roasting color as roasting control parameter plays a crucial part in the development of essential organoleptic attributes: it can be used to change the bitterness to acidity ratio (taste); it effects the formation and degradation of the different volatiles (aroma); and it conditions the completion of all pyrolysis reactions and, therefore, the final composition of roasted coffee and the partial or total development of many organoleptic characteristics.

On the other hand, caffeine content does not seem to have any direct effect on coffee sensory quality, because, despite its marked bitter taste, the amount of caffeine present in coffee

* Corresponding author (e-mail consuelo.pizarro@unirioja.es; telephone +34 941299626; fax +34 941299621).
[†] University of La Rioja.
[§] Università di Genova.

10.1021/jf071139x CCC: $37.00
Published on Web 08/14/2007

beverages accounts for only a relatively small proportion (<10%) of perceived bitterness. Nevertheless, as caffeine content displays a marked difference interspecies, the amount of caffeine present in a specific variety or blend of coffee may become an index of its quality. The relevance of the physiological effects of caffeine on the human body (stimulation of the central nervous system) should also be noted (caffeine is the main alkaloid present in coffee and it must therefore be determined).

Thus, the proven relevance of both roasting color and caffeine content as coffee quality parameters justifies the need for suitable techniques to accurately determine them and thus establish roasted coffee quality. Despite their reliability, the established analytical methods usually employed in the coffee industry for assessment of these descriptors may be fairly elaborate, costly, and/or time-consuming, requiring sample preparation and even chemical manipulations, so that key product quality parameters should be determined offline in a laboratory. Therefore, for practical reasons, simpler and faster methods, such as those based on spectroscopic techniques, capable of being easily implemented in coffee routine analyses to provide real-time measurements, would provide a very attractive and useful alternative tool for quality control and process improvement.

In this context, over the past decades, the application of near-infrared spectroscopy (NIRS) as a fast and nondestructive alternative analytical tool for the compositional, functional, and sensory analysis of both raw materials and final products in the food and agricultural industries has become widespread thanks to the advances in chemometrics. Thus, a large number of NIR applications have been reported in relation to quality and authenticity studies of many food and agricultural commodities, partly focused on the rapid and noninvasive determination of distinct types of chemical constituents and process parameters (*8*−*15*). As far as coffee quality assurance is concerned, the feasibility of using near-infrared spectroscopy, in combination with multivariate classification analysis, for coffee authentication purposes has been examined with notable success (*16*−*21*). Nevertheless, strangely, few studies are found in the literature specifically focused on the development of NIR-based calibration models aimed at simultaneously determining multiple quality features in coffee samples (*22*−*25*).

For this reason, the present study tried precisely to examine the potential of near-infrared spectroscopy, in combination with multivariate calibration techniques, to be used as a fast and accurate online monitoring tool to provide near real-time predictions of two parameters—roasting color and caffeine content—which are essential for both controlling the roasting process and assessing roasted coffee quality and quality changes.

To accomplish this aim, it must be taken into account that the actual applicability of NIRS to the online monitoring of coffee quality parameters would crucially depend on the robustness and reliability associated with the multivariate calibration models finally constructed, which would be, as well, limited by certain particular properties inherent to NIR spectral data: (1) the very high ratio between the number of variables and number of samples increases the risk of chance correlations and overfitting; (2) redundancy or collinearity between variables can lead to a model with poor predictive ability (although it is also true that the synergy of correlated variables can improve the stability of the model); (3) interference of systematic variation in the spectra unrelated to the modeled response will not only degrade the reliability of predictions but also affect its robustness over time. Such unwanted variation (noise) present in NIR data can come from different sources, including scatter effects, baseline drifts, or wavelength regions of low information content.

Thus, the application of suitable preprocessing ("correction") methods, aimed at resolving all of these limitations and improving the robustness and prediction performance of multivariate calibration models based on NIR spectra by removing useless spectral variation and extracting relevant information from the data, should be considered a crucial step prior to the development of calibration models (*26*−*28*).

Standard methods for calibration of near-infrared spectra, such as partial least-squares, based on the assumption that a set of underlying latent variables drives the changes in the system under study, have been traditionally applied using the full set of available spectral wavelengths without performing any previous feature selection (*29*). In recent years, it has been instead shown that an efficient variable (wavelength) selection prior to the calibration step can be greatly beneficial in providing more robust, reliable, and parsimonious models and considerably simplifying data acquisition and analysis (*30*, *31*).

Obviously, chemometrics offers a broad variety of alternative spectral tools, besides variable selection, to optimize PLS calibrations. In particular, there have been many attempts to develop spectral pretreatment methods to minimize the effects of variations in the spectral data that are not related to the chemical information contained. However, although many of these spectral filtering methods have often proven to improve calibration model performance, it must be taken into account that they ignore the fact that there might be spectral regions that do not contain any relevant information so that their presence will inevitably complicate the regression model.

In general, PLS performance will be improved when nonessential spectral features that contain predominantly noise are removed. Once such useless variables have been discarded in such a way that noise constitutes a minor fraction of the signals, there would not be a pressing need for selecting individual wavelengths to pursue an improvement in model predictive ability, because the error on the response as measured by the reference technique would be the limiting factor. Nevertheless, under these conditions, variable selection might provide simpler and more robust calibration models with at least the same reliability. On the other hand, the application of effective variable selection methods in quality control applications, such as that here presented, to generate highly parsimonious models offers additional advantages in terms of economy, with a notable reduction in the number of required sensors and the resulting simplification with regard to spectra acquisition and interpretation that this fact implies.

Therefore, all in all, the scope of this work was to explore the feasibility of applying variable selection techniques to extract from NIR spectra a minimum number (maximum parsimony) of informative predictors that will serve as the basis for developing robust and reliable calibration models to measure in real time two essential quality attributes of roasted coffees: roasting color and caffeine content. Several variable (wavelengths) selection techniques have been tested in the search for models with as high as possible performances (1) including both methods based on PLS regression, such as iterative predictor weighting (IPW) (*32*), interactive stepwise elimination (ISE) (*33*), and uninformative variable elimination (UVE) (*34*); (2) and approaches based on ordinary least-squares (OLS) regression, such as genetic algorithms coupled with OLS regression (GA-OLS) (*35*, *36*), and stepwise orthogonalization of predictors (SELECT) (*37*, *38*).

An important part of this study, it should be emphasized the novel application of an "old" technique recently revisited, stepwise orthogonalization of predictors (SELECT), that has barely been applied in regression problems despite the fact that it could provide significant advantages over other variable selection techniques commonly used in spectroscopy. For each response study, a comparison is made of the respective results yielded by the PLS model based on the "full spectrum" approach (i.e., using the full NIR wavelength range) and the diverse models constructed after feature selection.

It should be clarified that the aim of this study was not to provide definitive and immutable NIR calibration models for the on-line monitoring of roasting color and caffeine content in roasted coffee samples, but to propose an effective methodology capable of accomplishing this task and proving its reliability to predict both quality parameters. Although the data set used in this work was designed to cover insofar as possible the great variability inherent in commercially available coffee samples by considering different roasting conditions and degrees, and varied origins, it is quite clear that the dynamic nature of the coffee market and the particular needs and production lines of a given company would demand a more exhaustive or specific collection of calibration samples to develop suitable regression models with the strategy here presented.

MATERIALS AND METHODS

Variable Selection Techniques. The large number of available techniques for elimination of useless predictors can be classified into three main categories:

(a) Subset Selection. A number of regression models are built by different subsets of predictors; the performance of the constructed models is then evaluated, and it is used to search for other subsets. The most important example of this class of methods is selection by means of genetic algorithms (GA) coupled with OLS or with PLS regression.

(b) Dimension-wise Selection. Dimension-wise techniques work on a single dimension, in such a way that a biased regression model is progressively built by the addition of individual predictors, as in stepwise ordinary least-squares (SOLS) regression and the SELECT-OLS method, or by the addition of factors (principal components of PCR or latent variables of PLS regression).

(c) Model-wise Elimination. A wide variety of techniques, which can adopt different particular selection strategies and procedures, are part of this group of methods. The regression model can be developed many times with all of the predictors but with only a fraction of the available objects. Then, the useless predictors would be eliminated on the basis of the value and/or the dispersion of their regression coefficient in the regression model. Alternatively, many regression models can be developed with all of the objects but only a fraction of the predictors, so that the useless predictors would be eliminated according to their participation in models with better prediction ability.

To face the calibration problem proposed in the present work, different selection techniques will be applied, which have been chosen on the basis of several considerations, including the availability of related software, the simplicity of their theoretical fundamentals (so that they might be easily understood by chemists without a high chemometric background), and our experience in the field. Thus, methods of subset selection (GA-OLS), dimension-wise selection (SELECT-OLS), and model-wise elimination (IPW, ISE and UVE) have all been used.

This paper does not aim to describe in detail the fundamentals of all variable selection methods that have been tested to develop improved and reliable NIR regression models able to ensure an accurate determination of the caffeine content and roasting color in roasted coffee for quality assurance purposes, moreover, taking into account that some of these methods are quite well-known and have been largely applied in multivariate calibration. A detailed description of the variable selection techniques applied in the present work can be found in their respective references. Nevertheless, considering the recent introduction of SELECT method in its revised form, a careful reading of the original paper (*38*) is particularly recommended to gain a clear understanding of the procedure carried out by SELECT.

Samples. The data set used in the present study originally comprised 83 roasted coffee samples from varied origins and varieties (36 *arabica* and 47 *robusta* coffees), which were processed under different roasting conditions. In addition, 108 blends of *arabica* and *robusta* coffee varieties were prepared in the laboratory by combining the three coffee samples most representative of each variety that had been previously selected. The application of PCA on mean-centered NIR spectra provided an effective and easy-to-implement tool for selecting representative samples from *arabica* and *robusta* varieties. **Figure 1** shows a bidimensional representation of PC1 and PC2 scores accounting for 88.33% of the variance in the roasted coffee NIR spectral data, labeled according to their coffee variety: (1) *arabica* coffees; (2) *robusta* coffees. Two sample groups appeared slightly separated by the first bisectrix of the two component axes, suggesting the presence of two different clusters just associated with the two varieties considered. Thus, the centroid and the two extreme samples within each class (marked with a circle in **Figure 1**) were selected to later generate on their basis suitable coffee blends with a *robusta* content in the final blends ranging from 0 to 60% (w/w).

For each sample, the corresponding reference values of both a chemical parameter closely related to coffee sensory characteristics (caffeine content) and a process-variable fundamental for controlling the roasting degree (roasting color) were obtained. The reference values corresponding to each response analyzed were later used to develop the calibration models proposed in this study.

Each roasted coffee sample was obtained by an individual roasting process in which the amount of coffee to be roasted ranged from 12 to 16 kg (working at six charge levels), whereas the interval of final roasting color, in which all samples were included, presented minimum and maximum values of 48 and 92 (arbitrary units), respectively. Green coffee beans were roasted in a pilot-scale roaster of 20 kg capacity. Roasting temperature and time are closely related to roasting degree. The temperature used in this study ranged from 200 to 230 °C, whereas the length of the roasting time period ranged from 12 to 20 min. These relatively broad temperature and time ranges were chosen in order to be able to obtain different roasting degrees, with light, medium, and dark roasts. Stirring was continued throughout the process to guarantee a uniform heating and roasting of the beans. The roasting degree for each sample was established by determining the color of the beans. At the end of each roasting cycle, when the beans reached the desired color, they were quenched with water to stop the roasting process. The fact that no attempt was made to limit the roasting degree to narrower levels in the data set used was due to the need for including coffee samples covering a wide range of roasting conditions to represent, insofar as possible, the great variability found in the coffee market and to ensure the validity and applicability of the calibration models developed for monitoring roasting color in real time. On the other hand, the other chemical response studied, the caffeine content, ranged from 1.00 to 2.35% (w/w).

Determination of Chemical and Process Variables. The caffeine contents of roasted coffee samples were determined using the corresponding official method proposed by the AOAC (979.11). Nevertheless, the applied analytical method also underwent a laboratory validation study to determine a number of quality parameters such as accuracy, precision (repeatability and reproducibility), specificity, selectivity, linear range, and detection and quantification limits. Thus, a measurement of the accuracy and precision of the validated reference method was obtained by means of the relative standard deviation expressed in percentage equal to 1.25 (% RSD). This precision measurement might be considered to be extremely relevant, because it is important to remember that the prediction error provided by a regression model can never be lower than the experimental error associated with the analytical method used to determine the reference values.

The determination of roasting color was based on the measure of the reflectance diffused at a 0° angle by the surface of a sample under a light at 45° using a spectral colorimeter. Two calibration standards

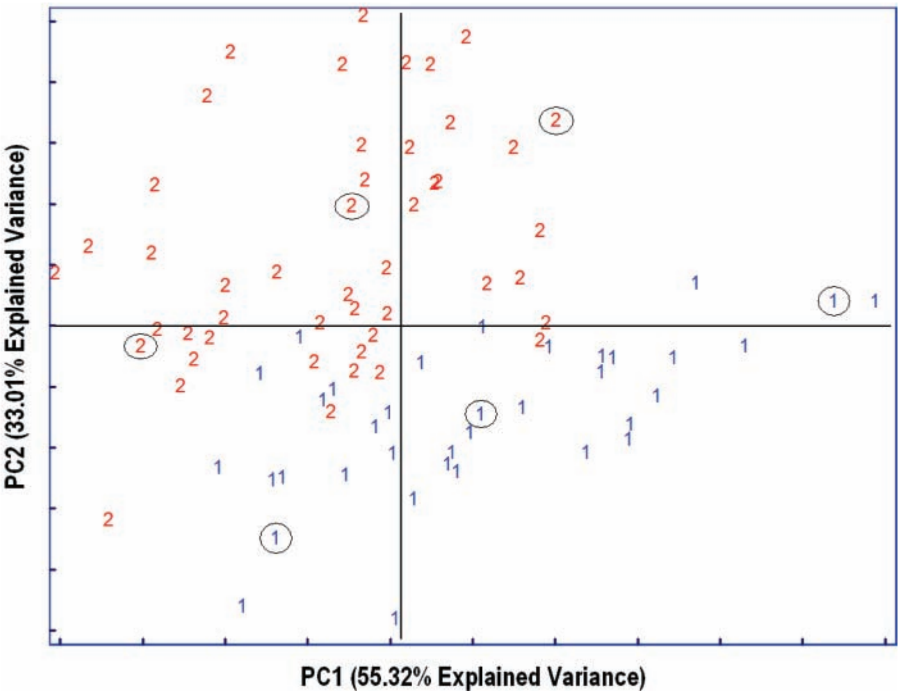

Figure 1. Scores of the 83 roasted coffee samples from *arabica* (labeled 1) and *robusta* (labeled 2) pure varieties on the first two principal components explaining the variability in the NIR spectral data. Objects marked with a circle represent the centroid and the extreme samples within each class.

(corresponding to extreme color values, 25 and 105 ua) were always memorized by the instrument before use in order to obtain a precise measure.

Apparatus and Software. Color measurements were obtained using a Dr. Lange Spectro Color 45°/0° spectral colorimeter (Dr. Bruno Lange GmbH & Co. KG, Düsseldorf, Germany), capable of covering a measuring range from 400 to 700 nm, with a 45°/0° viewing geometry (circular) and a measured area corresponding to 10 mm illuminated/8 mm measured.

The HPLC equipment consisted of an HP 1100 series liquid chromatograph (Hewlett-Packard GmbH, Chemical Analysis Group Europe, Waldbronn, Germany) with a high-pressure gradient pump, vacuum degasser, autosampler, thermostated column compartment, diode array detector, and an HP Chemstation data processing system (Hewlett-Packard) to perform peak purity analyses. The column used was a Zorbax SB-C18, 250 × 4.6 mm i.d. with a 5 μm particle size (Hewlett-Packard GmbH, Waldbronn, Germany). The stable bond packaging was obtained by chemical bonding of a sterically protected octyl stationary phase into a specially prepared high-purity Zorbax Rx porous silica microsphere, suitable for working at low pH values.

NIR spectra were recorded on a near-infrared spectrophotometer NIRSystems 5000 (FOSS, NIRSystems) equipped with a reflectance detector and a sample transport module. The instrument was controlled by a compatible PC, and Vision 2.22 (FOSS, NIRSystems) was used to acquire the data.

Preprocessing of the data, PLS, IPW-PLS, ISE-PLS, UVE-PLS, and GA-OLS calibrations were carried out with the chemometric software V-PARVUS (*39*). Likewise, the updated version of the stepwise orthogonalization of predictors is also contained in the SELECT program of V-PARVUS.

Recording of NIR Spectra. Reflectance spectra were obtained directly from untreated samples. Due care was taken to ensure that the same amount of sample was always used to fill the sample cell. Each spectrum was obtained from 32 scans performed at 2 nm intervals within the wavelength range of 1100−2500 nm, with five replicates for each individual sample. The samples were decompacted between recordings. An average spectrum was subsequently obtained from the replicates collected for each coffee sample.

Data Processing. The wavelength range from 1100 to 2200 nm was selected as the working region. The wavelength range of 2200−2500 nm, in which the signal/noise ratio decreases considerably, was removed in both cases analyzed, after it was verified that the inclusion of this specific wavelength range in the regression model construction would have a harmful influence on their quality. The data matrix containing NIR spectra of roasted coffee samples was subjected to preliminary studies to investigate the presence of possible outlier data that could have a detrimental effect on the quality of the results. None of the diagnostic tools applied (PCA, residual, and leverage plots) revealed the existence of anomalous data.

The whole available data set was randomly split into two independent subsets: a calibration set with 115 samples (used to develop the calibration models) and a test set with 76 samples (never used in the regression models development but to evaluate their actual predictive ability). The main caution taken to select a suitable composition of the external test set was to verify that the contained objects included samples of both pure varieties and different compositional blends and that they uniformly covered the range of values for both responses studied. The test set used was the same for all methods applied and models constructed.

Separate regression models were constructed between calibration NIR spectra and both the roasting color and the caffeine content of roasted coffee samples by the diverse methods employed. First, an initial PLS model was built for each roasted coffee parameter studied using the whole NIR wavelength range of 1100−2200 nm. Then, each variable selection technique evaluated was applied to select a subset of significant NIR absorption bands to be used in the subsequent development of a simplified regression model for predicting each property, in such a way that it would be possible to compare the results obtained with both those resulting from conventional PLS and the other methods tested. All models were developed from the column-wise centered data matrix, regardless of whether an additional preprocessing method was also applied on spectral variables. In an attempt to minimize physical contributions incorporating irrelevant information into spectra, first derivation was also applied to NIR spectral data to test its effect on the quality of the final regression models. To avoid an enhancement of noise when spectra were derived, data were smoothed by using the Savitzky−Golay moving window averaging method, using a cubic degree polynomial with a window size of 13 points. The optimal complexity of every calibration model was assessed by cross-validation (all models were built by cross-validation using five deletion groups).

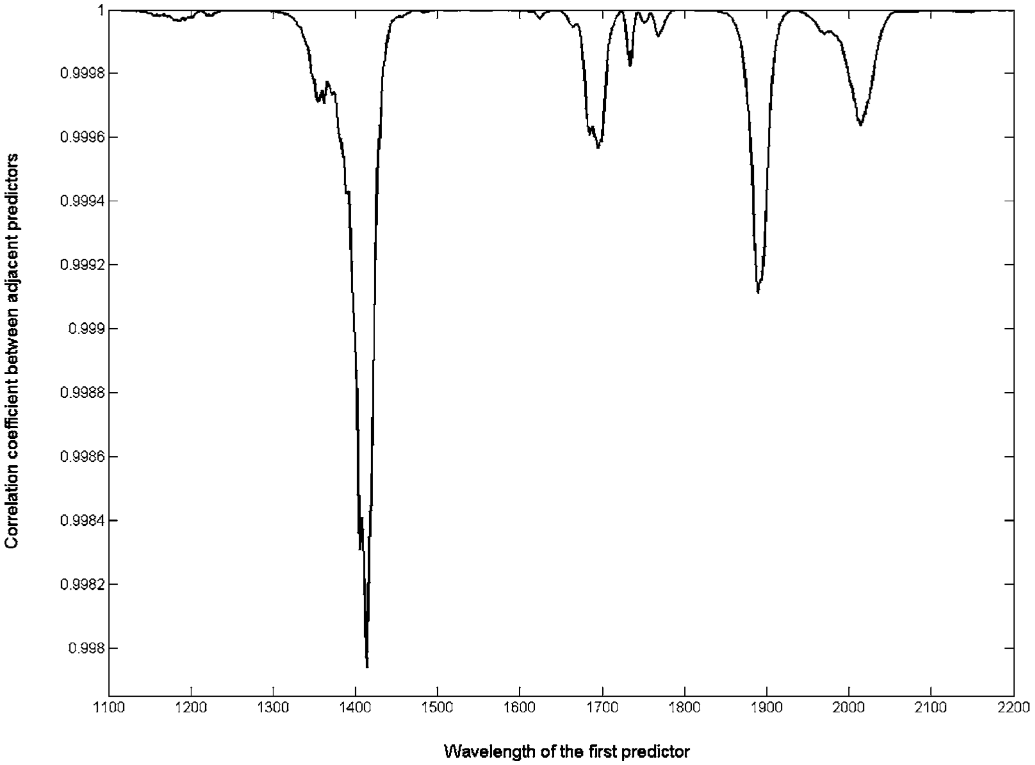

Figure 2. Correlation coefficient between contiguous wavelengths computed from original NIR spectra of roasted coffee samples.

The actual predictive ability of all the constructed regression models was also validated by testing their performance on the external test set, to control and avoid possible over-fitting.

The leave-one-out parameters provided in the case of SELECT-OLS models were obtained from the predictors selected in the final run. On the other hand, other different prediction parameters could be also obtained in the diverse cross-validation cycles used to evaluate the stability of the selection of useful predictors. However, in these evaluation runs the selected predictors were not necessarily the same or the same as the optimal selections obtained in the final run (with all of the objects). Thus, these additional prediction parameters were reported as "Complete-CV" in order to remark that SELECT-OLS seems to be much more penalized than other techniques due to its particular implementation of the cross-validation procedure.

The quality of the results provided by the different regression models constructed was compared according to the root-mean-square error (RMSE) of the residuals obtained, defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \qquad (1)$$

where y_i is the measured response value, $\hat{y}_i$ the computed response value, and n the total number of samples in the set. RMSE is termed RMSEC in calibration, RMSECV in cross-validation, and RMSEP in external prediction. These parameters have the advantage of being dimensionally comparable to the studied response. RMSE expressed as percentage can also be used taking into account the response range in its calculation

$$\%\ \mathrm{RMSE} = \left(\frac{\mathrm{RMSE}}{\max(y_{\mathrm{cal}}) - \min(y_{\mathrm{cal}})}\right) \times 100 \qquad (2)$$

where y_{cal} is the corresponding response value in the calibration set. In this way, results corresponding to dimensionally uncomparable responses may be confronted. For this reason, % RMSE was also used to evaluate the quality of the different regression models constructed.

RESULTS AND DISCUSSION

Correlation Structure of NIR Spectra. From a chemometric point of view, the spectral data have remarkable characteristics, which make necessary their treatment by specific methods. In particular, a common situation in spectroscopy is that there are linear or near-linear relationships among the predictors (which is referred to as a multi-collinearity). Thus, in a NIR spectrum, the variables (absorbances at wavelengths) are highly correlated with each other, with correlation coefficients of >0.999 in large intervals of the spectrum. Among all of the correlations, those between adjacent predictors have special significance, because as contiguous predictors correspond to absorbances at close wavelengths, they usually yield severe collinearities. **Figure 2** might serve to prove this correlation structure for the roasted coffee NIR spectra considered in the present study, as it graphically shows the values of the correlation coefficient between contiguous wavelengths (>0.9979 in all cases).

In this way, in view of the very large correlation coefficients observed, it might be stated that the random independent error on the predictors was very small, taking into account that the square of the correlation coefficient between two adjacent predictors represented a measure of the variance of the error of the regression between them, and that this variance could never be less than the variance of the independent random error contained in the data. The standard deviation of the regression error between adjacent predictors, shown in **Figure 3**, has clearly a structure, with maxima in correspondence with the maxima of the derivative of the mean spectrum. A similar structure is always obtained with NIR data (*37*, *38*). Random error cannot show a structure, so that it must have a standard deviation of clearly <0.0001, the range of the irregularities in **Figure 3**.

On the other hand, when a random Gaussian noise $N(0, \sigma)$ was added to the original spectra of roasted coffee samples, with the standard deviation of the normal added noise σ in a rather large interval, it could be observed how the correlation structure of the data set was destroyed (**Figure 4a**). In terms of PLS regression performance, the cross-validation error of the PLS models constructed for both modeled responses remained about constant when a moderate amount of random independent

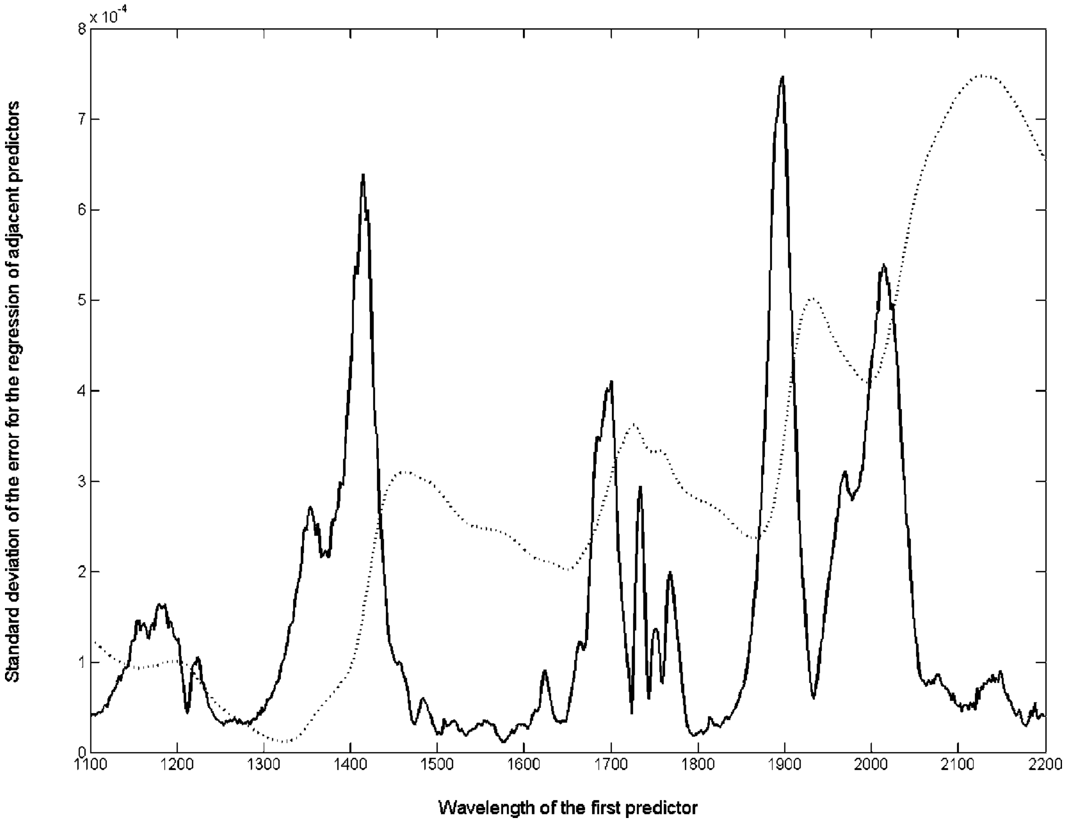

Figure 3. Standard deviation of the residuals for the regression of each wavelength on the contiguous wavelength.

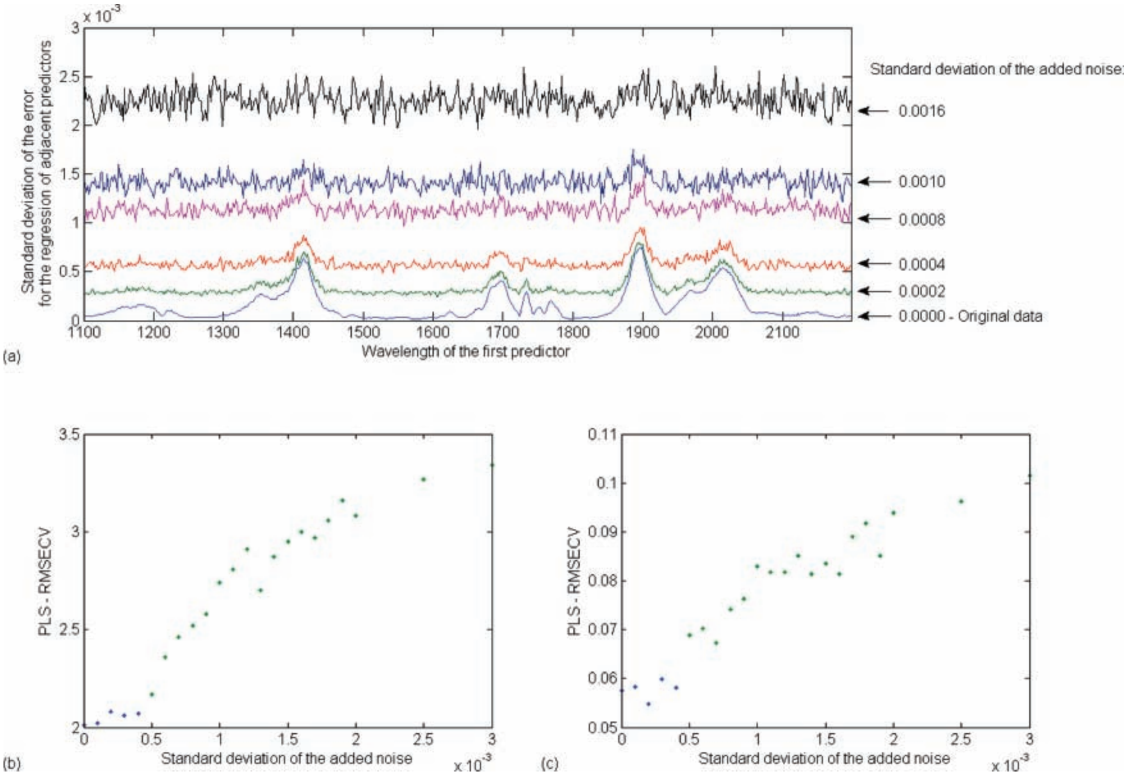

Figure 4. Random Gaussian noise *N*(0, σ) added to original NIR spectra of roasted coffee samples: (**a**) standard deviation of the residuals for the regression of each wavelength on the contiguous wavelength as a function of the standard deviation of the added noise; (**b**) cross-validation errors (RMSECV) obtained with PLS regression as a function of the σ of the added noise using the roasting color as response variable; (**c**) cross-validation errors (RMSECV) obtained with PLS regression as a function of the σ of the added noise using the caffeine content as response variable.

noise was added to spectra (σ < 0.0005) (**Figure 4b**,c). With increased noise, the prediction error increased for both response variables, but not as drastically as might be expected (taking into account that the added noise was much greater than the

Table 1. Calibration (RMSEC), Cross-Validation (RMSECV), and External Prediction (RMSEP) Errors, Percentages of Explained Variance, and Complexities Corresponding to the Regression Models Obtained with Each Variable Selection Technique Applied To Model the Roasting Color of Roasted Coffee Samples, Using Five CV Groups

method	pretreatment	predictors	complexity[a]	% expl var (cal)	RMSEC (% RMSEC)	% expl var (CV)[b]	RMSECV (% RMSECV)	RMSEP (% RMSEP)
PLS	first derivative	551	10 LVs	97.22	1.44 (3.27)	96.63	1.58 (3.59)	1.68 (3.82)
IPW-PLS[c]	first derivative	23	9 LVs	96.82	1.54 (3.50)	96.10	1.70 (3.86)	1.72 (3.91)
ISE-PLS	first derivative	133	10 LVs	97.89	1.25 (2.84)	97.32	1.41 (3.20)	1.62 (3.68)
UVE-PLS[d]	first derivative	300	10 LVs	97.54	1.35 (3.07)	97.00	1.49 (3.39)	1.66 (3.77)
GA-OLS[e]	first derivative	7	7 λs	97.42	1.38 (3.14)	96.97	1.50 (3.41)	1.74 (3.95)
SELECT-OLS[f]	first derivative	9	9 λs	97.21	1.44 (3.27)	96.72	1.57 (3.57)	1.62 (3.68)

[a] Minimum RMSECV for PLS-based techniques. [b] LOO-CV for SELECT-OLS model. [c] Ten IPW cycles, exponent of the importance to obtain the predictor weights = 0.7, cutoff level = 0.001. [d] Normal UVE, 300 noisy predictors. [e] Probability of mutation = 0.005, 2% Elitism, 1 GA run, 50 cycles. [f] Complete-CV: 93.63% CV-explained variance, 2.17 RMSECV.

Table 2. Calibration (RMSEC), Cross-Validation (RMSECV), and External Prediction (RMSEP) Errors, Percentages of Explained Variance, and Complexities Corresponding to the Regression Models Obtained with Each Variable Selection Technique Applied To Model the Caffeine Content of Roasted Coffee Samples, Using Five CV Groups

method	pretreatment	predictors	complexity[a]	% expl var (cal)	RMSEC (% RMSEC)	% expl var (CV)[b]	RMSECV (% RMSECV)	RMSEP (% RMSEP)
PLS	first derivative	551	10 LVs	99.63	0.0266 (1.99)	99.42	0.0333 (2.49)	0.0320 (2.39)
IPW-PLS[c]	first derivative	27	9 LVs	99.62	0.0269 (2.01)	99.56	0.0290 (2.17)	0.0377 (2.82)
ISE-PLS	first derivative	199	10 LVs	99.77	0.0209 (1.56)	99.69	0.0245 (1.83)	0.0269 (2.01)
UVE-PLS[d]	first derivative	270	10 LVs	99.69	0.0244 (1.83)	99.52	0.0304 (2.27)	0.0301 (2.25)
GA-OLS[e]	first derivative	7	7 λs	99.36	0.0352 (2.63)	99.28	0.0372 (2.78)	0.0403 (3.02)
SELECT-OLS[f]	mean centering	17	17 λs	99.87	0.0157 (1.17)	99.80	0.0198 (1.48)	0.0195 (1.46)

[a] Minimum RMSECV for PLS-based techniques. [b] LOO-CV for SELECT-OLS model. [c] Ten IPW cycles, exponent of the importance to obtain the predictor weights = 0.7, cutoff level = 0.001. [d] Normal UVE, 100 noisy predictors. [e] Probability of mutation = 0.005, 2% Elitism, 1 GA run, 50 cycles. [f] Complete-CV: 99.44% CV-explained variance, 0.0330 RMSECV.

actual noise in the data set), due to the partial compensation of the random errors prompted by the synergism of the predictors in the PLS latent variables (reduced sensitivity to noise).

In light of these facts, it was possible to reach the following conclusions:

(1) The random independent error of the NIR spectra of roasted coffee samples was very small.

(2) The prediction error on both response variables was controlled by the error of the corresponding reference technique and/or by systematic errors in the spectra.

Therefore, as the hypothesis ("the error on the predictors is randomly independent") that frequently suggests the use in multivariate calibration of techniques that retain all of the available information in the block of predictors to take advantage of the synergism of highly correlated variables to decrease the variance of the error (such as PLS regression) was not verified, the application of techniques that select a minimum number (maximum parsimony) of useful predictors seemed to be suitable. In fact, the parsimony principle precisely states that "entities should not be multiplied needlessly", which applied to multivariate calibration would mean that the simplest of two competing models is to be preferred.

Regression Models for Prediction of Roasting Color and Caffeine Content. The results obtained in both calibration and prediction after application of the diverse variable selection techniques tested to model both the roasting color and the caffeine content of roasted coffee samples, after selecting the most suitable data pretreatment and model complexity, are summarized in **Tables 1** and **2**, respectively. These results were also compared to those provided by the PLS model based on the full spectral range to better evaluate their actual quality. Despite the different natures of the two response variables considered in the present study, the conclusions that could be drawn were rather similar in relation to the comparative performance of the methods applied.

The application of a spectral pretreatment such as first derivation showed relative success in correcting, at least partially, the systematic variation arising from scattering or baseline shifts, because, for almost all calibration methods evaluated, its previous use provided improved results in comparison with direct development of regression models on absorbance spectra (only mean-centered).

When PLS was applied to the first-derivative spectra of roasted coffee samples for predicting both their roasting color and caffeine content, a high number of PLS components (10 LVs) was needed to develop calibration models capable of tackling the interference and redundancy of NIR signals with a suitable predictive accuracy.

In broad strokes, for both response variables studied, the applied selection techniques seemed to be separated into conservative (ISE, UVE) and parsimonious (IPW, GA-OLS, SELECT-OLS) methods.

The prediction performances of both conservative methods were quite similar or slightly improved compared to PLS regression developed from all of the predictors, but the complexities associated with the final PLS models could not be simplified. Thus, the still high number of predictors retained as useful made it necessary to include a very large number of latent variables in model construction to achieve a good prediction of both parameters, because information about background and noise was still present in selected predictors.

When the roasting color was considered as a roasted coffee quality parameter to be modeled, the application of IPW provided a high-quality calibration model, exhibiting a notable predictive ability, comparable to that achieved with conservative techniques and usual PLS regression, but reducing considerably the number of significant wavelengths taken into account in the model development and decreasing the observed overfitting degree. However, although the use of IPW-PLS to model the caffeine content also fostered a very substantial reduction in the number of useful predictors (only 27 wavelengths were considered in the PLS model development), this simplification in model complexity was not accompanied by an improvement in model performance, leading to even worse results in terms of predictive ability compared with the full-spectrum approach, also increasing the overfitting degree, probably due to a high residual noise level in the spectral bands selected by IPW for predicting this particular response.

In view of the low calibration and cross-validation errors and the extremely reduced complexity associated with the OLS model based on the wavelengths selected as significant by GA when the roasting color was considered as the response of interest, this highly parsimonious model might seem an improved solution with respect to the original PLS model. Nevertheless, the results obtained from the use of the external test set (with objects never used in the selection of predictors) proved that the GA-OLS method actually overestimated the prediction ability for predicting the roasting color, in such a way that this external performance measurement should be considered to be a more reliable estimate of its efficiency. Likewise, GA-OLS showed a certain failure in extracting individual predictors representing the underlying features responsible for a successful prediction of the caffeine content of roasted coffee samples, because the OLS regression model developed from a reduced subset of only seven variables selected as optimal solution by GA yielded the poorest results, in both calibration and prediction, among all of the applied methods.

Furthermore, both the flexibility and the computing time requirements of each method are additional factors to be considered in selecting the most suitable approach to be applied. The computing time is very large for GA-based techniques and for ISE (when only the worst predictor is eliminated in each cycle). On the contrary, SELECT-OLS (establishing a reasonable limit on the maximum number of selectable predictors allowed) and IPW (generally, 10−12 cycles would be enough) can be regarded as the methods that require less computation time.

All in all, and after examination of the resultant errors in both calibration and prediction, as well as the corresponding complexities for all models evaluated, it could be claimed that the best regression models for both analyzed responses were achieved when the stepwise orthogonalization of predictors (SELECT) was applied on NIR spectra to generate a reduced set of decorrelated wavelengths on the basis of their correlation coefficients with the response considered, so that only relevant (but not redundant) information was taken into account in each final OLS model development. In this way, the SELECT-OLS model constructed for assessing the roasting color resulted in a high predictive ability (3.68% RMSEP) with very low model complexity (only nine predictors deflated from redundant information) and with no alarming signs of overfitting (3.27% RMSEC). Likewise, when the SELECT method was applied on raw data to model and predict the caffeine content of roasted coffees, the optimal OLS model obtained from selected orthogonal variables prompted notably improved calibration and prediction results (1.17% RMSEC and 1.46% RMSEP) not only with respect to the PLS model constructed from first-derivative whole spectra but also with regard to the models developed by other selection methods, reducing the final complexity to 17 significant wavelengths. The particular influence of each selected band on the SELECT-OLS model finally built for predicting each quality parameter of roasted coffees was given by their respective correlation weights (**Tables 3** and **4**). The optimal complexities for both high-quality SELECT-OLS models proposed here were properly selected according to the complete validation strategy implemented in the recently updated version of the method (**Figure 5**). The larger validation errors obtained in both cases in the complete CV procedure performed by SELECT-OLS indicated that the selection of the useful predictors in the cross-validation cycles was different from the final selection run with all of the calibration objects in the training set.

Table 3. Order of Selection of the Significant/Decorrelated Wavelengths and the Respective Correlation Coefficients with the Response Corresponding to the SELECT-OLS Regression Model Developed from First-Derivative NIR Spectra To Model and Predict the Roasting Color of Roasted Coffees

order of selection	predictor index	wavelength (nm)	correlation
1	510	2118	0.865
2	277	1652	0.254
3	188	1474	0.249
4	59	1216	0.170
5	343	1784	0.138
6	16	1130	0.062
7	107	1312	0.148
8	323	1744	0.135
9	143	1384	0.092

Table 4. Order of Selection of the Significant/Decorrelated Wavelengths and the Respective Correlation Coefficient with the Response Corresponding to the SELECT-OLS Regression Model Developed from Original NIR Spectra To Model and Predict the Caffeine Content of Roasted Coffees

order of selection	predictor index	wavelength (nm)	correlation
1	131	1360	0.8298
2	56	1210	0.1354
3	149	1396	0.4472
4	461	2020	0.2442
5	393	1884	0.0582
6	59	1216	0.0489
7	160	1418	0.0527
8	1	1100	0.0312
9	285	1668	0.0789
10	294	1686	0.0896
11	286	1670	0.0348
12	348	1794	0.0602
13	289	1676	0.0404
14	443	1984	0.0359
15	297	1692	0.0137
16	282	1662	0.0184
17	45	1188	0.0166

The great compression rates resulting from the application of the optimization procedure carried out by SELECT (only 9 and 17 from a total of 551 predictors were selected to model roasting color and caffeine content, respectively), together with the high reliability and robustness of the subsequently developed OLS models, demonstrated the great effectiveness of SELECT-OLS as a correction and compression (feature selection) method. Moreover, the comparison and progressive reduction in the residual variance, before selection of any variable and after each successive decorrelation cycle was completed (until the optimal model complexity was reached in each case and the residual variance was negligible), can be considered as additional proof of the method efficiency (**Figure 6**).

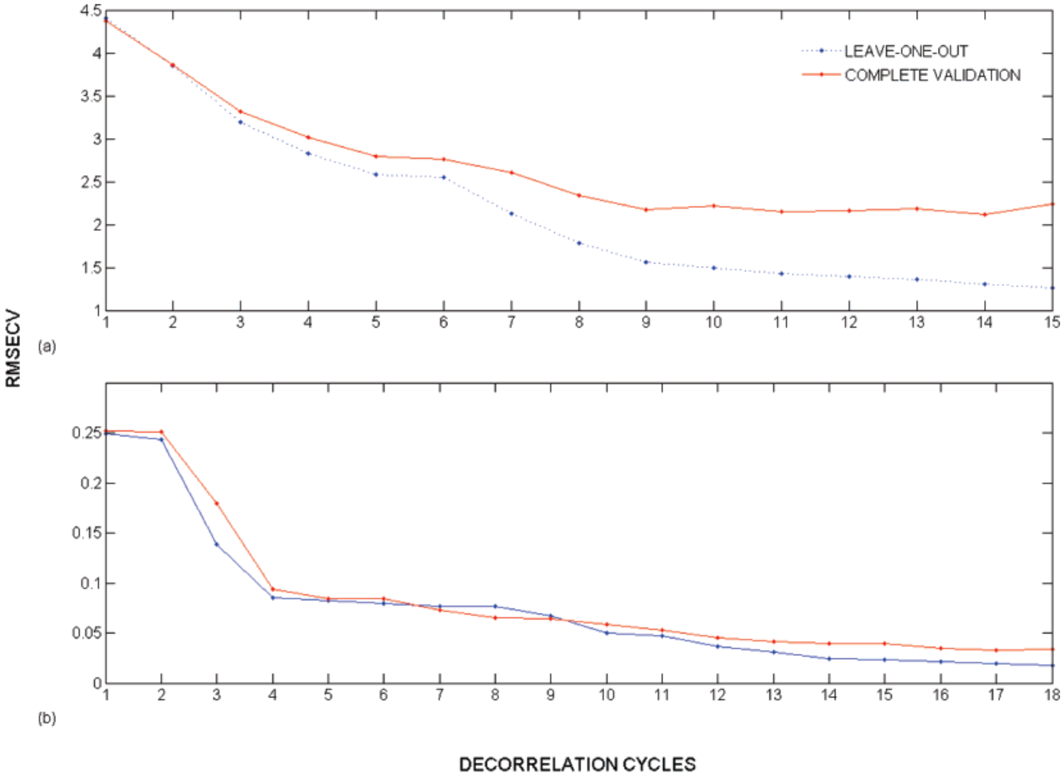

Figure 5. Leave-one-out and complete validation predictions provided by SELECT as a function of the number of selected predictors: (**a**) results from the model based on mean-centered first-derivative spectra using the roasting color as response variable; (**b**) results from the model based on mean-centered spectra using the caffeine content as response variable.

Stepwise Orthogonalization of Predictors: Chemical Assignments of Selected Near-Infrared Absorption Bands. The variable subset selection carried out by SELECT-OLS was based on statistical criteria, that is, neither spectroscopic nor chemical considerations were taken into account to constitute the final subset of decorrelated wavelengths considered as relevant for modeling each roasted coffee parameter. However, apart from confirming the suitability of the OLS regression models proposed in the present study, it was possible to determine the chemical assignments of the NIR bands previously selected by SELECT for predicting the two quality attributes analyzed (*9*). The following discussion is not intended to be an exhaustive compilation of all the potential wavelength regions and the related compounds that could contribute to each response studied, but only a chemical reasoning focused on the final variable selection provided by SELECT-OLS.

After application of the stepwise decorrelation of predictors, working on first-derivative NIR spectra, nine wavelengths were selected as useful for developing a subsequent reliable and parsimonious OLS model for determining the roasting color of roasted coffee. **Table 5** provides an overview of all these relevant absorption bands, including both the identification of the bond vibrations involved and the respective chemical assignments. The major degradation product during roasting is water (70%), in such a way that the proportion of moisture released is proportional to the degree of roasting. On the other hand, quantitatively, the major acids in coffee are chlorogenic acids (CGA). Although the contribution of these acids to perceived acidity is secondary, they are largely degraded during the roasting process, mainly into quinic acid, increasing their decomposition jointly with roasting degree. Thus, the selection of certain bands assignable to both water and chlorogenic acid structures can be easily explained by considering their common use as practical indices of actual roasting degree, as in the case of the roasting color. Likewise, the considerable importance attached to various selected wavelengths, which can be assigned to typical vibrations of other coffee constituents, such as carbohydrates and amino acids, can be consistently justified by bearing in mind that the organic losses of these compounds are also strongly affected by the degree of roasting. It deserves to be particularly emphasized that thanks precisely to the close connection between the roasting color (directly related to roasting degree) and the particular chemical composition of roasted coffee samples (strongly dependent on the chemical changes occurring at roasting) it was possible to develop, in the present work, reliable calibration models based on NIR spectra (which do not include the visible region) for assessing an optical property (such as roasting color).

Finally, the 17 λs considered to be important for the prediction of roasted coffee caffeine content as selected by SELECT method on the basis of mean-centered NIR spectra (after stepwise orthogonalization of predictors), as well as their corresponding assignments to characteristic absorptions of caffeine structure, are shown in **Table 6**.

In summary, many variable (wavelength) selection methods widely used in NIR calibration are mainly focused on “noise” removal by selection of a reduced number of significant predictors showing a high correlation with the response to be modeled, so as to enhance model predictive ability. However, many of these techniques may frequently ignore the fact that the optimal wavelength combination finally selected for modeling could still contain redundant information (collinearity between useful predictors). Instead, the SELECT method applied in this study takes into account this argument (which can be particularly problematic in the development of calibration models from individual predictors) and, probably, herein may lie the reason for its potential advantages over other traditional variable selection approaches. The stepwise orthogonalization

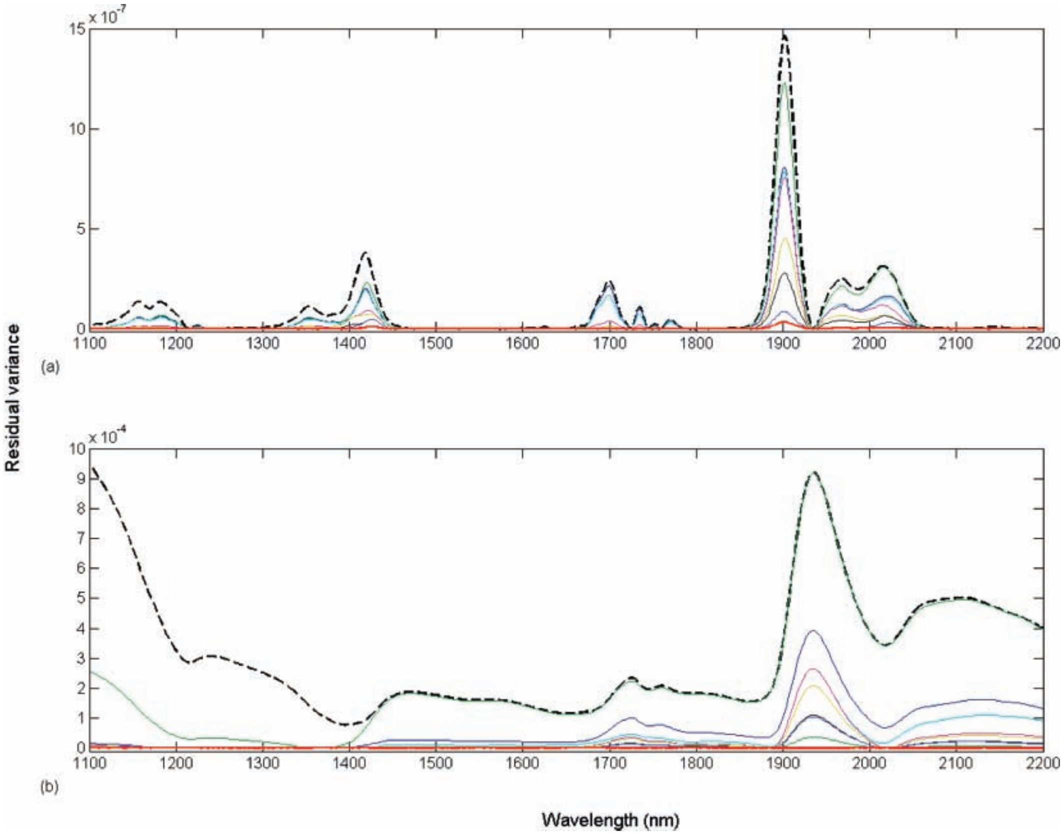

Figure 6. (**a**) Variance of the original variables (before selection of any wavelength) (rough dotted line) and residual variance of the decorrelated variables after 1–8 (fine solid lines) and 9 (rough solid line) selections obtained when working on first-derivative spectra, using the roasting color as response variable. (**b**) Variance of the original variables (rough dotted line), and residual variance of the decorrelated variables after 1–8 (fine solid lines) and 17 (rough solid line) selections obtained when working on original spectra, using the caffeine content as response variable.

Table 5. Chemical Assignments of Significant Near-Infrared Bands Selected by SELECT-OLS for Modeling Roasting Color of Roasted Coffees

selected wavelength (nm)	bond vibration	assignment
1130	C—H str second overtone (Ar)	CGA
1216	C—H str second overtone (CH_2)	carbohydrates amino acids
1312	2 × C—H str + C—H def (CH_3)	carbohydrates amino acids CGA
1384	2 × C—H str + C—H def (CH_2)	carbohydrates amino acids
1474	O—H str first overtone	H_2O
1652	C—H str first overtone (Ar)	CGA
1744	C—H str first overtone (CH_2 asym)	carbohydrates amino acids
1784	C—H str first overtone (CH_2 sym)	carbohydrates amino acids
2118	N—H str + C═O str O—H str + O—H def	amino acids CGA

Table 6. Chemical Assignments of Significant Near-Infrared Bands Selected by SELECT-OLS for Modeling Caffeine Content of Roasted Coffees

selected wavelength (nm)	bond vibration
1100	C—H str second overtone (═CH)
1188	C—H str second overtone (–CH_3 asym)
1210	C—H str second overtone (–CH_3 sym)
1216	C—H str second overtone (–CH)
1360	2 × C—H str + 2 × C—H def (–CH_3)
1396	2 × C—H str + 2 × C—H def (═CH)
1418	2 × C—H str + 2 × C—H def (–CH)
1662/1668/1670/1676	C—H str first overtone (═CH)
1686/1692	C—H str first overtone (–CH_3 asym)
1794	C—H str first overtone (–CH_3 sym)
1884	C═O str second overtone
1984	C═O str second overtone
2020	amide II + amide III

of predictors goes beyond a simple variable selection based on the predictors correlation with a variable response, because it simultaneously carries out two essential needs in multivariate calibration—data correction and feature selection (data compression)—thanks to the combined application in the same method of an orthogonalization algorithm to decorrelate the predictor variables and a "forward" stepwise variable selection procedure. Thus, the SELECT-OLS method is aimed at not only selecting a minimum number (maximum parsimony) of informative predictors closely related to the considered response but also avoiding any redundancy on the subset of relevant variables used to construct the final regression model. Although, in principle, it could be thought that a parsimonious technique such as SELECT loses the benefit of the potential synergism of correlated predictors, this loss is not significant when the main source of error is the determination of the response with a reference method (once the systematic variation in the spectra is minimized).

Thus, in this paper, near-infrared spectroscopy, combined with the above-mentioned powerful wavelength selection method (stepwise orthogonalization of predictors) and with OLS regression, has been shown to provide a very suitable strategy for modeling and predicting two roasted coffee attributes (roasting color and caffeine content) of great relevance from a quality assurance point of view. The use in industrial scale of the simple and reliable calibration models proposed in the present study (developed from a minimum number of significant and uncorrelated predictors) has the potential of dramatically reducing analytical time, efforts, and costs of assessing these roasted coffee quality parameters, allowing a near real time determination to be used for online monitoring of the coffee roasting process and for controlling coffee quality and possible quality changes.

The promising results obtained will allow a similar methodology to be considered in future applications related to online quantification of further coffee quality parameters.

LITERATURE CITED

(1) Clarke, R. J.; Macrae, R. *Coffee, Vol. 1: Chemistry*; Elsevier Applied Science Publishers: London, U.K., 1985.

(2) Dalla Rosa, M.; Nicoli, M. C.; Lerici, C. R. Caratteristiche qualitative del cafè espresso in relazione alle modalitá di preparazione. *Ind. Aliment.* **1986**, *25*, 629−633.

(3) Clarke, R. J.; Macrae, R. *Coffee, Vol. 2: Technology*; Elsevier Applied Science Publishers: London, U.K., 1987.

(4) Illy, A.; Viani, R. *Espresso Coffee: The Chemistry of Quality*; Academic Press: London, U.K., 1995.

(5) Nunes, F. M.; Coimbra, M. A.; Duarte, A. C.; Delgadillo, I. Foamability, foam stability, and chemical composition of espresso coffee as affected by the degree of roast. *J. Agric. Food Chem.* **1997**, *45*, 3238−3243.

(6) Andueza, S.; Maeztu, L.; Dean, B.; de Peña, M. P.; Bello, J.; Cid, C. Influence of water pressure on the final quality of arabica espresso coffee. Application of multivariate analysis. *J. Agric. Food Chem.* **2002**, *50*, 7426−7431.

(7) Franca, A. S.; Mendonça, J. C. F.; Oliveira, S. D. Composition of green and roasted coffees of different cup qualities. *LWT Food Sci. Technol.* **2005**, *38*, 709−715.

(8) Hildrum, K. I.; Isaksson, T.; Næs, T.; Tandberg, A. *Near-Infrared Spectroscopy: Bridging the Gap between Data Analysis and NIR Applications*; Ellis Horwood: Chichester, U.K., 1992.

(9) Osborne, B. G.; Fearn, T.; Hindle, P. H. *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*, 2nd ed.; Longman Scientific and Technical: Harlow, U.K., 1993.

(10) Williams, P.; Norris, K. *Near-Infrared Technology in the Agricultural and Food Industries*, 2nd ed.; American Association of Cereal Chemists: St. Paul, MN, 2001.

(11) Burns, D. A.; Ciurczak, E. K. *Handbook of Near-Infrared Analysis*, 2nd ed.; Dekker: New York, 2001.

(12) Workman, J. A review of process near infrared spectroscopy: 1980−1994. *J. Near Infrared Spectrosc.* **1993**, *1*, 221−245.

(13) Downey, G. Authentication of food and food ingredients by near infrared spectroscopy. *J. Near Infrared Spectrosc*. **1996**, *4*, 47−61.

(14) Kurowski, C.; Timm, D.; Grummisch, U.; Meyhack, U.; Grunewald, H. The benefits of near infrared analysis for food product quality. *J. Near Infrared Spectrosc*. **1998**, *6*, 343−348.

(15) Sahni, N. S.; Isaksson, T.; Næs, T. In-line near infrared spectroscopy for use in product and process monitoring in the food industry. *J. Near Infrared Spectrosc*. **2004**, *12*, 77−83.

(16) Downey, G.; Boussion, J.; Beauchêne, D. Authentication of whole and ground coffee beans by near infrared reflectance spectroscopy. *J. Near Infrared Spectrosc*. **1994**, *2*, 85−92.

(17) Downey, G.; Boussion, J. Authentication of coffee bean variety by near-infrared reflectance spectroscopy of dried extract. *J. Sci. Food Agric*. **1996**, *71*, 41−49.

(18) Downey, G.; Spengler, B. Compositional analysis of coffee blends by near infrared spectroscopy. *Ir. J. Agric. Food Res*. **1996**, *35*, 179−188.

(19) Downey, G.; Briandet, R.; Wilson, R. H.; Kemsley, E. K. Near- and mid-infrared spectroscopies in food authentication: coffee varietal identification. *J. Agric. Food Chem.* **1997**, *45*, 4357−4361.

(20) Esteban-Díez, I.; González-Sáiz, J. M.; Pizarro, C. An evaluation of orthogonal signal correction methods for the characterisation of arabica and robusta coffee varieties by NIRS. *Anal. Chim. Acta* **2004**, *514*, 57−67.

(21) Esteban-Díez, I.; González-Sáiz, J. M.; Sáenz-González, C.; Pizarro, C. Coffee varietal differentiation based on near-infrared spectroscopy. *Talanta* **2007**, *71*, 221−229.

(22) Pizarro, C.; Esteban-Díez, I.; Nistal, A. J.; González-Sáiz, J. M. Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Anal. Chim. Acta* **2004**, *509*, 217−227.

(23) Esteban-Díez, I.; González-Sáiz, J. M.; Pizarro, C. Prediction of roasting colour and other quality parameters of roasted coffee samples by near infrared spectroscopy. A feasibility study. *J. Near Infrared Spectrosc*. **2004**, *12*, 287−298.

(24) Esteban-Díez, I.; González-Sáiz, J. M.; Pizarro, C. Prediction of sensory properties of espresso from roasted coffee samples by near infrared spectroscopy. *Anal. Chim. Acta* **2004**, *525*, 171−182.

(25) Huck, C. W.; Guggenbichler, W.; Bonn, G. K. Analysis of caffeine, theobromine and theophylline in coffee by near infrared spectroscopy (NIRS) compared to high-performance liquid chromatography (HPLC) coupled to mass spectrometry. *Anal. Chim. Acta* **2005**, *538*, 195−203.

(26) Swierenga, H.; Wülfert, F.; de Noord, O. E.; de Weijer, A. P.; Smilde, A. K.; Buydens, L. M. C. Development of robust calibration models in near-infrared spectrometric applications. *Anal. Chim. Acta* **2000**, *411*, 121−135.

(27) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 131−150.

(28) Zeaiter, M.; Roger, J. M.; Bellon-Maurel, V. Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods. *Trends Anal. Chem.* **2005**, *24*, 437−445.

(29) Geladi, P.; Kowalski, B. R. Partial least squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(30) Forina, M.; Lanteri, S.; Cerrato-Oliveros, M. C.; Pizarro-Millán, C. Selection of useful predictors in multivariate calibration. *Anal. Bioanal. Chem*. **2004**, *380*, 397−418.

(31) Nadler, B.; Coifman, R. R. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration. *J. Chemom.* **2005**, *19*, 107−118.

(32) Forina, M.; Casolino, C.; Pizarro, C. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *J. Chemom.* **1999**, *13*, 165−184.

(33) Boggia, R.; Forina, M.; Fossa, P.; Mosti, L. Chemometrics study and validation strategies in the structure-activity relationships of a new class of cardiotonic agents. *Quant. Struct.−Act. Rel.* **1997**, *16*, 201−206.

(34) Centner, V.; Massart, D. L.; de Noord, O. E.; De Jong, S.; Vandeginste, B. M.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem*. **1996**, *68*, 3851−3858.

(35) Lucasius, C. B.; Kateman, G. Genetic algorithms for large-scale optimization in chemometrics: an application. *Trends Anal. Chem*. **1991**, *10*, 254−261.

(36) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267−282.

(37) Cerrato-Oliveros, M. C.; Forina, M.; Casale, M.; Pizarro, C. *Stepwise Orthogonalization of Predictors in NIR Spectroscopy*; Euroanalysis XIII (European Conference on Analytical Chemistry): Salamanca, Spain, 2004.

(38) Forina, M.; Lanteri, S.; Casale, M.; Cerrato-Oliveros, M. C. Stepwise orthogonalization of predictors in classification and regression techniques: an "old" technique revisited. *Chemom. Intell. Lab. Syst*. **2007**, *87*, 252−261.

(39) Forina, M.; Lanteri, S.; Armanino, C.; Cerrato-Oliveros, M. C.; Casolino, C. *V-PARVUS 2004, an Extendable Package of Programs for Explorative Data Analysis, Classification and Regression Análisis*; Dipartimento di Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova: Genova, Italy, 2004.

Received for review April 19, 2007. Revised manuscript received June 28, 2007. Accepted July 2, 2007. We thank the Ministry of Science and Technology (Project 2FD1997-0491) for financial support.

JF071139X