Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 514 (2004) 57–67
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

An evaluation of orthogonal signal correction methods for the characterisation of *arabica* and *robusta* coffee varieties by NIRS

I. Esteban-Díez, J.M. González-Sáiz, C. Pizarro*

Department of Chemistry, University of La Rioja, C/Madre de Dios 51, 26006 Logroño (La Rioja) Spain

Received 12 January 2004; received in revised form 8 March 2004; accepted 8 March 2004

Abstract

Two orthogonal signal correction methods (OSC and DOSC) were applied on a set of 83 roasted coffee NIR spectra from varied origins and varieties in order to remove information unrelated to a specific chemical response (caffeine), which was selected due to its high discriminant ability to differentiate between *arabica* and *robusta* coffee varieties. These corrected NIR spectra, as well as raw NIR spectra and three chemical quantities (caffeine, chlorogenic acids and total acidity), were used to develop separate classification models accordingly using the potential functions method as a class-modelling technique in order to evaluate their respective capacities to discriminate between coffee varieties and the influence of these pre-processing methods on the classification of the coffee samples into their corresponding variety class. The transformation of roasted coffee NIR spectra by means of an orthogonal signal correction method, taking into account in this correction a chemical response closely related to the sample origin, prompted a notable improvement in the specificity of the constructed classification models.

Keywords: Multivariate classification; NIR spectroscopy; Orthogonal signal correction; Roasted coffee; Potential functions

1. Introduction

Coffee is the world's most popular beverage after water, with over 400 billion cups consumed annually. At least 66 species of the genus *Coffea L.* have been identified so far, and two of these varieties are economically and commercially important: *C. arabica L.* (*arabica* coffee) and *C. canephora Pierre* (*robusta* coffee) [1]. At present, most commercially available coffee beverages are produced from *arabica* and *robusta* roasted beans or blends of these two. Both coffee varieties differ not only from a botanical standpoint but also in terms of quality. These differences are recognised commercially and *robusta* usually sells at prices 20–25% lower than *arabica*, more appreciated by the consumers. Therefore, it is easy to understand the huge relevance of the availability of suitable analytical methods, combined with proper chemometrical tools, in order to differentiate between these two coffee varieties. The two species exhibit considerable differences in their botanical, genetic, agronomical, chemical and morphological characteristics. After roasting and grinding, the visual criterion cannot be used and other methods must be used in order to differentiate between varieties. The differences detected in the chemical composition of *arabica* and *robusta* coffees can be useful for classification purposes, since several chemical parameters show a marked inter-variety difference. Some of the chemical descriptors used to characterise both varieties include: metal content [2], volatile components [3], chlorogenic acid and caffeine [4], fatty acid profiles [5], sterolic profile [6], diterpenic alcohols [7], tocopherols and triglycerides [8]. Nevertheless, despite the good results provided by these approaches, it is important to bear in mind that many analytical reference methods used to determine the significant chemical descriptors to be considered in the development of a classification model may be quite elaborate and/or time-consuming. It would be useful to be able to construct an accurate classification model from measurements obtained using a method as fast, clean and inexpensive as, for example, near-infrared spectroscopy has demonstrated to be. In fact, during the last decade one of the most common applications of near-infrared (NIR) spectroscopy combined with pattern recognition methods has been to discriminate between samples belonging to one of several distinct groups based on spectral properties [9–26]. Likewise, the potential

* Corresponding author. Tel.: +34-941-299592; fax: +34-941-1299587.
E-mail address: consuelo.pizarro@dq.unirioja.es (C. Pizarro).

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.03.022

of NIR reflectance spectroscopy for discriminating between *arabica* and *robusta* coffees has also been investigated with relative success [27–29]. Nevertheless, although it has been proven that original NIR spectra of coffee samples might be directly used in order to develop classification models with good discrimination abilities between pure coffee varieties, the observed values of inter-classes specificity were not so high as it would be desirable to ensure that extreme samples within each variety to be unequivocally classified into their right category or to avoid potential errors when working with coffee blends. In fact, certain effects that occur inherently in diffuse reflectance NIR spectroscopy for solid samples, such as light scattering and influence of particle size, could be responsible for perturbations in spectra (baseline shifts, slope changes and curvilinearity). As a result, NIR spectra contain not only chemical but also physical information about samples and measuring conditions, which may be irrelevant and can mask the chemical information in the spectra (including information closely related to sample origin), and might deteriorate classification models developed from raw NIR spectra. Therefore, the application of a suitable pre-processing method, aimed at minimising the contribution of physical effects to NIR spectra and thus enhancing the chemical information contained, could be seen as an important step in model development and improvement.

For all these reasons, the aim of this study is precisely to propose a strategy for developing improved and reliable classification models for discriminating between *arabica* and *robusta* coffee varieties based on their NIR spectra. The potential functions method was selected as class-modelling technique for this study, since it is a powerful method with certain specific features that enable a very comprehensive analysis of both numerical and graphical results to be performed. In this way, as an attempt to improve the classification models constructed on the basis of original roasted coffee NIR spectra, two orthogonal signal correction methods (OSC and DOSC) were applied on these raw spectra in order to remove information not related to a specific chemical response with a high modelling power to discriminate between coffee varieties. This specific descriptor was selected from among three chemical parameters (caffeine, chlorogenic acid and total acidity), which generally show a notable variation between varieties, taking into account the respective results yielded by the classification models constructed from these individual chemical responses. For evaluating the effect of the orthogonal corrections on sample classification, the results obtained before and after transforming the spectra were analysed and compared.

2. Theory

2.1. Orthogonal signal correction pre-processing methods

In recent years, there have been several attempts to deal with the problem of correcting spectral data before they are used as the basis to develop a calibration or classification model, beyond a simple mathematical filtering. Wold et al. introduced a novel spectral data treatment technique called orthogonal signal correction (OSC) [30]. Since the introduction of OSC, a number of different approaches that have attempted to improve or modify it, have been presented in literature [31–34]. Likewise, Westerhuis et al. developed another similar orthogonal signal correction technique called direct orthogonal signal correction (DOSC) [35]. The aim of all these orthogonal signal correction methods is the same, i.e. to correct the $\boldsymbol{X}$ data matrix by removing information that it is orthogonal to a response matrix $\boldsymbol{Y}$, i.e. information not related to a response of interest. These pre-processing methods are applied jointly to all spectra in the calibration set. Later, the correction performed on the $\boldsymbol{X}$ matrix can be applied to an external evaluation set in order to check the real prediction ability of the model constructed from the corrected data. The main difference between OSC and DOSC lies in the procedure applied to compute the orthogonal directions to be removed:

- OSC provides a PLS-based solution, in such way that the condition that weights should be computed in order to minimise the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is imposed;
- DOSC finds an exact solution to the orthogonality constraint in the calculation of the orthogonal components using an approach based solely on least squares steps.

The specific algorithms used in both methods are described in depth in the respective original works [30,35].

2.2. Classification method

The classification of roasted coffee samples by variety requires a method that yields a positive identification, i.e. a sample should be classified as belonging to a class only if it is similar enough to that considered class [36]. For this reason a class-modelling technique, such as the potential functions methods, was selected for use in this particular study.

2.2.1. Potential functions methods

Since the appearance of potential functions as a classification method in analytical chemistry (1950s), several classification methods and clustering procedures have been developed based on that method [37–39], also being applied for more than simply classification purposes, such as selecting representative subsets of samples [40].

The classification methods based on potential functions have been properly modified in order to obtain the corresponding class-modelling techniques (used in the present study). It is beyond the scope of this paper to describe fundamentals of potential functions methods, so for further descriptions more specific articles are recommended.

3. Experimental

3.1. Samples

The data set used in the present study comprised 83 roasted coffee samples from varied origins and varieties (36 *arabica* and 47 *robusta* coffees), which were processed under different roasting conditions. For each sample, three chemical parameters (caffeine, chlorogenic acids and total acidity) were determined in order to later study their respective capacities as inter-variety classifiers. As regards the specific levels for each of these chemical quantities corresponding to the coffees used in this study, caffeine content ranged from 1.00% to 2.35% (w/w), chlorogenic acid content ranged from 2.15% to 4.50% (w/w), and total acidity ranged from 6.3 ml to 23.9 ml (0.1 N alkali required to neutralise acidity of 100 g sample).

This data set was split into two independent subsets: a calibration set with 67 samples and a test set with 16 samples (7 *arabica* and 9 *robusta* coffees). The main cautions taken in order to select a suitable composition of the external test set were to include representative samples of both varieties and to verify that the contained samples uniformly covered the whole range of values for all chemical descriptors considered.

3.2. Apparatus and software

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, thermostatted column compartment, diode array detector and an HP Chemstation data processing system (Hewlett-Packard) to perform peak purity analyses. The column used was Zorbax SB-C_{18}, 250 mm × 4.6 mm i.d. with 5 μm particle size (Hewlett-Packard GmbH, Waldbronn, Germany). 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.

Spectrophotometric measurements at 324 nm were performed on a Milton Roy Spectronic 1201 spectrophotometer (Milton Roy Company, Rochester, NY, USA) equipped with a modified Czerny–Turner monochromator with holographic grating, a photomultiplier tube and tungsten–halogen/deuterium light sources.

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 NIR Systems, Raamsdonksveer, The Netherlands) was used to acquire the data.

Data pre-processing treatments and potential functions class-modelling technique were applied by means of V-PARVUS 2004 (Forina et al., Dipartimento di Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova, Italy). The OSC and DOSC routines were implemented in MATLAB 6.5 (Mathworks, Natick, USA). Specifically, the OSC method developed by Wise et al., available in PLS-Toolbox 2.1 for use with MATLAB, was used for the OSC calculations. Data for isopotential lines obtained from PARVUS were later mapped using Surfer 8 (Golden Software, Inc.).

3.3. Determination of chemical descriptors

Total acidity of roasted coffee samples was determined using an official method proposed by the AOAC (920.92). The two remaining parameters (caffeine and chlorogenic acid contents) were determined by analytical methods designed in our own laboratory, implying some adaptation with respect to the respective official method. All these analytical methods (including the official method) underwent a full validation study to determine accuracy, precision (repeatability and reproducibility), specificity, selectivity, linear range, detection and quantification limits. Thus, a measurement of the accuracy and precision of all the validated reference methods was obtained by means of the relative standard deviations expressed in percentage, which were equal to 1.25, 1.15 and 1.28 (% R.S.D.) when considering caffeine, chlorogenic acids and total acidity, respectively.

3.3.1. Caffeine

The analytical method used to obtain the caffeine content value for each sample involved accurately weighing 7 g roasted coffee (P_{sample}) and transferring it to a weighed two-neck flask (P_{flask}). One hundred millilitres of H_2O was then added, and the sample was heated under reflux for 45 min, cooled to room temperature, adding H_2O to wash walls, and then the flask (P_{solution}) was weighed again. An aliquot of this solution was centrifuged (at 11,000 rpm for 10 min) and 7.5 ml (P_{aliquot}) of the clear liquid transferred to a 250 ml (V_{flask}) volumetric flask, diluting to volume. It was then filtered through a 0.22 μm filter and transferred to a vial. The sample was injected in the HPLC system provided by a diode array detector, and the signal read at 276 nm. The previously calibration performed on the basis of standard solutions enabled us to obtain the caffeine concentration (C_{cal}). The percentage of caffeine in the roasted coffee sample was then obtained using the following expression:

$$(\%)_{\text{caffeine}} = \frac{C_{\text{cal}} \cdot V_{\text{flask}} \cdot \rho_{H_2O}(P_{\text{solution}} - P_{\text{flask}})}{P_{\text{aliquot}} \cdot P_{\text{sample}} \cdot 1000 \cdot 1000} \cdot 100 \quad (1)$$

Chromatographic separation was optimised using a gradient mobile phase, starting from 100% mili-Q water, gradually increasing the acetonitrile percentage (gradient grade) to 25% in 4 min, and keeping this solvent percentage constant until the end of the analysis (7 min). The stabilisation time between consecutive injections was fixed at 2 min.

3.3.2. Chlorogenic acid

The analytical method used to obtain the chlorogenic acid content value for each sample involved accurately weighing 7 g roasted coffee (P_{sample}) and transferring it to a weighed two-neck flask (P_{flask}). One hundred millilitres of H_2O was then added and the sample was heated under reflux for 45 min. It was then cooled to room temperature, adding H_2O to wash walls, and the flask ($P_{solution}$) was weighed again. An aliquot of this solution was centrifuged (at 11,000 rpm for 10 min) and 7.5 ml ($P_{aliquot}$) of the clear liquid transferred to a 250 ml (V_{flask}) volumetric flask, diluting to volume. This solution was divided into two separate aliquots:

- Part A: 5 ml (V_1) filtrate was transferred to 50 ml (V_2) volumetric flask, diluted to volume with H_2O and absorbance A at 324 nm was determined.
- Part B: 50 ml sample solution was transferred to 100 ml volumetric flask. 1 ml saturated KCH_3COO solution and 5 ml basic $Pb(CH_3COO)_2$ solution were added with swirling. Flask was placed in boiling H_2O bath 5 min, swirling occasionally. It was then removed, cooled under tap, and placed in ice–H_2O bath stirring mechanically for 1 h. Next, flask was removed, warmed to room temperature, and diluted to volume with H_2O. Sample was filtered through fluted paper, discarding first 25–50 ml filtrate. Absorbance B at 324 nm was immediately determined.

The following values were determined from the standard curve:

- Apparent concentration of chlorogenic acid in solution taken for A measurement without Pb treatment (C_A);
- Apparent concentration in filtrate after Pb treatment (C_B).

From the latter value 0.00045 mg/ml was subtracted to correct for the solubility of lead chlorogenate.

In this way, the corrected concentration of chlorogenic acid and the respective percentage of chlorogenic acid in roasted coffee were obtained:

$$C_{corr} = C_A - \left(\frac{C_B - 0.00045}{5}\right) \quad (2)$$

$$(\%)_{chlorogenic} = \frac{C_{corr} \cdot V_{flask} \cdot V_2 \cdot \rho_{H_2O}(P_{solution} - P_{flask})}{P_{aliquot} \cdot P_{sample} \cdot V_1 \cdot 1000 \cdot 1000} \cdot 100 \quad (3)$$

3.3.3. Total acidity

Ten grams of sample (P_{sample}) were treated in Erlenmeyer with 75 ml (V_1) 80% alcohol, stopper, and left to stand for 16 h, shaking occasionally. It was then filtered and 10 ml (V_2) of the filtrate transferred to a beaker, diluted to ca 100 ml (V_3) with H_2O, and titrated with 0.1 N alkali, using phenolphthalein. The result was expressed as ml (V_{alkali}) 0.1 N alkali required to neutralise acidity of 100 g sample:

$$\text{acidity} = \frac{V_{alkali} \cdot F \cdot V_3 \cdot V_1}{P_{sample} \cdot V_2} \cdot 100 \quad (4)$$

3.4. 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 up the sample cell. Each spectrum was obtained from 32 scans performed at 2 nm intervals within the wavelength range 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 replicates.

3.5. Validation methods

Usually, the class-modelling technique based on potential functions applied in this work validates the predictive ability of classification models constructed by cross-validation, since when working with potential functions is not recommended to waste objects to make up an external evaluation set. However, it is known, that orthogonal signal correction methods can produce a notable overfitting when applied on the spectra forming the training set.

For this reason, although all potential funcions classification models were constructed by cross-validation, we decided to also validate the actual predictive abilities of resulting models by testing their perfomance on an external test set, simply to control and avoid a possible overfitting that could inherently appear in these orthogonal signal correction methods.

3.6. Data processing

The models used to classify roasted coffee samples were constructed by using the potential functions method in its modified form as a class-modelling technique. The optimal value of the smoothing parameter was selected by means of a cross-validation procedure, in such a way that the amplitude of each individual potential, defined by this smoothing parameter, was the same for all the objects in the category (fixed potential functions). Model boundaries were computed from the estimate of the equivalent determinant. The class-models were constructed at a level of significance corresponding to 95%. The results provided by each classification model were evaluated by means of cross-validation, using five cancellation groups in all cases. The same a priori class probability was applied to both categories (equal to one). When classification models were developed on the basis of NIR data, given the large number of spectroscopic variables, a prior step of dimensionality reduction computing a small number of principal components was required. One crucial step in modelling based on NIR spectra is the selection of the optimal number of PCs that must be used in model development. Thus, the PCs to include in the model were chosen according to the highest correct classification rate, in terms of both classification and validation, comparing the performance of class-models from 1 to 5 PCs, when working on raw and corrected spectra. The wavelength range

2200–2500 nm, where the signal-noise ratio decreases considerably, was removed from the spectral matrix. Data were always centred, and when OSC was used as a pre-processing method, the spectra were additionally transformed into their first derivative spectra (using cubic smoothing with a window size of thirteen points) before the mean centering step, since this preliminary treatment led to a deeper correction. When orthogonal signal correction methods (OSC and DOSC) were applied to remove information not related to caffeine content, the number of orthogonal components to be subtracted varied from 1 to 3, in order to determine the optimal correction degree to be used. The quality of the results provided by the different class-models constructed was compared according to several evaluation parameters:

- Classification (prediction) rate

$$TR = \frac{\sum_c m_{cc}}{N} \quad (8)$$

- Category rate

$$R_c = \frac{m_{cc}}{N_c} \quad (9)$$

These equations were applied in both classification and prediction, where m_{cc} is the correct classification number and N the total classification or prediction number (during cross-validation an object is classified many times). Graphical tools, such as isopotential lines and Coomans plots, were also used to analyse the goodness of the models.

4. Results and discussion

4.1. Original NIR spectra

Table 1 summarises the classification and prediction rates corresponding to the class-models developed on the basis of mean-centred original NIR spectra of roasted coffee samples using the potential functions method from 1 to 5 PCs for the two-class problem analysed. The table shows the influence on the number of PCs used to construct the class-model, in such a way that it was necessary to include at least 4 PCs in the model to reach maximum correct classification rates in both calibration and prediction, i.e., to obtain a satisfactory classification model of coffee varieties.

These numerical results can be also confirmed graphically. In a Coomans plot, the axes represent categories. The 'coordinates' on these axes provide a measurement of distance of each sample from class-models. In this way, each class model is defined by a class space delimited by the critical distances corresponding to each class. Each model will accept samples whose distance to the corresponding centroid is lower than the critical distance. Samples located in the lower left square of the diagram lie in a uncertainty zone (accepted by both models). Bearing in mind the categories studied here, the vertical rectangle that appears in this type of plot corresponded to *arabica* coffees, whereas the vertical one corresponded to *robusta* ones. Fig. 1 shows Coomans diagrams corresponding to classification models constructed from raw NIR spectra with complexities from 1 to 3 PCs.

Table 1
Percentages of correctly classified samples

PCs	Classification (%)			C.V. (%)			Prediction (%)		
	RC1	RC2	TR	RC1	RC2	TR	RC1	RC2	TR
Mean centred NIR spectra									
1	31.03(20)	78.95(8)	58.21(28)	34.48(19)	81.58(7)	61.19(26)	42.86(4)	88.89(1)	68.75(5)
2	89.66(3)	92.11(3)	91.04(6)	93.10(2)	92.11(3)	92.54(5)	100	77.78(2)	87.50(2)
3	96.55(1)	89.47(4)	92.54(5)	96.55(1)	89.47(4)	92.54(5)	100	77.78(2)	87.50(2)
4	100	100	100	100	100	100	100	100	100
5	100	97.37(1)	98.51(1)	100	97.37(1)	98.51(1)	100	100	100
OSC-corrected NIR spectra (1 O-LV)									
1	72.41(8)	86.84(5)	80.60(13)	68.97(9)	86.84(5)	79.10(14)	100	77.78(2)	87.50(2)
2	100	100	100	100	100	100	100	100	100
OSC-corrected NIR spectra (2 O-LVs)									
1	100	100	100	100	100	100	100	100	100
DOSC-corrected NIR spectra (1 O-PC)									
1	93.10(2)	92.11(3)	92.54(5)	93.10(2)	92.11(3)	92.54(5)	100	77.78(2)	87.50(2)
2	100	89.47(4)	94.03(4)	100	89.47(4)	94.03(4)	100	77.78(2)	87.50(2)
3	100	100	100	100	100	100	100	100	100
DOSC-corrected NIR spectra (2 O-PCs)									
1	100	97.37(1)	98.51(1)	100	97.37(1)	98.51(1)	100	88.89(1)	93.75(1)
2	100	100	100	100	100	100	100	100	100
DOSC-corrected NIR spectra (3 O-PCs)									
1	100	100	100	100	100	100	100	100	100

Total Rate (TR) and category rates (RC1 and RC2) both in classification, cross-validation and prediction, working on original and corrected spectra. The number of samples incorrectly classified appears in brackets.

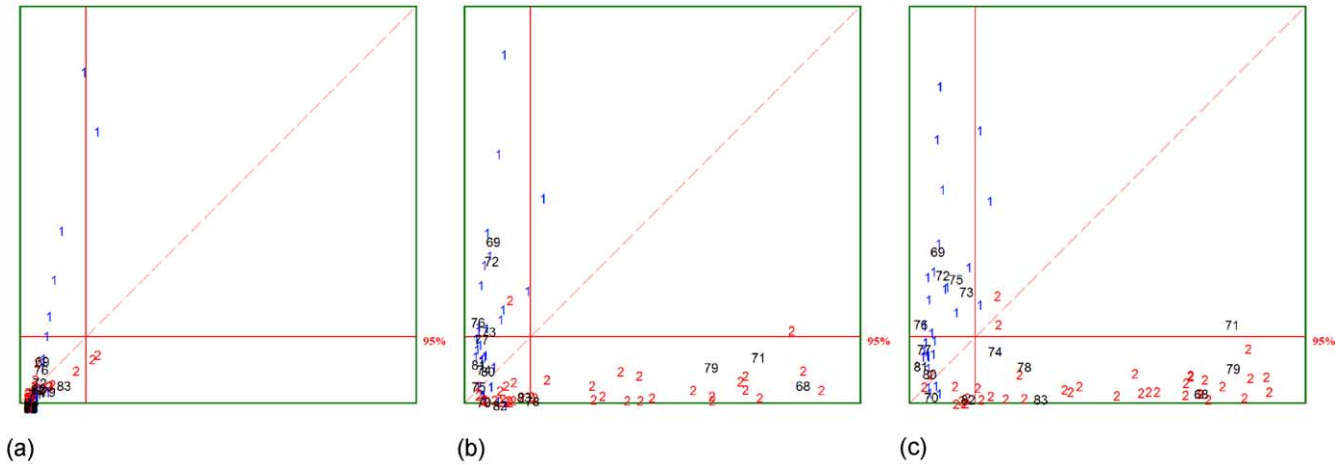

Fig. 1. Coomans plots corresponding to potential functions class-models developed from mean-centred NIR spectra using: (a) 1 PC, (b) 2 PCs and (c) 3 PCs. Calibration set samples are labelled according to their respective category index. Object indexes from 68 to 83 represent samples constituting test set.

The large number of samples plotted in the zone common to the two models demonstrated the low degree of specificity of these models. The addition of a fourth component for the development of the class-model involved a notable improvement in specificity (Fig. 2(a)), which confirmed the selection of 4 PCs as optimal complexity working on unpreprocessed NIR spectra. The Potential Functions Method also enabled us to obtain potentials for contour plots (isolines) taking into account all the objects and categories, which was very useful as a visualising method. Fig. 2(a) displays the isopotential lines plot corresponding to the potential functions model of both *arabica* (circles) and *robusta* (triangles) varieties. As can be seen, class-models relating to both coffee varieties did not appear perfectly separate, showing a high degree of overlapping.

In sight of the numerical and graphical results provided by the optimal classification model developed from raw NIR spectra, several remarks should be made. In spite of the good discrimination ability exhibited by the selected model, which might be considered satisfactory for classification of coffee varieties, the relatively short distance between *arabica* and *robusta* class-models (exhibited in the Coomans plot and confirmed by the clear overlapping between ispotential lines) could reveal some potential problems for classification of extreme samples within each category or blends of varieties in future. This fact serves for giving more sense to the aim pursued in the present study, i.e., trying to improve the final classification model in terms of both specificity and sensitivity in order to make possible a more accurate practical application.

4.2. Chemical variables: selection of a discriminant descriptor

In view of the relative specificity problems, which class-models developed on the basis of original NIR spectra exhibited, we decided to pre-process these spectra by means of a number of orthogonal signal correction methods as an attempt to improve classification performance. In order to apply these treatments, a significant variable had to be selected with a high discriminant power to discriminate between coffee varieties to perform on this basis the corresponding orthogonal correction. Three chemical parameters (caffeine, chlorogenic acid and total acidity) were pre-selected as good inter-variety descriptors, and were used to develop separate potential functions class-models in order to select the most suitable one to carry out the subsequent correction.

As can be observed (Table 2), the most discriminant descriptor was caffeine, providing not only 100% correct classifications in both classification and prediction, but also a great specificity between varieties, visually confirmed in isopotential lines plot (Fig. 3(a)). This result would appear to be logical since caffeine only experiences a very small

Table 2
Percentages of correctly classified samples

Chemical variable	Classification (%)			C.V. (%)			Prediction (%)		
	RC1	RC2	TR	RC1	RC2	TR	RC1	RC2	TR
Caffeine	100	100	100	100	100	100	100	100	100
Chlorogenic acids	55.17 (13)	68.42 (12)	62.69 (25)	48.28 (15)	68.42 (12)	59.70 (17)	71.43 (2)	77.78 (2)	75.00 (4)
Total acidity	93.10 (2)	97.37 (1)	95.52 (3)	96.55 (1)	97.37 (1)	97.01 (2)	100	88.89 (1)	93.75 (1)

Total rate (TR) and category rates (RC1 and RC2) both in classification, cross-validation and prediction, working on chemical variables. The number of samples incorrectly classified appears in brackets.

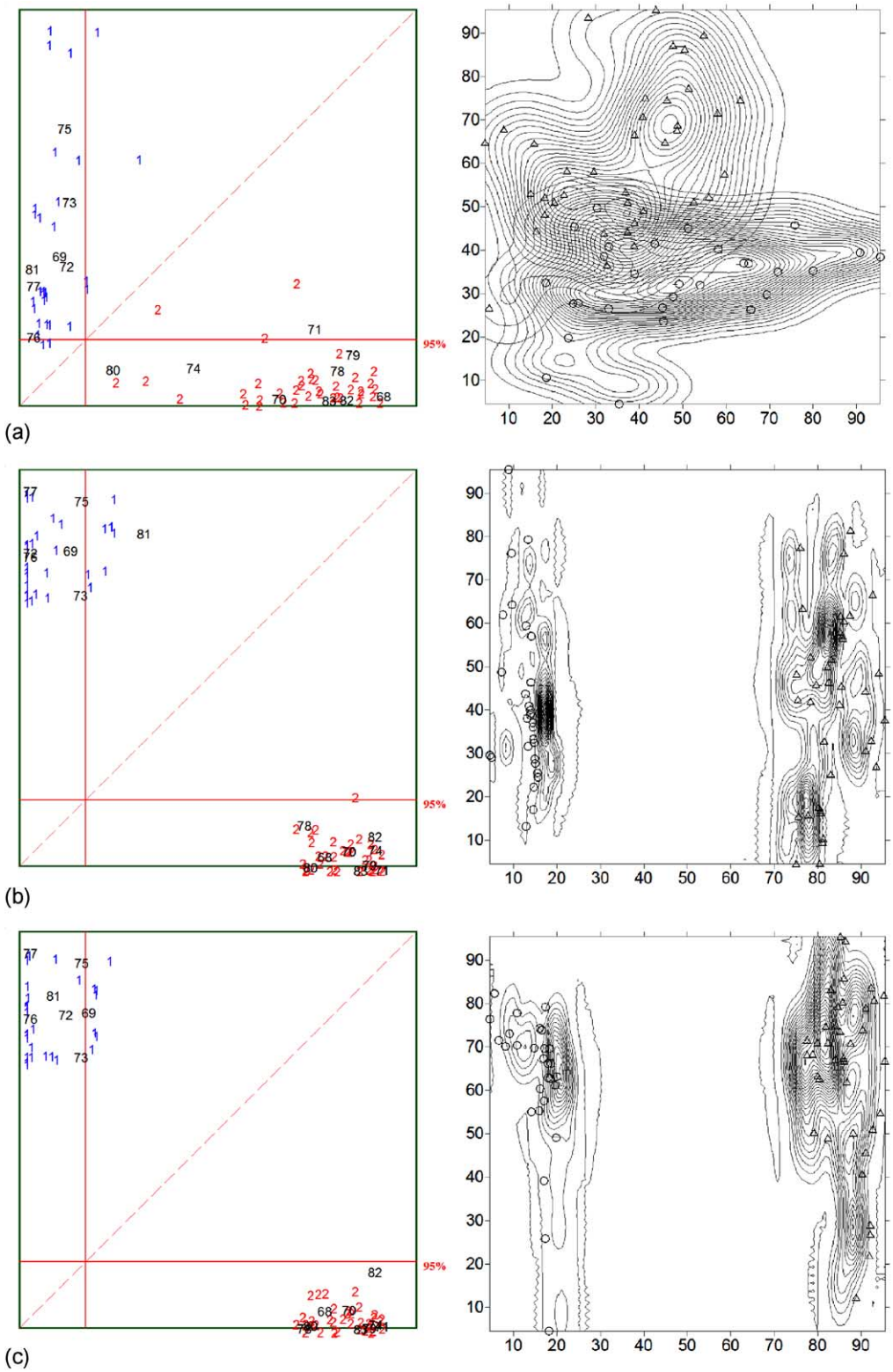

Fig. 2. Coomans and isopotential lines plots corresponding to (a) 4 PCs class-model developed from mean-centred NIR spectra, (b) 1 PC class-model constructed from mean-centred NIR spectra after removing two orthogonal LVs by OSC and (c) 1 PC class-model developed from mean-centred NIR spectra after subtracting three orthogonal PCs by DOSC.

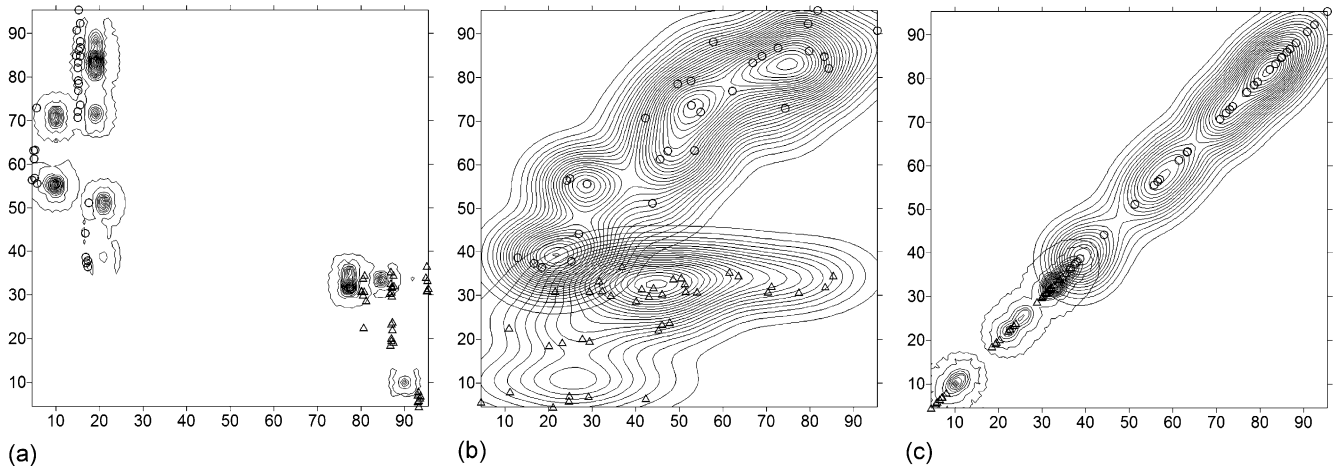

Fig. 3. Isopotential lines plots corresponding to class-model developed by potential functions methods using as discriminant parameters: (a) caffeine content, (b) chlorogenic acid content, and (c) total acidity of roasted coffee samples.

decrease at roasting, maintaining a broad separation interval between *arabica* and *robusta* varieties in comparison with the whole range of caffeine values. In contrast, the transformations occurring at roasting for chlorogenic acids are very significant, and may affect their values in coffee samples to a great or lesser extent, in such way that the gap between varieties can be significantly reduced. This may account for the mediocre results obtained when using chlorogenic acid content as a discriminant variable (Fig. 3(b)). Likewise, the results obtained using total acidity as a discriminant parameter were quite good in terms of correct classifications, but with a lower specificity between varieties (see Fig. 3(c)). It is true that one of the main factors influencing the total acidity of roasted coffee samples is just the variety to which it belongs. However, this is not the only factor because others, such as processing method and roasting degree, can play a decisive role. For this reason a clear separation between class models was not observed; a common zone appeared that corresponded to samples with intermediate acidity from both varieties (overlapping between isopotential lines).

4.3. Corrected NIR spectra

Once selected caffeine content as discriminant variable to be used for correcting NIR spectra, OSC and DOSC were applied varying the number of orthogonal factors (LVs or PCs) to be removed from original data from 1 to 3. Corrected spectra were then used to construct the respective potential functions class-models. The results, expressed as correct classification and prediction rates, are shown in Table 1.

Considering the spectra corrected by OSC, it can be seen that after the removal of two orthogonal latent variables the complexity of the resulting class-model was reduced to only one component, maintaining an excellent discriminant power between categories (100% classification and prediction rates). These numerical results were visually confirmed by Coomans and isopotential lines plots (Fig. 2(b)), showing a high degree of inter-class specificity and a patently clear separation between classes, similar to that obtained when using caffeine content as a discriminating descriptor and considerably improved with regard to the model constructed from raw spectra.

When applying DOSC, the optimal potential functions model was obtained after subtracting three orthogonal components, again providing excellent results in terms of both classification and prediction (always with the minimum complexity) comparable to those obtained when using OSC as pre-processing method or caffeine as a modelling parameter (Fig. 2(c)). As in the case of spectra corrected by the OSC method, the specificity between class-models showed a notable improvement in relation to the model developed from unprocessed spectra.

4.4. Spectral profiles

The scatter effects inherent in near-infrared spectroscopy can be substantial and prompt the expansion of the absorbance interval for individual wavelengths, which can be appreciated at first sight on the spatial distribution of the spectral profiles corresponding to roasted coffee samples along the whole spectral range (Fig. 4(a)). Thus, the physical information contained in spectra often masks significant chemical information as it can be spectral differences between *arabica* and *robusta* coffee varieties.

The main objective of this study focused precisely on minimising these physical effects, which actually have a harmful influence on the quality of the final class-models.

The numerical and graphical results obtained in this study have already demonstrated the efficiency of applying orthogonal signal correction methods (taking into account, in order to perform this correction, a chemical descriptor with a high discriminant power between coffee varieties) prior to the development of a reliable and stable classification model. However, the comparison between the distributions

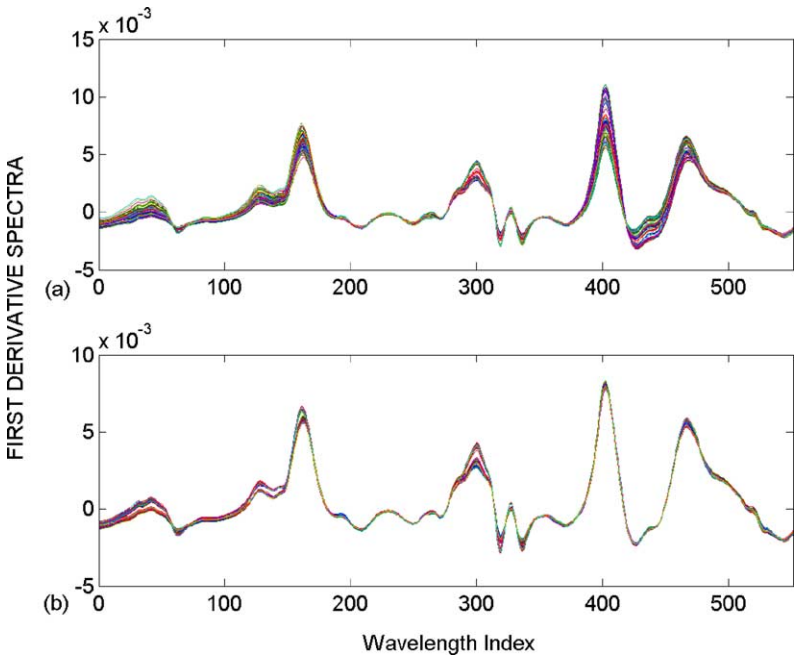

Fig. 4. First derivative spectra of the roasted coffee data set: (a) without any other pre-treatment and (b) after applying OSC.

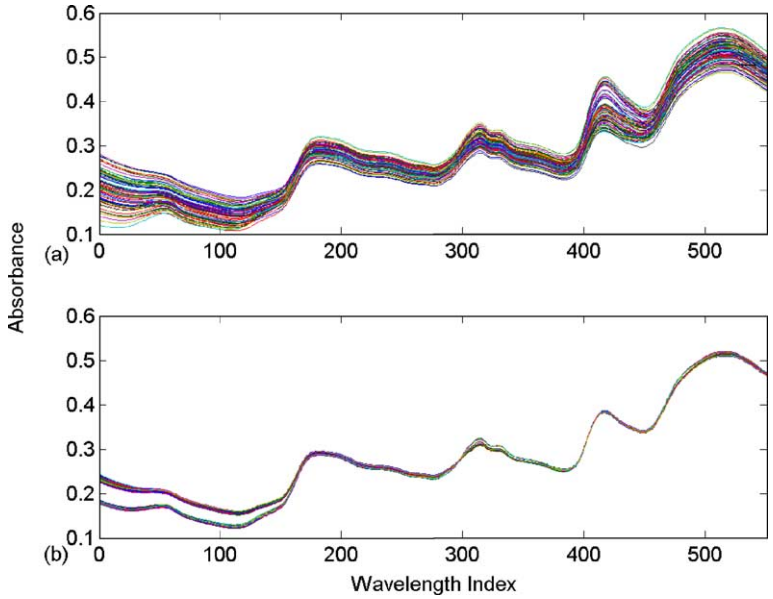

Fig. 5. Spectra of roasted coffee data set: (a) without pre-treatement and (b) after applying DOSC.

of spectral profiles before and after correcting the roasted coffee NIR spectra could still reveal more evidence of the goodness of the methodology proposed here.

The first conclusion that may be drawn by comparing original (Fig. 4(a)), first derivative (Fig. 5(a)), OSC-corrected (Fig. 5(b)) and DOSC-corrected (Fig. 4(b)) spectral profiles was that orthogonal signal correction methods were capable of considerably minimising spectral inter-sample differences, since systematic variation not related to caffeine had been rejected, increasing the overlapping degree among spectra. However, when we looked more closely at the corrected spectral profiles, we could see that the spectral differences that remained were due simply to differences between *arabica* and *robusta* coffee varieties. Thus, in zones where spectra did not overlap, two groups of spectra appeared, each one corresponding to a different coffee variety.

5. Conclusions

This study has shown that the transformation of roasted coffee NIR spectra using an orthogonal signal correction method, taking into account in this correction a chemical response closely related to sample origin, prompted a sub-

stantial improvement in the quality of the constructed classification models in comparison with the model developed on the basis of original spectra. Even though, it is true that classification models developed from unprocessed NIR spectra can be sufficient for providing a satisfactory classification of pure *arabica* and *robusta* coffees, the improved class-models constructed from corrected spectra have proven to be a more efficient and accurate tool for unequivocally discriminate between both varieties, due to the great inter-categories specificity achieved. Moreover, the results obtained using the methodologies proposed in this study were as good as those obtained with the classification model constructed using a chemical descriptor with a high modelling power, in terms of both classification/prediction ability and stability.

On the other hand, it should be clarified that the introduced strategy would not imply any additional effort in relation to analytical determinations in future samples to be characterised. The measurement of caffeine content values was only required for samples which form the calibration set. Once calibration NIR spectra have been corrected by applying the preferred orthogonal signal correction method, the same correction can be directly applied to future sample spectra without need for measuring any reference caffeine content value. In this way, from a practical application standpoint, the classification methodology proposed would only rely on NIR measurements. Therefore, in view of the results obtained, it could be stated that the advantages that the improved classification models offer, in terms of both specificity and reliability, compensate the need for measuring an extra chemical property (caffeine content) in the case of the calibration samples.

The promising results obtained in this study will a similar procedure to be considered in future applications to quantify different blends of varieties in order to identify fraudulent mixtures.

Acknowledgements

The authors thank the Ministry of Science and Technology (Project No. 2FD1997-0491), the Autonomous Government of La Rioja – *Consejería de Educación, Cultura, Juventud y Deportes* (Project No. ACPI2000/08) and the University of La Rioja (Research grant FPI-2001) for their financial support, as well as Professor Michele Forina for providing us with the last version of Parvus package.

References

[1] A. Illy, R. Viani, Espresso coffee: The Chemistry of Quality, Academic Press, London, 1996.
[2] M.J. Martin, F. Pablos, A.G. Gonzalez, Characterization of *arabica* and *robusta* roasted coffee varieties and mixture resolution according to their metal content, Food Chem. 66 (1999) 365–370.
[3] C.P. Bicchi, M.P. Ombretta, G. Pellegrino, A.C. Vanni, Characterization of roasted coffee and coffee beverages by solid phase microextraction-gas chromatography and principal component analysis, J. Agric. Food Chem. 45 (1997) 4680–4686.
[4] M.J. Martin, F. Pablos, A.G. Gonzalez, Discrimination between *arabica* and *robusta* green coffee varieties according to their chemical composition, Talanta 46 (1998) 1259–1264.
[5] M.J. Martin, F. Pablos, A.G. Gonzalez, M.S. Valdenebro, M. Leon-Camacho, Fatty acid profiles as discriminant parameters for coffee varieties differentiation, Talanta 54 (2001) 291–297.
[6] F. Carrera, M. Leon-Camacho, F. Pablos, A.G. Gonzalez, Authentication of green coffee varieties according to their sterolic profile, Anal. Chim. Acta 370 (1998) 131–139.
[7] N. Frega, F. Bocci, G. Lercker, Determinazione del caffè *robusta* nelle miscele commerciali con l'*arabica*, Indust. Aliment. 34 (1995) 705–708.
[8] A.G. Gonzalez, F. Pablos, M.J. Martin, M. Leon-Camacho, M.S. Valdenebro, HPLC análisis of tocopherols and triglycerides in coffee and their use as authentication parameters, Food Chem. 73 (2001) 93–101.
[9] D.G. Evans, C.N.G. Scotter, L.Z. Day, M.N. Hall, Determination of the authenticity of orange juice by discriminant analysis of near infrared spectra: a study of pretreatment and transformation of spectral data, J. Near Infrared Spectrosc. 1 (1993) 33–44.
[10] B.G. Osborne, B. Mertens, M. Thompson, T. Fearn, The authentication of basmati rice using near infrared spectroscopy, J. Near Infrared Spectrosc. 1 (1993) 77–83.
[11] W.J. Krzanowski, Communication: the authentication of basmati rice using near infrared spectroscopy: some further analysis, J. Near Infrared Spectrosc. 3 (1995) 111–117.
[12] P.J. Gemperline, L.D. Webber, F.O. Cox, Raw materials testing using soft independent modelling of class analogy analysis of near-infrared reflectance spectra, Anal. Chem. 61 (1989) 138–144.
[13] N.K. Shah, P.J. Gemperline, Combination of the Mahalanobis distance and residual variance pattern recognition techniques for classification of near-infrared reflectance spectra, Anal. Chem. 62 (1990) 465–470.
[14] P.J. Gemperline, N.R. Boyer, Classification of near-infrared spectra using wavelength distances: comparison to the Mahalanobis distance and residual variance methods, Anal. Chem. 67 (1995) 160–166.
[15] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, Spectral transformation and wavelength selection in near-infrared spectra classification, Anal. Chim. Acta 315 (1995) 243–255.
[16] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Comparison of regularised discriminant analysis, linear discriminant analysis and quadratic discriminant analysis applied to NIR data, Anal. Chim. Acta 329 (1996) 257–265.
[17] W. Wu, D.L. Massart, Regularised nearest neighbour classification method for pattern recognition of near infrared spectra, Anal. Chim. Acta 349 (1997) 253–261.
[18] Q. Guo, W. Wu, D.L. Massart, The robust normal variate transform for pattern recognition with near-infrared data, Anal. Chim. Acta 382 (1999) 87–103.
[19] P.J. de Groot, G.J. Postma, W.J. Melssen, L.M.C. Buydens, Selecting a representative training set for the classification of demolition waste using remote NIR sensing, Anal. Chim. Acta 392 (1999) 67–75.
[20] B.M. Smith, P.J. Gemperline, Wavelength selection and optimization of pattern recognition methods using the genetic algorithm, Anal. Chim. Acta 423 (2000) 167–177.
[21] P.J. de Groot, G.J. Postma, W.J. Melssen, L.M.C. Buydens, Validation of remote, on-line, near-infrared measurements for the classification of demolition waste, Anal. Chim. Acta 453 (2002) 117–124.
[22] M. Blanco, J. Pagès, Classification and quatitation of finishing oils by near infrared spectroscopy, Anal. Chim. Acta 463 (2002) 295–303.
[23] R. De Maesschalck, A. Candolfi, D.L. Massart, S. Heuerding, Decision criteria for soft independent modelling of class analogy applied to near infrared data, Chemom. Intell. Lab. Sys. 47 (1999) 65–77.

[24] U.G. Indahl, N.S. Sahni, B. Kirkhus, T. Næs, Multivariate strategies for classification based on NIR-spectra–with application to mayonnaise, Chemom. Intell. Lab. Sys. 49 (1999) 19–31.
[25] J. Luypaert, S. Heuerding, S. de Jong, D.L. Massart, An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream, J. Pharm. Biomed. Anal. 30 (2002) 453–466.
[26] K.I. Hildrum, T. Isaksson, T. Naes, A. Tandberg, Near infra-red spectroscopy. Bridging the gap between data analysis and NIR applications, Ellis Horwood, Chichester, 1992.
[27] G. Downey, J. Boussion, D. Beauchêne, Authentication of whole and ground coffee beans by near infrared reflectance spectroscopy, J. Near Infrared Spectrosc. 2 (1994) 85–92.
[28] G. Downey, J. Boussion, Authentication of coffee bean variety by near-infrared reflectance spectroscopy of dried extract, J. Sci. Food Agric. 71 (1996) 41–49.
[29] G. Downey, R. Briandet, R.H. Wilson, E.K. Kemsley, Near- and mid-Infrared spectroscopies in food authentication: coffee varietal identification, J. Agric. Food Chem. 45 (1997) 4357–4361.
[30] S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-Infrared spectra, Chemom. Intell. Lab. Sys. 44 (1998) 175–185.
[31] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra, Chemom. Intell. Lab. Sys. 44 (1998) 229–244.
[32] C.A. Andersson, Direct orthogonalization, Chemom. Intell. Lab. Sys. 47 (1999) 51–63.
[33] T. Fearn, On orthogonal signal correction, Chemom. Intell. Lab. Sys. 50 (2000) 47–52.
[34] B.M. Wise, N.B. Gallagher, http://www.eigenvector.com/MATLAB/OSC.html.
[35] J.A. Westerhuis, S. de Jong, A.K. Smilde, Direct orthogonal signal correction, Chemom. Intell. Lab. Sys. 56 (2001) 13–25.
[36] D.J. Hand, Discrimination and Classification, John Wiley&Sons, Chichester, 1981.
[37] D. Coomans, I. Broeckaert, Potential pattern recognition in chemical and medical decision making, Research Studies Press, Letchworth, 1986.
[38] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, J. Chemom. 5 (1991) 435–453.
[39] D.L. Massart, L. Kaufman, The interpretation of analytical chemical data by the use of cluster analysis, Wiley, New York, 1983.
[40] C. Pizarro-Millán, M. Forina, C. Casolino, R. Leardi, Extraction of representative subsets by potential functions method and genetic algorithms, Chemom. Intell. Lab. Sys. 40 (1998) 33–52.