Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 509 (2004) 217–227
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy

Consuelo Pizarro [a,*], Isabel Esteban-Díez [a], Antonio-José Nistal [b], José-María González-Sáiz [a]

[a] *Department of Chemistry, University of La Rioja, C/Madre de Dios 51, 26006 Logroño (La Rioja), Spain*
[b] *Laboratorio del Ebro, Centro Técnico Nacional de Conservas Vegetales, C/Santa Gema 56, Apartado 21, 31570 San Adrián (Navarra), Spain*

Received 4 April 2003; received in revised form 31 October 2003; accepted 7 November 2003

Abstract

In near-infrared (NIR) measurements, some physical features of the sample can be responsible for effects like light scattering, which lead to systematic variations unrelated to the studied responses. These errors can disturb the robustness and reliability of multivariate calibration models. Several mathematical treatments are usually applied to remove systematic noise in data, being the most common derivation, standard normal variate (SNV) and multiplicative scatter correction (MSC). New mathematical treatments, such as orthogonal signal correction (OSC) and direct orthogonal signal correction (DOSC), have been developed to minimize the variability unrelated to the response in spectral data. In this work, these two new pre-processing methods were applied to a set of roasted coffee NIR spectra. A separate calibration model was developed to quantify the ash content and lipids in roasted coffee samples by PLS regression. The results provided by these correction methods were compared to those obtained with the original data and the data corrected by derivation, SNV and MSC. For both responses, OSC and DOSC treatments gave PLS calibration models with improved prediction abilities (4.9 and 3.3% RMSEP with corrected data versus 7.1 and 8.3% RMSEP with original data, respectively).

Keywords: Roasted coffee; Near-infrared spectroscopy; Multivariate calibration; Spectral pre-processing; Orthogonal signal correction

1. Introduction

The total solids content, that is, the overall concentration of a beverage (including suspended solid materials and emulsified lipids and solutes) is the most important feature of the chemical composition of espresso and roasted coffee. On the other hand, three of the main organoleptic features of a cup of espresso are flavour, mouthfeel and aftertaste, and they all depend on the quantity of oil present as a dispersion of small droplets that not only dissolve important flavour components, but also increase the viscosity of the beverage [1]. Therefore, it is quite important to ensure an accurate determination of the ash content and total lipids in roasted coffee, particularly in terms of quality assurance.

To determine the ash content in roasted coffee, the coffee industry uses an analytical reference method that requires carbonising the sample until constant weight is achieved, whereas the reference method for the quantification of the total lipids involves a Soxlhet extraction and desiccation until constant weight is attained [2]. These reference methods are perfectly reliable but, in practice, they are also time-consuming methods that use organic solvents, and this is a clear disadvantage from an environmental point of view.

The use of near-infrared spectroscopy (NIRS) to estimate the different features of the samples has become widespread thanks to the progress of multivariate calibration [3]. Among other advantages, in NIR spectroscopy the measurements are non-destructive and it is possible to use solid samples without any sample pre-treatment. Typically, with NIR spectra, the analytical information is contained in fine spectral variations that are usually dominated by features such as light scattering, background noise and baseline drift. These unwanted variations can have adverse effects on the development of a calibration model and produce inaccurate or biased results. Therefore, when working with NIR data, an important decision is whether a pre-processing method is necessary, and if this is the case, the selection of a suitable pre-processing method is an important step in the model building. For example, NIR spectra are subject to large

* Corresponding author. Tel.: +34-941-299626; fax: +34-941-299621.
E-mail address: consuelo.pizarro@dq.unirioja.es (C. Pizarro).

0003-2670/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2003.11.008

baseline shifts due to the reflectance mode in which they are usually recorded [4,5]. This problem is particularly relevant for solid powdered samples, such as roasted coffee, which contain materials of varying particle size distributions. The use of adequate pre-processing methods minimizes the contribution of physical effects to the NIR spectra.

The mathematical treatments most commonly used to correct the spectral variations due to physical changes are derivation to different orders [6,7], standard normal variate (SNV) and multiplicative scatter correction (MSC) [8–10]. The first derivative is commonly used to eliminate baseline offset variations within a set of spectra. As it is a constant (zero order) term added to a function $f(w)$ (the spectrum), offset C is eliminated by taking the derivative with respect to w (wavelength). On the other hand, the linear baseline of a spectrum is described by the first order equation $aw + C$ (a is slope, w the wavelength, and C the offset), which adds to a function $f(w)$ (the spectrum). As it was shown earlier, the calculation of the first derivative with respect to w eliminates the offset term. However, the slope term becomes a constant term in the first derivative. It is a common practice, therefore, to take the second derivative with respect to w so as to eliminate both the offset and slope. The spectral offset and slope may vary within a set of spectra for several reasons including, particle size differences among samples, varying particulate levels among liquid samples, or small changes in instrument response due to short term variations in lamp intensity, detector response, or instrument temperature. Irrespectively of the order, both derivatives lead to an increased spectral resolution at the expense of a decreased signal-to-noise ratio. SNV, developed by Barnes et al. [11], is a scatter correction method used to normalize spectra when the effective pathlength varies among the samples of a data set. Such a pathlength variation can occur when measuring the spectra of granular or powdery samples if the sample presentation in a cell is not fully reproducible, or if the particle size varies among the samples. Each spectrum is mean centred and then divided by its standard deviation, so that the new spectra are centred in 0 and their standard deviations are 1, with a common scale for all the spectra. All the math pre-treatments discussed so far are based on and applied to individual spectra; they operate on data points of a given spectrum, and give results determined by the unique features of that spectrum. By contrast, MSC, developed by Geladi et al. [9], is set-dependent: it is a scatter correction method based on a related set of spectra, where the correction is carried out based on the assumption that all the samples have the same scatter coefficient at all the NIR wavelengths. In MSC, the mean spectrum is calculated from all the spectra in a defined data set. Then, a least squares linear regression is performed on absorbance values of the sample spectrum versus those at corresponding wavelengths in the mean spectrum. This operation provides a linear equation with a defined intercept and slope. Next, the value of the intercept is subtracted from every data point in the spectrum. Finally, each absorbance value in the resulting spectrum is divided by the value of the slope. Using the mean spectrum, the same set of operations is performed on every spectrum in the data set. In this way, MSC tries to separate multiplicative and additive effects of the scatter in NIR measurements, minimizing spectral variations that are not due to the analyte concentration. It has been demonstrated that MSC and SNV are linearly related, and thus should give similar results [12].

2. Theory

2.1. Orthogonal signal correction (OSC)

There is no proof that any of the signal processing techniques discussed in the previous section will remove only irrelevant information from the response matrix. For this reason, Wold et al. [13] developed orthogonal signal correction (OSC) to remove systematic variation from the predictor matrix $\boldsymbol{X}$ unrelated, or orthogonal, to the property matrix $\boldsymbol{Y}$. OSC is a method developed to reduce light scatter effects, and indeed more general interferences, whilst only removing the effects that have zero correlation with the reference value y. The idea is that all the information in the spectrum related to y should remain rather than be removed. The first step of the algorithm is to compute loading weights, $\hat{\boldsymbol{w}}$, so that the score vector $\hat{\boldsymbol{t}} = \boldsymbol{X}\hat{\boldsymbol{w}}$ describes as much as possible the variance of $\boldsymbol{X}$ under the constraint that it is uncorrelated to $\boldsymbol{Y}$, making $\hat{\boldsymbol{t}}$ as close as possible to the orthogonality to $\boldsymbol{Y}$. After having obtained this component, its effect is subtracted before computing any new component of the same type. The procedure can continue, but usually a very small number of "correction" components is needed. The residuals after OSC are used for regular PLS modelling. This treatment is applied at once to all the spectra in the calibration set. Then, the correction on the $\boldsymbol{X}$ matrix can be applied to an external evaluation set to validate the prediction ability of the calibration model built with the treated data. Since the introduction of the OSC method, a number of different attempts to improve the OSC method have been presented in the literature [14–21]. Furthermore, a MATLAB code for an OSC algorithm has been published on the Internet [22]; in this paper, this version has only been used to correct the set of evaluated spectra.

2.2. Direct orthogonal signal correction (DOSC)

The original OSC algorithm and all its modifications are not the only pre-processing methods to correct $\boldsymbol{X}$ by removing, at least, a part of the systematic variation unrelated to the studied responses. Recently, Westerhuis et al. [23] have proposed a new signal correction method, called direct orthogonal signal correction (DOSC). DOSC calculates components orthogonal to $\boldsymbol{Y}$ and describe the largest variation in $\boldsymbol{X}$. The method is developed using only simple least squares steps, and provides a theoretically exact solution to the problem set out by Wold.

In this work, the impact of data pre-processing on the determination of the ash content and lipids in roasted coffee by NIRS was studied in order to find a suitable pre-processing method capable of minimizing or suppressing systematic spectral differences. In turn, this would lead to an improved final calibration model, less complex and/or with an increased predictive ability and robustness. A comparison of the modelling power of PLS was made using unprocessed spectra, derivative spectra, SNV and MSC spectra, OSC-filtered spectra provided by Wise and Gallagher's algorithm, and DOSC-corrected spectra obtained by the Westerhuis algorithm.

3. Experimental

3.1. Apparatus and software

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

First and second derivatives, SNV on the spectra, PCA models and PLS calibrations were carried out with PARVUS (M. Forina, version 2000). Pre-processing by MSC was performed with Unscrambler 7.5 (CAMO, Trondheim, Norway). The OSC and DOSC routines were implemented in MATLAB 6.5 (MathWorks, Natick, USA).

3.2. Recording of 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 up 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 computed from the collected data.

3.3. Samples

The data set used in this study was composed of 83 roasted coffee samples of varied origins and varieties (35 *arabica* and 47 *robusta* coffees). The roasting degree for the various samples can be controlled by means of two significant parameters: the colour and the quantity (kg) put into the roaster. Thus, the samples of the data set were obtained from roasting processes where the quantity of roasted coffee ranged from 12 to 16 kg, working at six charge levels, whereas the colour ranged from 48 to 92 (arbitrary units).

In the case of the first response analysed, i.e. ash content, its value varies from 3.5 to 6.5% (w/w). The analytical method used to obtain the reference value for each sample involves to carbonise 5 g of coarsely ground coffee in a platinum capsule using a muffle-oven at 550 °C until the sample turned into white ash to constant weight. On the other hand, the second response studied, i.e., total lipids, ranges from 8.5 to 15.2% (w/w). Its reference method implies to weigh 2.5 g of ground sample in an extraction thimble, extract with ether in Soxlhet for 4.5 h, distil the excess ether and place the sample in oven under vacuum at 75 °C for 1.5 h. If necessary, the desiccation process is repeated until constant weight is reached.

Both reference methods were subjected to a validation study in laboratory, in terms of accuracy and precision, in such way that the variability of each one of the reference methods could be determined. With regard to the analysis reference method for the quantification of ash content its measured experimental error can be expressed as 2.10% CV, whereas in relation to the reference method for the determination of total lipids the observed variability was equal to 1.66% CV. These precision measurements are extremely important, as one has to keep in mind that the prediction error provided by a calibration model can never be lower than the experimental error associated to the respective reference method.

On the other hand, preliminary studies have been carried out in order to check the presence of possible outlier data which could have a detrimental effect on the quality of the calibration models. Thus, none of the diagnostics applied (residual and leverage plots, PCA) have shown the existence of anomalous samples (high leverage or high residual observations) in the data set used in this work.

3.4. Data processing

The whole wavelength range of 1100–2200 nm was selected as working region because none of the studied responses could be associated to defined spectral zones due to their global condition. The wavelength range of 2200–2500 nm, where the ratio signal-noise decreased considerably, was removed in all the cases. First derivative was applied using a cubic smoothing with a window size of 13 points. Partial least squares (PLS) was used for calibration within the abovementioned wavelength range. The data were centred before use. All the PLS regression models were constructed by cross-validation, using 10 cancellation groups in all cases. However, the comparison between the results provided by cross-validation and by the use of an external test set showed that the cross-validation led to an important over-fitting. For this reason, it was preferred to base the decisions concerning to the selection of the suitable calibration models (in terms of complexity and prediction errors) on the results observed in the external validation set. When OSC was used as pre-processing method, before the mean centering step, the spectra were transformed into their first derivative spectra, as this preliminary step improved significantly the quality of the model. All the PLS-models

tested were subject to external validation, performed on a set of 10 samples, randomly selected from the calibration matrix, and isolated as a test set. For both responses, it was verified that the samples in the external validation set covered perfectly the whole variable range. The quality of the results provided by the different pre-processing methods was compared using the root-mean-square error (RMSE) of the residuals obtained with the PLS model, defined as:

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

where y_i is the reference value, $\hat{y}_i$ the calculated value and n the total number of samples. RMSE is termed root mean square error in calibration (RMSEC) for the calibration set and root mean square error in prediction (RMSEP) for the prediction set. The number of latent variables to be used in each case was determined by the lowest RMSEP. These parameters had the advantage of being dimensionally comparable to the studied response.

4. Results and discussion

4.1. Spectral profiles

Due to physical differences among roasted coffee samples, which include, for example, particle size differences and variations in the compressing degree, the scatter effect can be significant and produce an expansion of the absorbance interval for the individual wavelengths. Therefore, the set of spectral profiles was clearly wider (Fig. 1a).

First derivative, MSC and SNV pre-treatments did not remove all the spectral displacements caused by the scatter effect; rather, they only reduced them at some spectral regions (Fig. 1b-d). On the other hand, as it can be seen, thanks to their respective plots, the similarity of the corrected spectra corresponding to MSC and SNV and, consequently, how both methods are closely related.

For the filtering by orthogonal signal correction methods (OSC and DOSC), Fig. 2 shows the spectra obtained after using the methods with each response variable considered, once the number of orthogonal latent variables to subtract from the original data had been optimised. As it can be observed, the spectral differences were significantly minimised compared to the raw spectra. At some wavelength intervals, the corrected spectra even overlapped, as the systematic variation unrelated to the analysed response had been rejected. The result provided by DOSC for the ash content was particularly interesting (Fig. 2b). This plot showed a high overlapping degree among spectra, except for a set of 10 samples, which appeared separate from the rest at some wavelength regions (1100–1400, 1700–1900, and 2000–2200 nm). The reason for this behaviour can be easily understood considering the ash content values of these samples: the ash content in this 10 samples was considerably higher than in the rest, and therefore, these values were separate from the group when information related to the response was considered.

4.2. Calibration models

The data set was split into two independent subsets, a calibration set with 73 samples and a test set with 10 randomly selected samples. The main caution which was taken in order to select a suitable test set was to verify that it covered appropriately and uniformly the whole range of values for both responses. The test set used was the same for all the pre-processing methods and models.

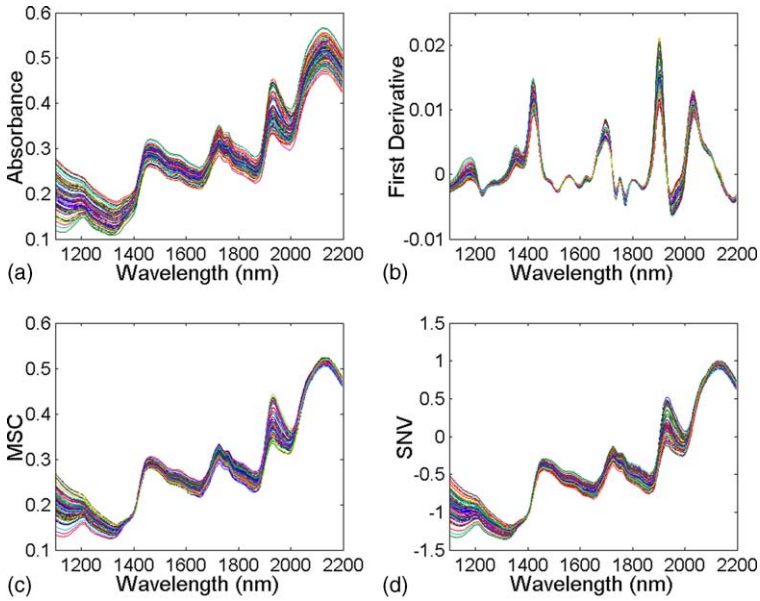

Fig. 1. (a) Original spectra; (b) first derivative spectra; (c) MSC spectra; (d) SNV spectra.

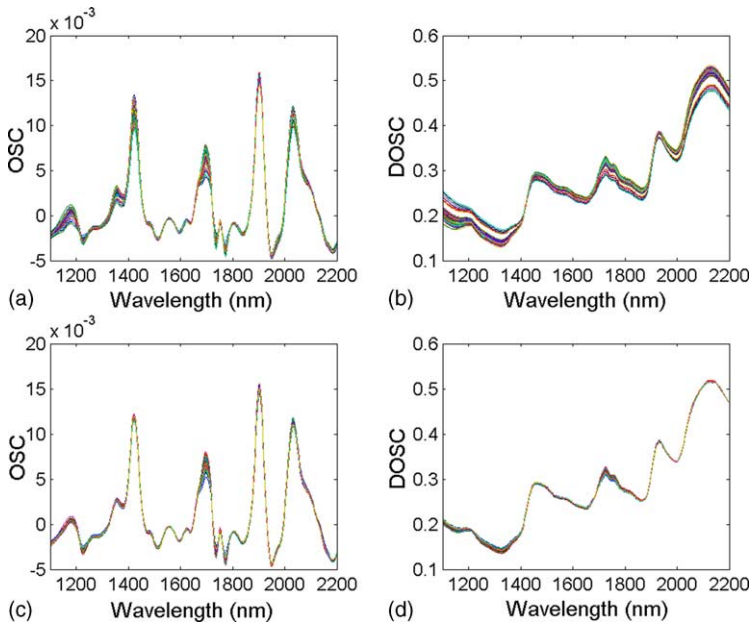

Fig. 2. (a) OSC and (b) DOSC spectra corrected for ash content as variable response. (c) OSC and (d) DOSC spectra corrected for total lipids as variable response.

It is interesting to be noticed that, despite what it would be thought a priori, it is not advisable to develop a separate regression model for each category of samples, because a large extent of the coffee commercialized in the market corresponds really to unknown blending of different varieties. In this way, it is much useful, from an actual point of view, to construct a global calibration model for predicting both responses.

When the orthogonal signal correction methods (OSC and DOSC) were evaluated, the calibration models obtained were as follows. First, an orthogonal component was obtained, and after subtraction from the raw data, a PLS model

Table 1
Calibration (RMSEC) and prediction (RMSEP) errors obtained by each data pre-processing method using the ash content as variable response

	PLS-LVs	RMSEC (%)	RMSEP (%)
Pre-processing method			
Original spectra	10	7.7	7.1
First derivative	9	6.0	5.9
Second derivative	9	5.9	5.9
MSC	10	6.8	7.7
SNV	10	6.9	8.3
OSC	**2**	**3.1**	**4.9**
DOSC	**4**	**4.7**	**4.9**
OSC-LVs			
1	8	4.1	5.4
2	**2**	**3.1**	**4.9**
3	1	2.0	6.5
4	1	1.2	8.7
5	1	0.9	9.5
6	1	0.6	10.0
DOSC-LVs			
1	3	5.0	5.8
2	3	5.0	5.7
3	5	4.7	5.0
4	**4**	**4.7**	**4.9**
5	3	4.8	5.4
6	2	4.9	5.6

OSC and DOSC have being applied varying the number of orthogonal components to be removed.

Table 2
Calibration (RMSEC) and prediction (RMSEP) errors obtained by each data pre-processing method using the total lipids as variable response

	PLS-LVs	RMSEC (%)	RMSEP (%)
Pre-processing method			
Original spectra	5	10.2	8.3
First derivative	9	7.0	7.2
Second derivative	9	6.3	6.2
MSC	10	8.4	10.9
SNV	10	8.5	11.4
OSC	**3**	**2.2**	**3.3**
DOSC	1	5.9	5.9
OSC-LVs			
1	9	4.5	5.6
2	5	3.6	4.1
3	**3**	**2.2**	**3.3**
4	1	1.2	4.2
5	1	0.5	5.7
6	1	0.2	6.6
DOSC-LVs			
1	8	5.6	6.4
2	7	5.5	6.4
3	4	5.4	6.4
4	3	5.8	6.3
5	4	5.8	6.3
6	**1**	**5.9**	**5.9**

OSC and DOSC have being applied varying the number of orthogonal components to be removed.

was built. The number of components to retain was chosen on the basis of the lowest prediction error. This was repeated for two orthogonal latent variables, etc. The PLS models were also obtained for the original data, the first and second derivative spectra, and the MSC and SNV spectra, without subtracting any orthogonal component.

The results obtained with the model selected for each treatment evaluated with their respective calibration and prediction errors are summarized in Table 1 using the ash content as variable response, and in Table 2 considering total lipids as variable response.

In order to find calibration models with a predictive ability as high as possible, the effect of the number of OSC and DOSC latent variables to be removed from the raw data was studied, testing their impact on the quantification of both responses (second and third sections in Tables 1 and 2).

Note that OSC can provide an overfitted solution with an extremely low calibration error but with a reduced predictive ability. This overfitting can be controlled by means of two parameters: the number of OSC factors (number of times that OSC is applied to a set of spectra) and the number of orthogonal latent variables (unrelated to the response) to be removed from the $\boldsymbol{X}$ matrix. In this paper, only the number of orthogonal latent variables was evaluated to determine the occurrence of this overfitting, as usually a single OSC run is enough to correct the data.

4.2.1. Ash content

When working on mean-centred original data, very poor results were obtained for the prediction set in spite of the large number of PLS latent variables (i.e., 10) used.

The comparison of calibrations based on first and second derivative spectra gave very similar results. Both optimal PLS models had nine components and, although the complexity of the models was similar to that obtained with the original data, their calibration and prediction errors were smaller (see Table 1).

In the light of the results shown in Table 1, the similarity between MSC and SNV is clear. However, the apparent improvement of the fitting ability provided by the models developed from MSC and SNV spectra was not confirmed by a corresponding improvement of their predictive ability. Therefore, it can be inferred from these results that both pre-treatments led to a significant overfitting, in such way that they were not suitable for this particular data set.

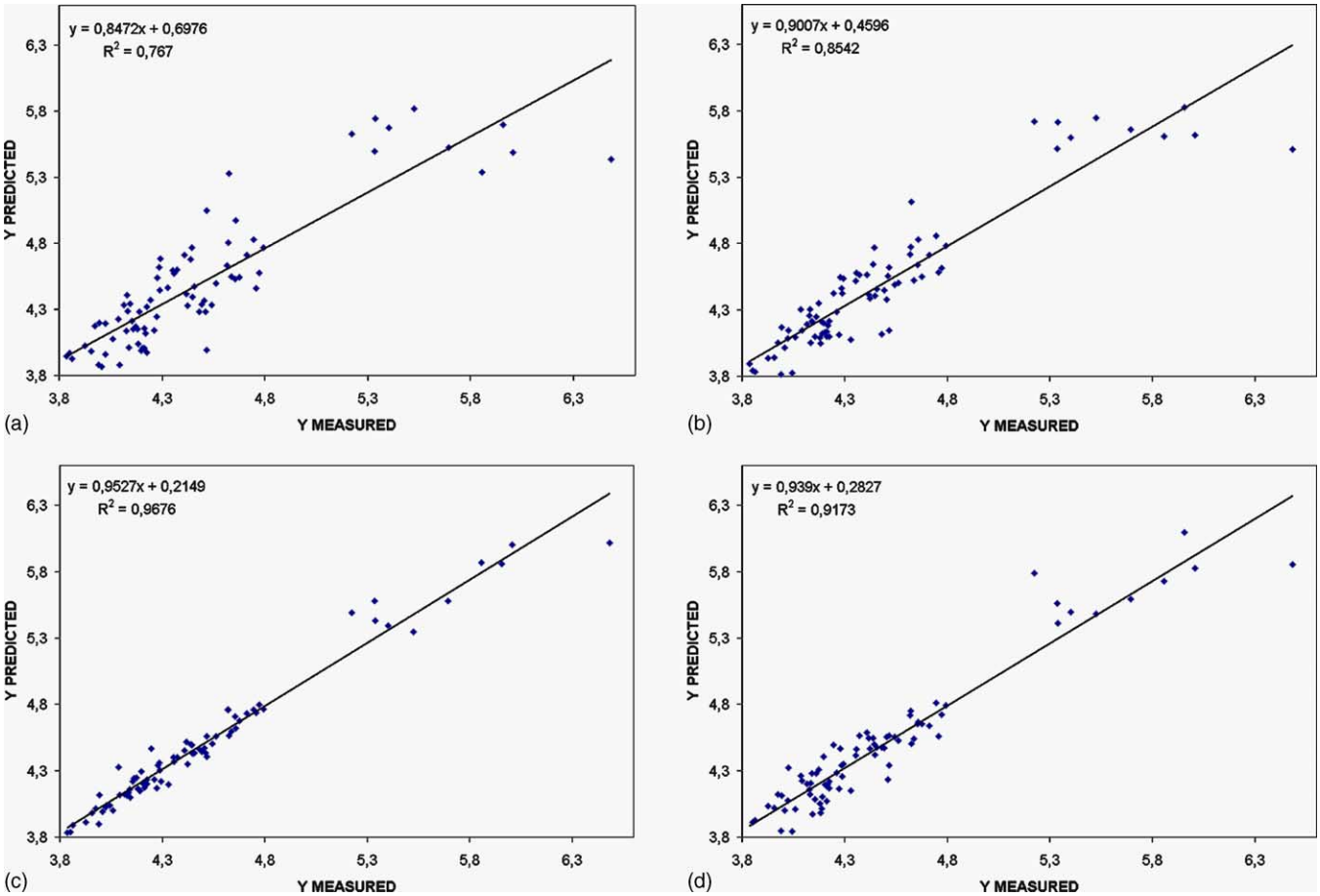

Fig. 3. Measured vs. predicted ash content values for the PLS models calculated from the roasted coffee data set. (a) Results from a model based on original data (only mean centered). (b) Results from a model based on mean centered first derivative spectra. (c) Results from a model based on OSC filtered spectra. (d) Results from a model based on DOSC corrected spectra.

After applying OSC, the calibration model obtained after removing two orthogonal latent variables had a significantly improved prediction ability (from 7.1% RMSEP with original data to 4.9% RMSEP with corrected data). From three orthogonal components, all the PLS models were obtained with a single PLS latent variable. However, the most suitable prediction results were still those obtained with two orthogonal components, as the removal of more than two orthogonal latent variables reduced the calibration error at the expense of increasing the prediction error, thus giving an overfitted solution (see second section in bold in Table 1).

For the data corrected by DOSC, third section in bold in Table 1 shows that the calibration and prediction errors were less dependent on the number of orthogonal latent variables removed from the raw data compared to OSC, since the various models built led to errors of the same magnitude. Nevertheless, considering the spectral profiles obtained after the DOSC treatment, it can be claimed that the removal of four DOSC-components minimized the scatter effects to a larger extent. Furthermore, this PLS model obtained with four PLS latent variables, after removing four DOSC components from the $\boldsymbol{X}$ matrix, gave the best prediction results using DOSC as pre-processing method (4.9% RMSEP), which implies a considerable improvement of the prediction ability compared to original data.

To clearly see the effect of the different pre-treatments on the calibration models, and the improvement achieved by the use of the orthogonal signal correction methods, Fig. 3 shows several plots of the regression line obtained fitting the predicted values for the ash content versus reference values for all samples in the data set, considering some of the pre-processing methods applied.

4.2.2. Total lipids

The negative impact of the scatter effects on the regression quality was proven by the poor predictive ability obtained when a PLS model was developed from original data to determine the total lipids (Table 2). This way, up to five PLS latent variables, both calibration error and prediction error decrease slightly, but from six latent variables an overfitting starts to appear. For this reason, a model with five PLS-components based on the original data could be taken as starting point for further comparisons, despite his low predictive ability (8.3% RMSEP).

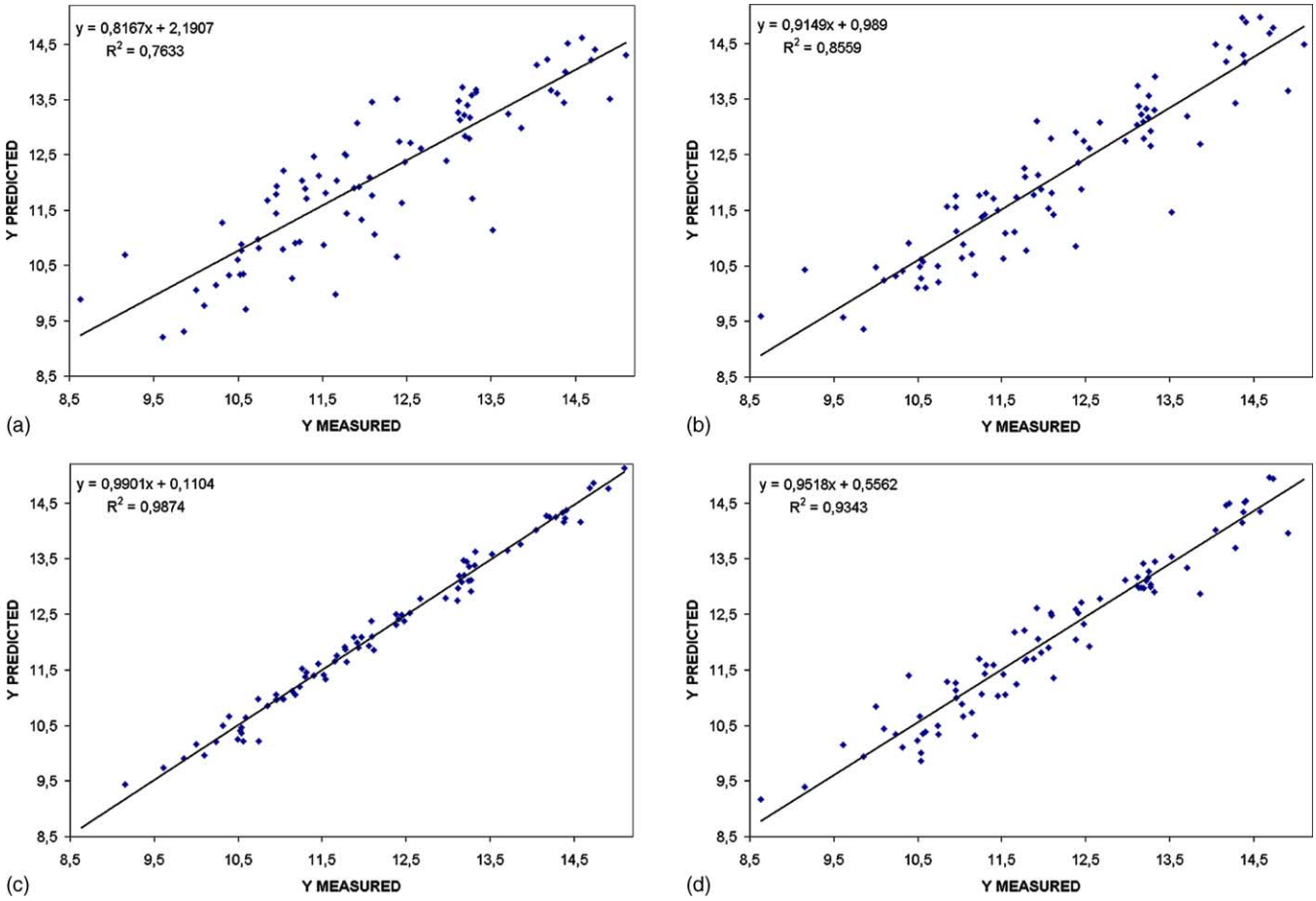

Fig. 4. Measured vs. predicted total lipids values for the PLS models calculated from the roasted coffee data set. (a) Results from a model based on original data (only mean centered). (b) Results from a model based on mean centered first derivative spectra. (c) Results from a model based on OSC filtered spectra. (d) Results from a model based on DOSC corrected spectra.

With first derivative spectra, a modest improvement of the predictive ability of the PLS model (7.2% RMSEP) was obtained, although using more PLS latent variables. On the other hand, when the total lipids were used as variable response, and in contrast with the ash content, the application of second derivative to spectra gave a PLS model with a significantly lower prediction error (5.6% RMSEP) compared to first derivative spectra and, of course, to raw spectra, although using more PLS components (see Table 2).

Once more, MSC and SNV behave similarly with the PLS models built from their respective corrected spectra. As in the case of the ash content quantification, a large overfitting occurred, and thus the results of both pre-processing methods cannot be considered reliable.

Considering the data filtered by OSC, it can be seen that when the number of orthogonal latent variable to be removed from the data increased, the complexity of the respective PLS models decreased, being necessary only one PLS component from the removal of four orthogonal LVs. Note that the fitting ability improved significantly once each orthogonal component had been removed. Nevertheless, this gradual fall in the calibration error was not accompanied by a similar decrease in the prediction error, and thus from four orthogonal latent variables subtracted, it was clear that the solution reached was overfitted. Therefore, the model with three PLS components, built on data obtained after the removal of three OSC components, was the most suitable to ensure a high predictive ability of the final model (see second section in bold in Table 2). The selected model implied a significant improvement of the prediction quality compared to the model based on non-pre-processed data (3.3% RMSEP versus 8.3% RMSEP).

Considering the results obtained after applying DOSC to the data set, no overfitting problems were observed, and all the PLS models developed by removing a number of orthogonal components showed similar errors, although considerably higher than OSC (third section in bold in Table 2). The model that used a single PLS component, built by subtracting six DOSC latent variables from the raw data, appeared to be the most suitable model, as it had the highest predictive ability (5.9% RMSEP) with the lowest level of complexity. Thus, the decrease in the prediction error compared to the original model can be considered as modest, but always greater than with the common applied pre-processing methods anyway.

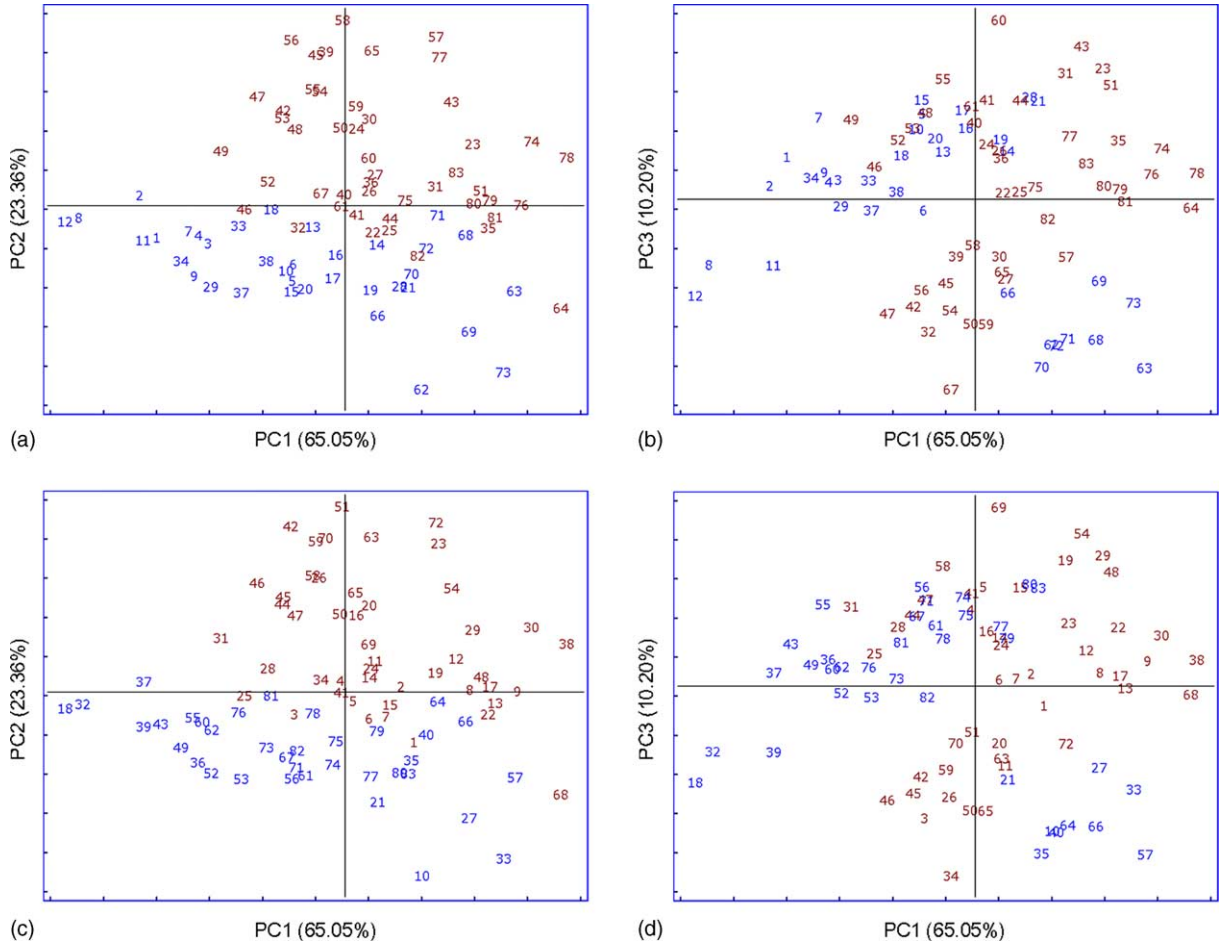

Fig. 5. Principal component analysis (PCA) from original spectra. The variance explained by each PC is shown in brackets. (a) PC1 vs. PC2, (b) PC1 vs. PC3 (samples ordered according to their ash content); (c) PC1 vs. PC2, (d) PC1 vs. PC3 (samples ordered according to their total lipids content).

Again, as in the case of ash content, the plots summarized in Fig. 4, containing the regression lines obtained fitting the predicted values for the total lipids versus their reference values for all samples, allow to see even more clearly the effect of the several pre-treatments on the regression models.

4.3. Principal component analysis (PCA)

Finally, a principal component analysis was carried out, both on original spectra and on OSC and DOSC corrected spectra for the studied responses in order to verify whether the promising results obtained with these orthogonal signal correction methods tallied with the conclusions drawn from the respective score plots.

In the plots that will be discussed next, an index was assigned to each sample following an increasing order as regards the respective values of the considered response. The score plots were interpreted according to the individual values of ash content and total lipids, as well as the coffee variety of each sample.

Fig. 5 shows the score plots after computing the first three principal components on the original spectra, which account for 98.6% of the variance in the data. However, clear patterns related to the studied responses could not be observed in the distribution of the samples along the different component axes. Only the second component showed a limited ability to discriminate between samples of different varieties. Therefore, it was concluded that these first components that captured most of the variance in the data were not very closely related to the modelled responses. Thus, it is logical to think that the respective regression models developed using these original data to determine both responses must have more PLS latent variables without necessarily giving a small enough prediction error.

It was expected to change this behaviour by applying orthogonal signal correction methods to remove the information unrelated to the studied response from the data matrix.

Taking the ash content as variable response and analyzing the score plots corresponding to the three first principal components computed from the corrected data after applying OSC, which accounted altogether for more than 95% of the variance in the data (Fig. 6a and b), it can be observed that the samples were perfectly arranged along the first component axis according to their ash content, and that this content decreased for objects placed on the left. Therefore, the direct connection between this maximum variance

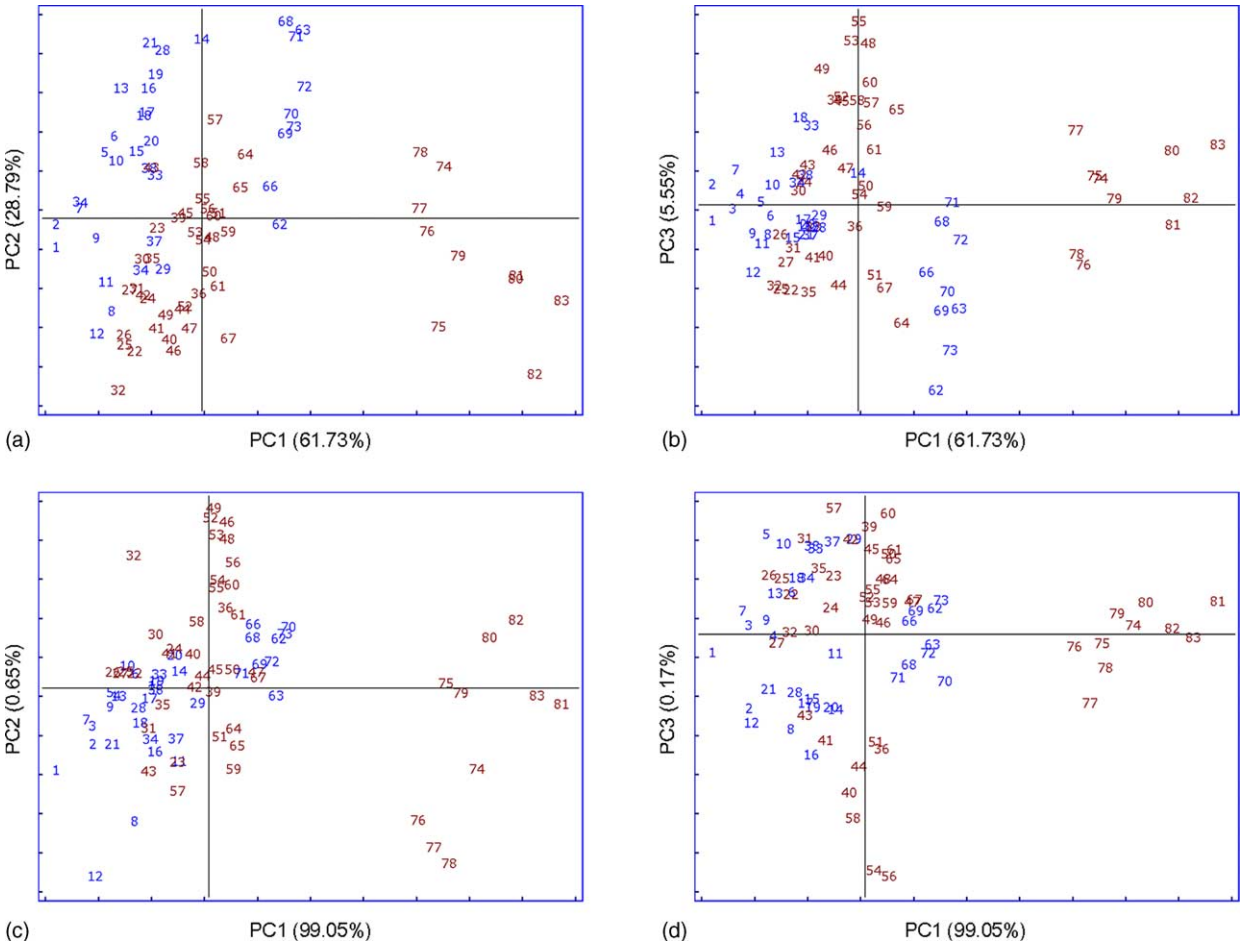

Fig. 6. Principal component analysis (PCA) from OSC and DOSC corrected spectra for ash content as variable response. The explained variance for each PC is shown in brackets. (a) OSC-PC1 vs. PC2; (b) OSC-PC1 vs. PC3; (c) DOSC-PC1 vs. PC2; (d) DOSC-PC1 vs. PC3.

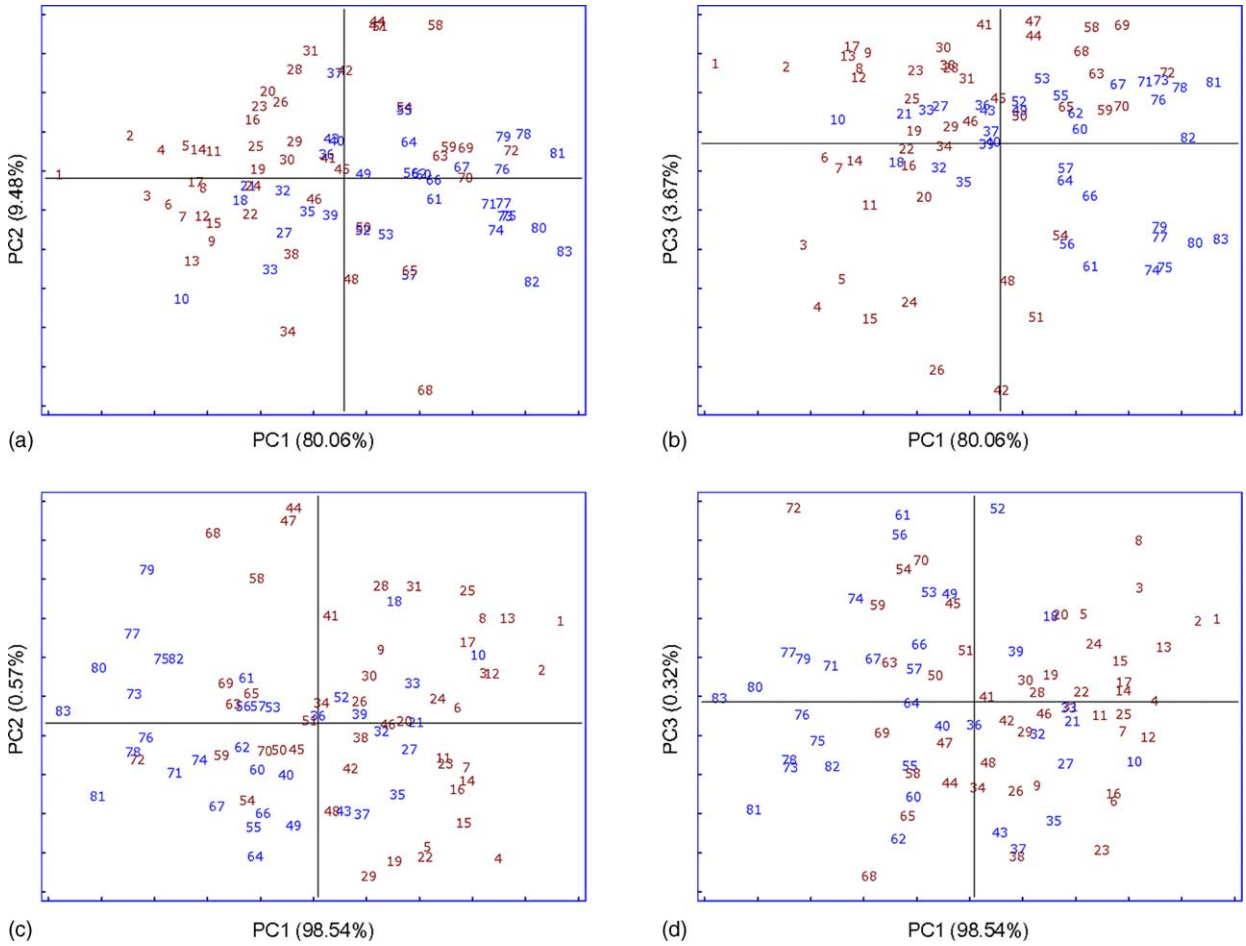

Fig. 7. Principal component analysis (PCA) from OSC and DOSC corrected spectra for total lipids as variable response. The explained variance for each PC is shown in brackets. (a) OSC-PC1 vs. PC2; (b) OSC-PC1 vs. PC3; (c) DOSC-PC1 vs. PC2; (d) DOSC-PC1 vs. PC3.

component (more than 60% of the variance in the data) and the modelled response was clear, and this proved the efficiency of OSC to correct the data.

On the other hand, when the orthogonal signal correction method selected to filter the data before developing a regression model for the ash content was DOSC, the first principal component alone accounted for 99% of the variance in the data (Fig. 6c and d). Moreover, it could be seen that, along this first axis, the samples were exactly arranged according to their ash content, with an increase from left to right. Thus, the behaviour of the samples along a maximum variance principal component could be a proof of the efficiency of DOSC.

When the pre-processing method applied was again OSC, but this time working with the total lipids as variable response, the conclusions drawn from the analysis of the corresponding score plots were the same than for the ash content. This time, the three first principal components accounted for 93.2% of the variance in the data (Fig. 7a and b), and the samples were again perfectly arranged along the first component axis, with an increase in the total lipid content from left to right. The only difference was that the first component directly related to the modelled response accounted for a larger variance (80%). Therefore, OSC proved once more its ability to correct the data.

The score plots obtained after applying DOSC to the raw data, taking the total lipids as variable response, also proved the efficiency of DOSC (Fig. 7c and d). The first principal component alone accounted for more than 98% of the variance in the data and the rest of the components were responsible for a small part of the data variability. Again, the first component of the samples was perfectly arranged along its axis according to their total lipid content, with an increase from right to left.

5. Conclusions

In this work, it has been confirmed that the pre-processing methods most usually applied (derivation, MSC, SNV) can reduce the spectral variability caused by scatter effects, or what it is the same, the variation unrelated to the modelled response, and that, in some cases, they lead to better calibration models compared to those obtained from original spectra. However, taking into account the results obtained from the data sets studied, it can be conclude that none of

these pre-treatments can remove completely all the systematic variability, and the optimal complexities of the built PLS models are still high. By contrast, OSC and DOSC have proven their relative effectiveness to correct coffee spectra in the studied data sets, both for the quantification of ash content and for the determination of total lipids, removing at least a portion of information unrelated to the response, leading to significantly improved and simpler calibration models that only require a few components (never more than four in the data sets used in this study) to model the data and providing predictive abilities much higher than other methods. Therefore, it can be claimed that both OSC and DOSC have proven to be quite appropriate pre-processing methods for the applications presented in this work, as they can be used to develop reliable regression models based on filtered data. Nevertheless, despite the promising results obtained in the present study, for the moment we must limit the scope of the conclusions which can be drawn from them in the absence of a deeper study which we intend to do in future.

Acknowledgements

The authors thank the Ministry of Science and Technology (Project no. 2FD1997-0491) and the University of La Rioja (Research Grant FPI-2001) for their financial support.

References

[1] A. Illy, E. Illy, R. Macrae, M. Petracco, M.R. Sondahl, S. Valussi, R. Viani, Espresso Coffee: The Chemistry of Quality, Academic Press, England, 1995.
[2] Official Methods of Analysis of AOAC International, 16th ed., AOAC International.
[3] K.I. Hildrum, T. Isaksson, T. Næs, A. Tandberg, Near-Infrared Spectroscopy: Bridging The Gap Between Data Analysis and NIR Applications, Horwood, England, 1992.
[4] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy, 2nd ed., Longman, Harlow, UK, 1993.
[5] D.L. Wetzel, Anal. Chem. 55 (1983) 1165.
[6] W.F. McClure, NIR News 4 (6) (1993) 12.
[7] W.F. McClure, NIR News 5 (1) (1994) 12.
[8] T. Isaksson, T. Næs, Appl. Spectrosc. 42 (1988) 1273.
[9] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (3) (1985) 491.
[10] T. Næs, T. Isaksson, B. Kowalski, Anal. Chem. 62 (1990) 664.
[11] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (5) (1989) 772.
[12] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, J. Near Infrared Spectrosc. 2 (1994) 43.
[13] S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemom. Intell. Lab. Syst. 44 (1998) 175.
[14] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, Chemom. Intell. Lab. Syst. 44 (1998) 229.
[15] C.A. Andersson, Chemom. Intell. Lab. Syst. 47 (1999) 51.
[16] T. Fearn, Chemom. Intell. Lab. Syst. 50 (2000) 47.
[17] J.A. Fernández Pierna, D.L. Massart, O.E. de Noord, Ph. Ricoux, Chemom. Intell. Lab. Syst. 55 (2001) 101.
[18] S. Wold, J. Trygg, A. Berglund, H. Antii, Chemom. Intell. Lab. Syst. 58 (2001) 131.
[19] M. Blanco, J. Coello, I. Montoliu, M.A. Romero, Anal. Chim. Acta 434 (2001) 125.
[20] J. Trygg, S. Wold, J. Chemom. 16 (2002) 119.
[21] R.N. Feudale, H. Tan, S.D. Brown, Chemom. Intell. Lab. Syst. 63 (2002) 129.
[22] B.M. Wise, N.B. Gallagher, http://www.eigenvector.com/MATLAB/OSC.html.
[23] J.A. Westerhuis, S. de Jong, A.K. Smilde, Chemom. Intell. Lab. Syst. 56 (2001) 13.