Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 544 (2005) 89–99
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Generalization of OWAVEC method for simultaneous noise suppression, data compression and orthogonal signal correction

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 22 October 2004; received in revised form 24 February 2005; accepted 24 February 2005
Available online 23 March 2005

Abstract

This paper presents modifications to our recently introduced pre-processing method, orthogonal WAVElet correction (OWAVEC), based on the combination of wavelet analysis and an orthogonal correction algorithm, described in detail in a former paper [I. Esteban-Díez, J.M. González-Sáiz, C. Pizarro, OWAVEC: a combination of wavelet analysis and an orthogonalization algorithm as a pre-processing step in multivariate calibration, Anal. Chim. Acta 515 (2004) 31–41], aimed at extending its applicability and at improving its performance; thanks to an additional use of OWAVEC as an effective data compression tool. The OWAVEC method uses the discrete wavelet transform (DWT) to decompose each individual signal into the wavelet domain, and then an orthogonalization algorithm is applied to the obtained wavelet coefficients matrix to remove the information not related to a considered response variable. Later, the corrected wavelet coefficients are ranked by their variance or by their correlation coefficient with the response variable, and the subset providing the most stable and reliable calibration model is finally selected (data compression). The new version of OWAVEC has been applied to two NIR data sets to test its performance. For both regression problems studied, high quality calibration models with very high compression ratios were obtained, providing improved predictive results and a considerably lower overfitting than other orthogonal signal correction methods. The generalized OWAVEC method presented here may be used as a global tool for simultaneous noise suppression, data compression and orthogonal correction of signals.

Keywords: Wavelet analysis; Orthogonalization; Data compression; Multivariate calibration; OWAVEC

1. Introduction

We recently introduced a new pre-processing method, orthogonal WAVElet correction (OWAVEC) [1], which attempted to lump together two important needs in multivariate calibration: signal correction and data compression. The distinctive features of OWAVEC with respect to other commonly applied pre-processing methods (e.g., derivation to different orders [2], standard normal variate (SNV) and multiplicative scatter correction (MSC) [3–5] or different orthogonal signal correction methods [6–13]) are focused on unifying in the same method, the huge potential that wavelet analysis has demonstrated for signal processing (smoothing/de-noising/data compression) [14–21], and the application of a particular orthogonal signal correction algorithm to the data prior to being used as the basis for developing a calibration model to model and predict a certain response, thus pursuing an enhancement of the resulting model predictive ability.

The first version of the OWAVEC method focused on signal correction, i.e., on the removal of information not related to a studied response in the wavelet domain (orthogonalization procedure) and the de-noising processes of the wavelet coefficients matrix. Nevertheless, today, effective data compression methods are increasingly necessary in multivariate calibration mainly due to the redundant nature of many types of collected signals (collinearity problems). The main beneficial effects derived from successful data compression include a reduction of memory storage, an improved generalization ability and robustness of the resulting calibration model and a simpler model representation. The aim of this paper is to describe the additional utility of OWAVEC for achieving effi-

* Corresponding author. Tel.: +34 941299626; fax: +34 941299621.
E-mail address: consuelo.pizarro@dq.unirioja.es (C. Pizarro).

0003-2670/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2005.02.076

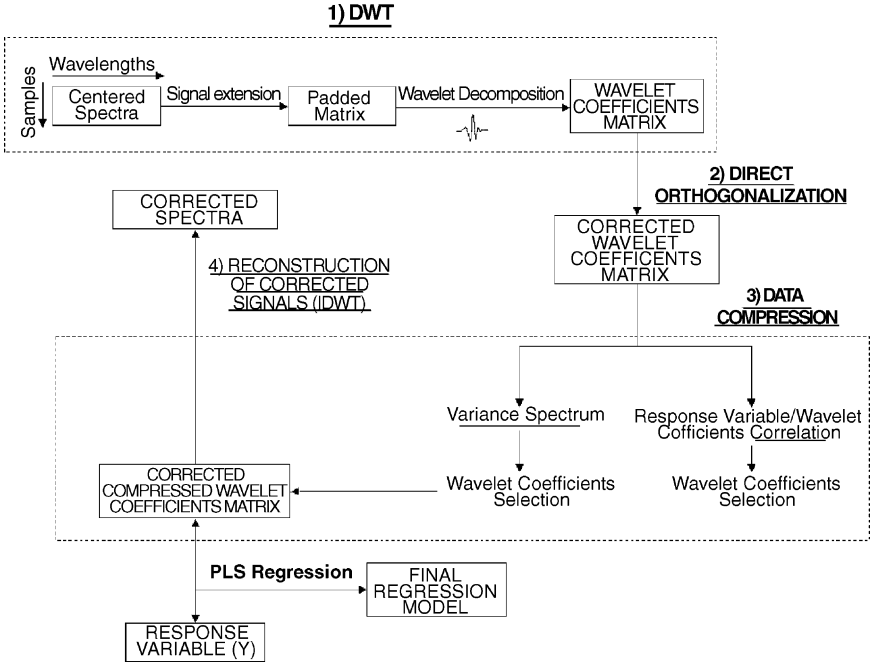

Fig. 1. An overview of the steps involved in the OWAVEC procedure. (1) The previously centered spectra are padded to the nearest length 2^n. Wavelet decomposition: each spectrum is transformed individually using a selected wavelet function. (2) Direct orthogonalization: the wavelet coefficient matrix is corrected by removing a part unrelated to the considered response. (3) Data compression: two different approaches are available; wavelet coefficients selection based on the variance spectrum or on the correlation coefficients vector. (4) Filtered spectra can be reconstructed using the IWT. PLS regression is performed on the basis of the corrected and compressed wavelet coefficients matrix obtained.

cient signal compression, thus improving the predictive ability of the calibration model later developed from the resulting corrected and compressed wavelet coefficients matrix, through an appropriate selection of wavelet coefficients, removing those coefficients that do not hold significant information. Two different approaches are available in the updated version of OWAVEC for achieving such wavelet coefficients compression, depending on the criterion chosen to select the significant coefficients (the variance spectrum of the coefficients matrix or the correlation coefficients computed between response variable and wavelet coefficients).

The generalized method presented here was tested on a set of roasted coffee NIR spectra using both ash content and chlorogenic acid content as response variables and the performance of the finally selected OWAVEC-PLS regression models (considered as the most suitable for modelling and predicting each studied response) was compared not only with that obtained from original data (using mean-centering as the only pre-processing method applied) but also with that provided by other signal correction methods (i.e., OSC [9] and DOSC [13]). Taking into account the huge amount of overlapping information present in NIR spectra, it is easy to understand that much redundant information will be present in NIR data, and that a great part of it can be irrelevant to model and predict a certain response variable. Thus, the application of suitable methods to compress NIR data and properly extract the most useful information is of great importance to produce more robust and reliable regression models. For this reason, we have decided to test OWAVEC performance as an effective compression tool for NIR spectra considering two different regression problems.

2. Methods: new OWAVEC features

Fig. 1 shows a schematic view of all the steps involved in the OWAVEC method as applied in the present study. On the face of it, four main stages can be perfectly differentiated in the procedure carried out by OWAVEC: wavelet decomposition of signals, direct orthogonalization of the wavelet coefficients matrix, corrected wavelet coefficients matrix compression and final reconstruction of corrected signals.

As a first step, OWAVEC uses the discrete wavelet transform (DWT) as wavelet transform technique to decompose each spectrum, through the use of a selected wavelet function in this transformation. In DWT the choice of different wavelet basis functions produces different decompositions of a signal in the resulting wavelet domain. Therefore, the selection of a suitable wavelet function is a crucial parameter, which will depend on the particular shape of the considered signal. In the present study, we used three wavelet functions corresponding to different families of orthogonal wavelet bases and wavelet order – Daubechies-4, Coiflet-2 and Symlet-8 – since they are probably among the most used wavelets and they all have several vanishing moments, which essentially means that polynomials of low degree, i.e., smooth signals, like NIR data, are well compressed. The DWT requires the length of the analysed signal to be dyadic, i.e., 2^n, where n

is an integer. Nevertheless, if this condition is not fulfilled a number of methods exist for extrapolating the signal, such as zero padding, symmetric extension or linear padding. In this report, all signal extrapolations were performed by linear padding, which involves connecting the last and the first values of the signal by means of a straight line. This applied extrapolation method takes advantage of the vanishing moments property inherent in most wavelets in such a way that minimal edge effects are caused. Another key parameter to be optimized when applying DWT for regression purposes is the optimal decomposition level. In order to identify the most suitable wavelet decomposition, depending on the wavelet function selected and the coefficients selection criterion later applied, several decomposition levels were tested (always respecting the maximum allowed wavelet decomposition level associated with a certain wavelet function and the particular extended length of the analysed signals). The choice of the final optimal decomposition level was based on several criteria, including compression rate and the predictive ability of the resulting calibration model, once the OWAVEC procedure had been completed.

The DWT can be seen as a special case of a more general wavelet transform technique: the wavelet packet transform (WPT). In the updated version of OWAVEC introduced here, we preferred to apply the DWT because of its simplicity and computation speed. Nevertheless, the OWAVEC method will continue to be improved and completed in the future, and a modification, which will apply the WPT with a best basis selection criterion is expected to be developed.

Once the wavelet coefficients matrix corresponding to the calibration set is computed, an orthogonalization algorithm is applied to deflate it of information not related to the particular response to be modelled later. Further details on all the steps involved in the orthogonalization procedure implemented in OWAVEC can be found in [1]. Once the orthogonalization procedure has been completed working on calibration set spectra and the spectra in the prediction set have been decomposed by wavelet analysis, the resulting matrix of weights of the original wavelet coefficients in the orthogonal subspace of these coefficients, $\boldsymbol{W}$, can be applied to obtain directly a corrected wavelet coefficients matrix for new data. A critical value to be optimized when performing wavelet coefficients matrix orthogonalization is the value of the tolerance factor used to compute the generalized inverse of the wavelet coefficients matrix. The use of a generalized inverse to compute the matrix of weights of the original wavelet coefficients in the orthogonal subspace of these coefficients implies the loss of the complete orthogonality constraint, which means that the information removed from the wavelet coefficients matrix is not 'necessarily' orthogonal to the response variable. The real deviation degree from orthogonality is precisely given by the tolerance factor value in such a way that if the tolerance is increased, then the information removed from the wavelet coefficients matrix slowly moves away from orthogonality. Allowing some correlation between the removed information and response variable can produce a lower fit in the resulting calibration model, but this is not necessarily true for predictions of a separate test set and prediction errors can decrease notably, due to the fact that non-stable directions in the wavelet coefficients matrix are not used in calculation. For this reason, different values of the tolerance factor have been tested in each case, in order to avoid overfitting, thus selecting that, which enhances the predictive ability of the regression model finally constructed on the basis of the resulting corrected and compressed wavelet coefficients matrix. Logically, the final choice of the most suitable tolerance factor value to be applied will depend on several parameters, including the wavelet basis function used, the particular wavelet decomposition level considered and the compression rate achieved.

Considering the wavelet coefficients matrix obtained after wavelet decomposition, we can realize that as we move further to the right, information about higher and higher frequencies in the original signal is encountered. In the case of NIR spectra, since they are usually smooth signals, it is expected that most low frequency wavelet coefficients (approximation coefficients) will be large, whereas wavelet coefficients representing higher frequencies (detail coefficients) will be close to zero. Therefore, there would be many wavelet coefficients with very small values, which could be regarded as uninformative and could be suppressed without any significant effect on the main information of the original signals.

Besides, if we take into account that once the orthogonalization algorithm is applied, the wavelet coefficients matrix would be deflated of information unrelated to the dependent variable, it is easy to understand that many wavelet coefficients considered originally as 'informative' could significantly modify their final size depending on their particular correlation degree with the response variable. As a result, the amount of non-significant wavelet coefficients, which could be discarded, would be expected to increase.

However, neither the wavelet transforms itself nor the orthogonalization procedure, later applied, provides an effective data compression. The wavelet transformed and corrected signal would have the same length than the original 'extrapolated' signal. Compression will be only achieved after removing the wavelet coefficients that do not contain relevant information. As previously noted, the main modification included in the OWAVEC method with respect to its original version is precisely the added alternative of performing real data compression in the wavelet domain, regardless of whether a specific signal de-noising procedure has been applied. The feature selection (data compression) carried out by OWAVEC is aimed at selecting a reduced subset of significant wavelet coefficients to develop on its basis a calibration model with an enhanced prediction ability, since irrelevant wavelet coefficients contained in the wavelet coefficients matrix will usually worsen the performance of the developed regression model and the precision of the prediction. This potential use of OWAVEC as a compression tool would substantially improve its practical usefulness in multivariate calibration.

The selection of which specific wavelet coefficients will be retained in the data compression step implemented in OWAVEC depends on the criterion adopted to estimate the relevance of each individual wavelet coefficient. Since we are interested in keeping those coefficients, which represent the underlying features responsible for a successful prediction, the value of the correlation coefficient computed between the response variable and each wavelet coefficient is a logical selection criterion to be used in compression. On the other hand, since the systematic variation in the corrected coefficients matrix should remain intact, the amount of variance captured by each coefficient can also be used as a compression criterion. Taking into account that the information not related to the response variable has been already removed in the orthogonalization step, both selection criteria would be expected to be closely related. However, due to the loss of the complete orthogonality constraint, the same results are not obtained when performing wavelet coefficients selection by applying each compression approach. Once the particular criterion to be used for selecting significant wavelet coefficients has been determined, another crucial parameter to be set is the final number of coefficients to be retained. A visual screening of the size distribution of the variance/correlation vector might suggest an approximate number of wavelet coefficients that may be used at the outset. As a first estimation, the positions of the 15 largest variance/correlation coefficients were extracted, and, from these, the effect of adding more coefficients to the PLS model was studied. Thus, after identifying the positions representing the selected number of the largest correlation/variance coefficients, the corresponding columns from the corrected wavelet coefficient matrix were extracted and fed into a corrected compressed wavelet coefficients matrix. The original positions of the extracted coefficients were saved for use in future compression of new spectra (test set). Later, these corrected and compressed wavelet coefficients matrices (obtained for both calibration and test sets), free of information not related to the studied response, would serve as the basis for the subsequent development of a robust and reliable calibration model.

The global result after consecutively applying the first three steps involved in OWAVEC (wavelet decomposition, direct orthogonalization and data compression) is a compressed version of a wavelet coefficients matrix, deflated of information not related to a response of interest, representing a set of input signals. Once this whole process has been iterated for spectra contained in the calibration set, the same procedure can be performed on the signals comprising the external test set, also leading to their corresponding corrected compressed wavelet coefficients matrix.

A clear difference was observed when checking the application order of the diverse steps involved in the new version of OWAVEC with respect to the order followed in the former version of the method; in the first approach, the orthogonalization procedure was applied to the wavelet coefficients matrix after performing the corresponding wavelet denoising, whereas now the orthogonalization algorithm was applied as a prior step to wavelet compression. The reason for this change in the application sequence was justified by the fact that it is more logical to select the final subset of significant coefficients, i.e., to perform the data compression, considering only information closely related to the studied response. Otherwise, if the compression procedure was applied prior to wavelet coefficients matrix orthogonalization, the information content not directly related to the considered response variable could interfere with the compression results and lead to a final wavelet coefficients selection that might be unsuitable for reliably modelling the response. If the aim of applying OWAVEC is to signal de-noising instead of data compression, this aspect is not so crucial and similar results are obtained regardless the order applied.

The final step involved in the procedure carried out by OWAVEC was related to the possibility of reconstructing the corrected spectra into the original domain from the respective corrected and compressed wavelet coefficients matrices by inverse wavelet transform, for both the calibration and evaluation sets. Wavelet transform is a linear transform, and a complete reconstruction of the signals can be always performed from all the computed wavelet coefficients. However, when an orthogonal set of basis functions is used, as in the present case, the signals can still be reconstructed even if some coefficients have been discarded. In this way, the corrected spectra can be reconstructed back into the original domain using the selected wavelet coefficients, to allow for chemical interpretation of the results. In this study, reconstructed NIR spectra, once pre-processed (corrected and compressed) by OWAVEC, were used mainly to evaluate the effect on the spectral profiles after applying the OWAVEC pre-processing method to the original signals.

3. Experimental

3.1. Data sets

The two data sets used to test the efficiency of OWAVEC consisted of 83 NIR spectra of roasted coffee samples from varied origins and varieties (36 *arabica* and 47 *robusta* coffees), processed under different roasting conditions. The first response studied was the ash content expressed as a percentage, within the range from 3.8 to 6.5% (w/w). This data set was randomly split into two separate subsets: a calibration set with 73 samples, and a test set with 10 samples. The main precaution taken when selecting a suitable composition of the external test set was to verify that the contained samples uniformly covered the whole range of response values. The second response analysed was the chlorogenic acid content in percentage terms within the range from 2.15 to 4.50% (w/w), for the samples contained in the data set. In this case, the data set was split into two independent subsets: a calibration set with 67 samples and a test set with 16 samples, properly selected in such a way that they appropriately and uniformly covered the whole response range.

Both data sets had previously been used in several studies carried out in our research group [1,22], which were aimed at testing the ability of certain pre-processing methods (including the first version of OWAVEC in the case of chlorogenic acids content) to correct data and to improve the final quality of the regression models developed thereafter. Therefore, such prior applications might serve as a useful reference to subsequently evaluate and compare the results provided by the regression models constructed after data pre-processing by the new version of OWAVEC presented here.

3.2. Software

The OSC and DOSC [13] routines were implemented in MATLAB®, version 6.5 [24], in the same way as OWAVEC, which was entirely developed using this technical computing language. Moreover, the steps of the OWAVEC procedure that involved the performance of wavelet analysis were carried out using the Wavelet Toolbox version 2 (for use with MATLAB®) [25]. The main objective of using the OSC and DOSC methods to pre-process the data was simply to serve as a reference point for evaluating the results provided by OWAVEC. Thus, the OSC method developed by Wise and Gallagher [9], available in PLS-Toolbox 2.1 for use with MATLAB® [26], was used for the OSC calculations.

For the development of all the calibration models, PLS regression was performed using the PLS program contained in V-PARVUS [23].

3.3. Recording of spectra

Reflectance spectra were obtained directly from untreated samples. Each spectrum was obtained from 32 scans within the wavelength ranges 1100–2500 nm (performed at 4 nm intervals) and 1100–2200 nm (performed at 2 nm intervals) when considering ash content and CGA content as response variables, respectively, with five replicates for each individual sample. Due care was taken to ensure that the same amount of sample was always used to fill up the sample cup. The samples were decompacted between the recordings. An average spectrum was subsequently computed from collected sample spectra replicates.

3.4. Data processing

Several pre-processing methods, such as mean-centering and two orthogonal signal correction methods (OSC and DOSC), were applied to the data sets, testing the quality of the respective calibration models constructed and comparing them to those developed after applying OWAVEC, to evaluate the performance of the new method introduced. After the application of all these tested pre-processing methods to each data set, corresponding PLS calibration models were developed between NIR spectra (or the pre-processed wavelet coefficients matrix in the case of OWAVEC) (calibration set) and the respective responses considered. The data were always centered before use. It is known that orthogonal signal correction methods can produce a notable overfitting when applied to the spectra forming the calibration set. For this reason, although the number of significant latent variables in all the regression models were assessed by cross-validation (all the PLS regression models were built by cross-validation using five deletion groups in all cases), we decided to also validate the actual predictive abilities of the obtained calibration models by testing their performance on an external test set, in order to control and avoid a possible overfitting. Thus, all PLS models tested were subjected to external validation on their corresponding test sets. When OSC was used as a pre-processing method, prior to the mean-centering step, the spectra were transformed into their first derivative spectra, as this preliminary step significantly improved the quality of the model. The root mean square error (RMSE) of the residuals obtained (termed RMSEC in calibration, RMSECV in cross-validation and RMSEP in external prediction) was used to evaluate and compare the respective goodness of the regression models constructed, since this parameter represented an objective measurement of the resulting model predictive ability, being dimensionally comparable to the studied response. Likewise, the percentage of RMSE over the response range was also computed in all cases, thus allowing results corresponding to both analyzed responses, not directly comparable in terms of dimensionality, to be compared.

In order to obtain regression models with as high a predictive ability as possible, the effect of the number of orthogonal LVs (OSC) or orthogonal PCs (DOSC) to be removed from the raw data was studied for both data sets. In all cases the selection of the optimum number of orthogonal factors to be removed was determined by cross-validation with five groups. It is important to bear in mind that to apply cross-validation with either OSC or DOSC, filtering matrices should be recalculated each time cross-validation samples are left out from the calibration set.

Spectral variables were always centered but no scaled prior to OWAVEC application, i.e., the wavelet coefficients matrix retained the original variance natural for NIR spectra. Consequently, the last optional step of corrected spectra reconstruction implemented in OWAVEC would imply a final rescaling operation, so the global shape of spectra will be preserved.

4. Results and discussion

In this section, we will demonstrate the efficiency of the modified, generalized version of the OWAVEC pre-processing method, highlighting its usefulness as a compression tool by comparing the PLS models developed from the wavelet coefficients matrix corrected and compressed by OWAVEC with both the original PLS models (constructed from mean-centered variables), and the PLS models built after applying several orthogonal signal correction methods (OSC and DOSC) to NIR spectra, for both studied responses.

Table 1
Calibration (RMSEC), cross-validation (RMSECV) and prediction (RMSEP) errors, and percentages of explained variance obtained with every data pre-processing method, using the set of roasted coffee samples and the ash content as response variable

	PLS-LVs	RMSEC	%RMSEC	%Explained variance (cal)	RMSECV	%RMSECV	%Explained variance (CV)	RMSEP	%RMSEP
Centering	10	0.2044	7.72	84.02	0.2480	9.37	76.42	0.1889	7.13
OSC (2 O-LVs)	2	0.0814	3.07	97.43	0.0890	3.36	96.96	0.1305	4.93
DOSC (4 O-PCs)	4	0.1240	4.68	94.03	0.1550	5.85	90.79	0.1308	4.94
OWAVEC (DB4/COV)	2	0.0820	3.10	97.39	0.1057	3.99	95.72	0.0821	3.10

Only the model yielding the best performance for each method is shown.

4.1. Prediction of ash content of roasted coffee samples

The external test set used represented 12% of total roasted coffee samples. The same calibration set and external evaluation set compositions were used for all the pre-processing methods applied and calibration models constructed.

Results from the original PLS model, developed from the mean-centered data to model the ash content of roasted coffee samples, are shown in Table 1. The best regression model gave 10 PLS-LVs and predicted ash content with a quite high prediction error (7.13% RMSEP). Additionally, Table 1 also shows the modelling results from the improved PLS models constructed on the basis of corrected NIR spectra pre-processed by OSC and DOSC methods, after selection of the most suitable number of orthogonal factors (LVs or PCs) to be removed in order to later obtain the model yielding the best performance. These PLS models developed after applying both orthogonal signal correction methods provided very similar results in terms of predictive ability (4.93–4.94% RMSEP for OSC and DOSC methods, respectively), and represented a considerable improvement in the resulting regression model quality compared to that of the original in both terms of model reliability and complexity. Nevertheless, note that, although the number of orthogonal latent variables to be removed was selected in order to control overfitting, OSC provided a relatively overfitted solution, since the resulting low calibration error was not accompanied by a similar decrease in the prediction error.

In spite of the unquestionable improvement in the quality of the developed models achieved by applying OSC and DOSC methods, the aim of this study was precisely to try to perform a more efficient signal correction and compression by applying OWAVEC with all the modifications that this involved. Thus, Table 2 summarizes the results from the PLS models based on the OWAVEC corrected and compressed wavelet coefficients matrix, comparing the various wavelet functions tested and both compression approaches applied. Of all these PLS models constructed after applying the updated OWAVEC pre-processing method, the use of Daubechies-4 as wavelet basis function, associated with a four-level decomposition structure, together with the application of the variance spectrum of the corrected coefficients matrix in the compression step, provided not only a notable compression rate (only 57 from a total of 375 wavelet coef-

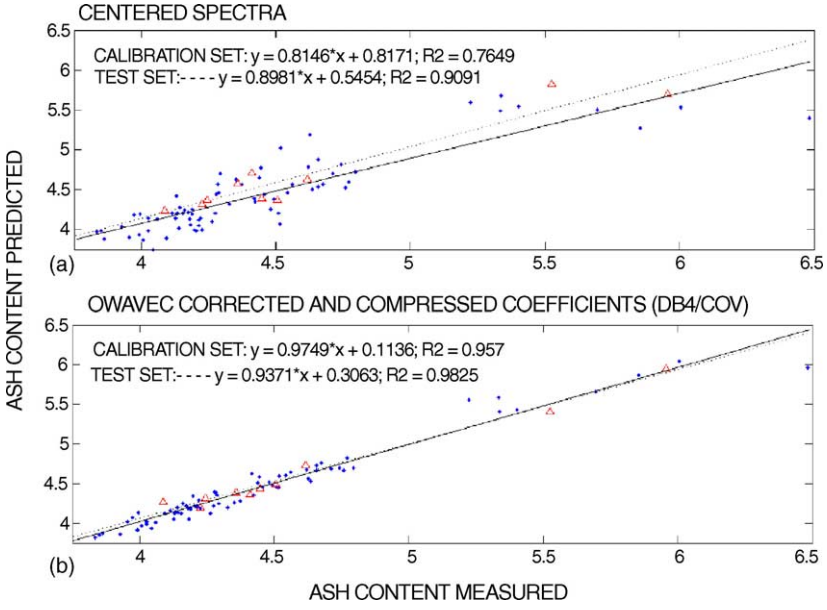

Fig. 2. Measured vs. predicted ash content values: (●) in cross-validation for calibration samples; (△) in external prediction for test samples. (a) PLS model based on original spectra (only mean-centered). (b) Improved PLS model based on the OWAVEC corrected and compressed wavelet coefficients matrix (once selected the most suitable one among all OWAVEC models evaluated).

Table 2
Calibration (RMSEC), cross-validation (RMSECV) and prediction (RMSEP) errors and percentages of explained variance of PLS models developed after the application of OWAVEC from corrected and compressed wavelet coefficients matrix, comparing the various wavelet functions tested and both compression approaches available, using ash content as response variable

Compression criterium	PLS-LVs	RMSEC	%RMSEC	%Explained variance (cal)	RMSECV	%RMSECV	%Explained variance (CV)	RMSEP	%RMSEP	LEVEL	TOL	NCOF	%Zero coeff.
Daubechies-4													
Correlation coefficient	2	0.0862	3.26	97.12	0.1058	4.00	95.71	0.0863	3.26	3	7.5e−4	79	78.53
Covariance spectrum	2	0.0820	3.10	97.39	0.1057	3.99	95.72	0.0821	3.10	4	7.5e−4	57	84.80
Coiflet-2													
Correlation coefficient	2	0.0866	3.27	97.09	0.1030	3.89	95.93	0.0854	3.23	3	7.5e−4	85	77.69
Covariance spectrum	2	0.0891	3.37	96.92	0.1025	3.87	95.97	0.0881	3.33	4	7.5e−4	48	87.76
Symlet-8													
Correlation coefficient	2	0.0926	3.50	96.67	0.1128	4.26	95.12	0.0912	3.44	3	1e−3	32	91.84
Covariance spectrum	2	0.0868	3.28	97.08	0.1036	3.91	95.88	0.0872	3.29	4	7.5e−4	50	87.68

Number of retained coefficients and percentage of zero coefficients are also shown (compression scores). Several OWAVEC parameters such as optimal wavelet decomposition level and tolerance factor used to compute the generalised inverse in the orthogonalization procedure are also included.

ficients were retained, 84.80% zero coefficients) but also a very high predictive power (3.10% RMSEP) with low model complexity (only two PLS-LVs). This particular model (selected from all OWAVEC-based models mainly to serve as a representative example of method performance, since all of them provided very similar results) represented a very notable improvement with respect to the case where no data preprocessing was performed (Fig. 2). However, an even more significant finding may have been the fact that OWAVEC led to improved (more reliable) PLS models in comparison to other orthogonal signal correction methods (see Table 1). The considerable reduction in overfitting (good agreement between results in calibration, cross-validation and external prediction is observed) is particularly remarkable, thus demonstrating the efficiency of the new version of OWAVEC presented here.

A special consideration should be made regarding ash content range of roasted coffee samples. The samples appeared to be clustered into two separate groups according to ash content. This existing gap observed only corresponds to the particular origin of certain samples that results in a considerably higher ash content than most roasted coffee samples. From a practical point of view, in order to develop a suitable calibration model for predicting ash content, these 'extreme' roasted coffee samples must be included in the calibration set for being representative. Nevertheless, this particular composition of the calibration set can lead to draw wrong conclusions when comparing results obtained in calibration and cross-validation. The apparent 'overfitting' observed should be carefully analysed, since it can be mainly due to the potential exclusion of some of these 'extreme' samples in cross-validation cycles.

4.2. Prediction of chlorogenic acids content of roasted coffee samples

The test set used represented 20% of total roasted coffee samples. The same calibration set and test set compositions were used for all the pre-processing methods applied and calibration models developed.

When original data were only mean-centered, a large number of latent variables (nine) had to be included to develop the calibration model in order to obtain a relatively low error in prediction when considering chlorogenic acids (CGA) content of roasted coffee samples as response variable (Table 3). Nevertheless, the fact that errors in both calibration and cross-validation were higher than the error obtained in external prediction demonstrated the lack of reliability of the original PLS model, thus confirming the need for pre-processing data prior to the development of a suitable model.

The results corresponding to the PLS models developed from NIR spectra after applying both OSC and DOSC pre-processing methods, once the most suitable number of orthogonal factors to be removed had been selected, are also included in Table 3. As can be observed, the final complexity of the respective regression models obtained af-

Table 3
Calibration (RMSEC), cross-validation (RMSECV) and prediction (RMSEP) errors and percentages of explained variance obtained with every data preprocessing method, using the set of roasted coffee samples and the CGA content as response variable

	PLS-LVs	RMSEC	%RMSEC	%Explained variance (cal)	RMSECV	%RMSECV	%Explained variance (CV)	RMSEP	%RMSEP
Centering	9	0.1299	5.61	94.27	0.1633	7.06	90.96	0.1118	4.83
OSC (2 O-LVs)	2	0.0674	2.91	98.46	0.0714	3.09	98.27	0.1247	5.39
DOSC (3 O-PCs)	1	0.0819	3.54	97.72	0.0825	3.56	97.70	0.1162	5.02
OWAVEC (DB4/CORR)	2	0.0716	3.09	98.26	0.0795	3.44	97.86	0.0707	3.05

Only the model yielding the best performance for each method is shown.

ter applying OSC and DOSC was significantly reduced. This fact, together with the improved calibration and cross-validation results, suggested a simplification and improvement with respect to the model constructed with unprocessed spectra, although a notable overfitting was observed in the resulting OSC and DOSC based models. The apparent decrease observed in terms of the predictive ability obtained in both cases was not too important considering the abovementioned unreliability displayed by the original PLS model.

Table 4 shows the results obtained from the various PLS models developed from the OWAVEC corrected and compressed wavelet coefficients matrix, both varying the wavelet function used and the compression criterion applied. Although either of these OWAVEC-PLS models could be directly considered as an improved solution to the problem studied here in terms of model reliability and robustness, the selection of Daubechies-4 as a wavelet basis function, associated with a decomposition structure of four levels, together with the use of the vector containing the correlation coefficients computed between the response variable (CGA content) and the individual wavelet coefficients in the compression procedure, resulted in an impressive compression rate (only 15 from a total of 579 wavelet coefficients were retained; 97.41% zero coefficients), also providing a very low error in external prediction (3.05% RMSEP) with no significant signs of overfitting (3.09% RMSEC versus 3.44% RMSECV). Therefore, the application of OWAVEC as a preprocessing method not only led to an important improvement in the final quality of the calibration model later developed with respect to the original PLS model (Fig. 3), but also with regard to the results obtained by other orthogonal signal correction methods (see Table 3), particularly in terms of model reliability.

4.3. Spectral profiles

The scatter effects that are inherently in near-infrared reflectance spectroscopy can be significant and produce an expansion of the absorbance interval for the individual wavelengths. This effect can be seen in Figs. 4a and 5a, which show the distribution of the original spectra corresponding to the roasted coffee samples along the spectral range analysed for each studied response.

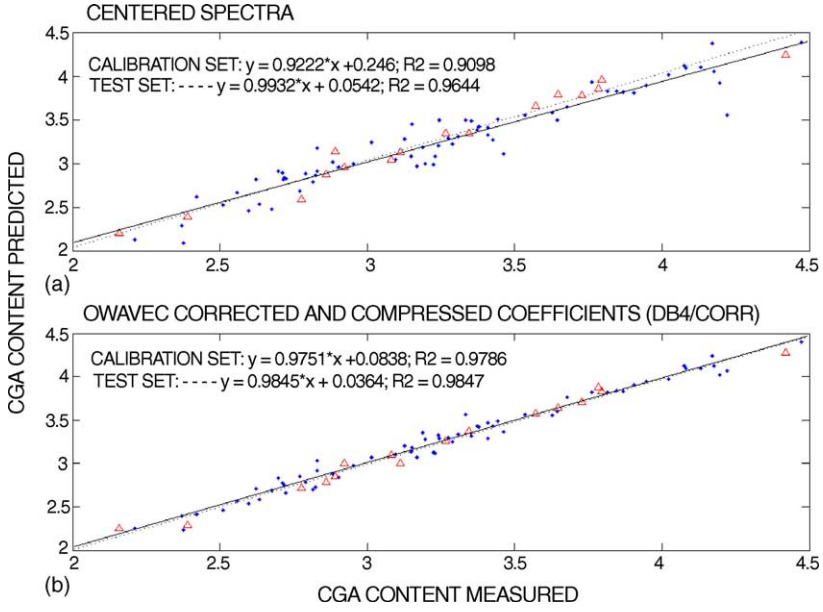

Fig. 3. Measured vs. predicted CGA content values: (●) in cross-validation for calibration samples; (△) in external prediction for test samples. (a) PLS model based on original spectra (only mean-centered). (b) Improved PLS model based on the OWAVEC corrected and compressed wavelet coefficients matrix (once selected the most suitable one among all OWAVEC models evaluated).

Table 4
Calibration (RMSEC), cross-validation (RMSECV) and prediction (RMSEP) errors and percentages of explained variance of PLS models developed after the application of OWAVEC from corrected and compressed wavelet coefficients matrix, comparing the various wavelet functions tested and both compression approaches available, using CGA content as response variable

Compression criterium	PLS-LVs	RMSEC	%RMSEC	%Explained variance (cal)	RMSECV	%RMSECV	%Explained variance (CV)	RMSEP	%RMSEP	LEVEL	TOL	NCOF	%Zero coeff.
Daubechies-4													
Correlation coefficient	2	0.0716	3.09	98.26	0.0795	3.44	97.86	0.0707	3.05	4	1e−3	15	97.41
Covariance spectrum	2	0.0684	2.96	98.41	0.0761	3.29	98.04	0.0742	3.21	2	9e−4	24	95.75
Coiflet-2													
Correlation coefficient	2	0.0694	3.00	98.36	0.0806	3.48	97.80	0.0791	3.42	4	9e−4	14	97.64
Covariance spectrum	2	0.0685	2.96	98.41	0.0759	3.28	98.05	0.0752	3.25	2	9e−4	23	95.99
Symlet-8													
Correlation coefficient	2	0.0739	3.19	98.15	0.0824	3.56	97.70	0.0844	3.65	4	1e−3	15	97.54
Covariance spectrum	2	0.0688	2.97	98.39	0.0764	3.30	98.02	0.0755	3.26	2	9e−4	24	95.87

Number of retained coefficients and percentage of zero coefficients are also shown (compression scores). Several OWAVEC parameters such as optimal wavelet decomposition level and tolerance factor used to compute the generalised inverse in the orthogonalization procedure are also included.

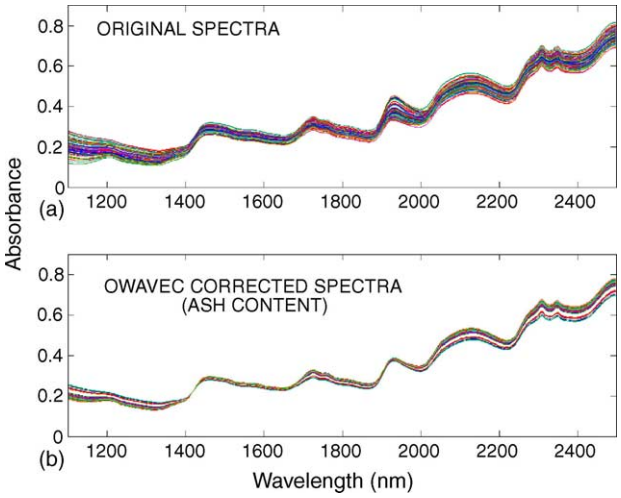

Fig. 4. NIR spectral profiles corresponding to all the roasted coffee samples in the data set (including both calibration and test samples). (a) Raw spectra; (b) reconstructed spectra after OWAVEC application, considering ash content as response variable and the corrected and compressed wavelet coefficients matrix yielding better results.

One of the main objective that OWAVEC attempts to accomplish is precisely the minimization of these scatter effects, by removing systematic variations not related to the response to be modelled and predicted that have a harmful influence on the quality of the resulting calibration model.

In this section we have attempted to demonstrate the usefulness and efficiency of OWAVEC for correcting and compressing data prior to the development of a reliable calibration model, mainly by expounding the numerical results. However, a graphical analysis of the spectral profiles, once the data have been pre-processed by OWAVEC, could also be interesting for evaluating the performance of the method and its effect on the original signals; and for this reason, we have addition-

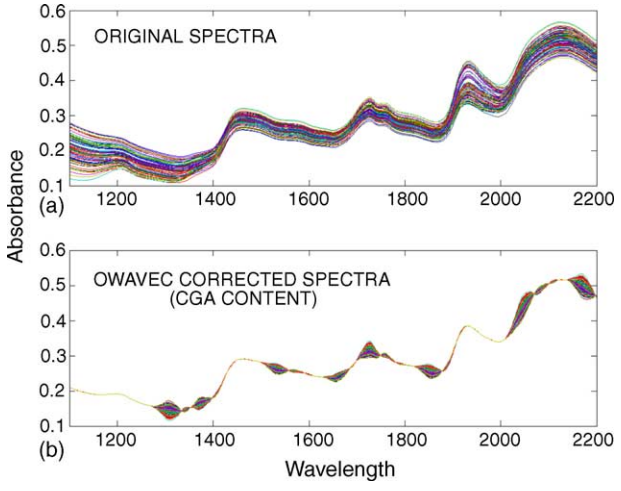

Fig. 5. NIR spectral profiles corresponding to all the roasted coffee samples in the data set (including both calibration and test samples). (a) Raw spectra; (b) reconstructed spectra after OWAVEC application, considering CGA content as response variable and the corrected and compressed wavelet coefficients matrix yielding better results.

ally reconstructed the corrected and compressed spectra in the original domain. Figs. 4b and 5b show the spectral profiles corresponding to all roasted coffee samples after application of OWAVEC, taking into account ash content and CGA content as response variables, respectively. The spectral profiles obtained after applying OWAVEC for ash content were particularly interesting, since despite the high degree of overlapping among spectra, a small set of 10 samples appeared clearly separate from the rest at some wavelength regions. The reason for this behaviour can be easily explained bearing in mind that only information related to ash content was contained in the corrected spectra and, as previously noted, that the ash content values in these particular samples were considerably higher than in the rest. Likewise, when considering CGA content as response variable to be modelled, the great compression rate achieved with the application of OWAVEC resulted not only in the minimization of spectral differences in certain spectral ranges to a great extent compared to raw spectra, but also in considering certain broad wavelength regions as being completely 'empty of valuable information'.

5. Conclusions

This study introduced an updated version of OWAVEC method incorporating certain modifications aimed at turning it into a more versatile pre-processing method. Without forgetting the potential use of OWAVEC for signal de-noising, the present study focused mainly on describing the additional usefulness of OWAVEC as an effective compression tool for use in multivariate calibration, by implementing a proper selection procedure of a limited number of significant coefficients representing the underlying features responsible for a reliable prediction of a certain response variable.

Several studies in which OWAVEC was used in regression analysis for signal correction and compression, were performed yielding very promising results. Thus, the application of the modified version of OWAVEC for pre-processing the two data sets used here to test its efficiency, comprising NIR spectra of roasted coffee samples, resulted in extremely high compression rates (84.80 and 97.41% zero coefficients using ash content and CGA content as response variables, respectively) without any loss of predictive power. Moreover, the orthogonalization algorithm implemented in OWAVEC was shown to successfully enhance the predictive ability of the calibration models subsequently constructed on the basis of the resulting corrected and compressed wavelet coefficients matrices. The regression models developed for modelling and predicting both ash and CGA contents of roasted coffee samples after pre-processing NIR spectra by OWAVEC were considerably better compared to the models developed from raw spectra (in terms of both predictive ability and model complexity), and they also provided notably improved results in comparison to the other orthogonal signal correction methods tested.

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 the Parvus package.

References

[1] I. Esteban-Díez, J.M. González-Sáiz, C. Pizarro, OWAVEC: a combination of wavelet analysis and an orthogonalization algorithm as a pre-processing step in multivariate calibration, Anal. Chim. Acta 515 (2004) 31–41.
[2] W.F. McClure, NIR News 5 (1) (1994) 12.
[3] T. Isaksson, T. Næs, The effect of multiplicative scatter correction and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[4] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and detrending of near infrared diffuse reflectance, Appl. Spectrosc. 43 (1989) 772–777.
[5] T. Næs, T. Isaksson, B. Kowalski, Locally weighted regression and scatter-correction for near-infrared reflectance data, Anal. Chem. 62 (1990) 664–673.
[6] S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[7] C.A. Andersson, Direct orthogonalization, Chemom. Intell. Lab. Syst. 47 (1999) 51–63.
[8] T. Fearn, On orthogonal signal correction, Chemom. Intell. Lab. Syst. 50 (2000) 47–52.
[9] B.M. Wise, N.B. Gallagher, http://www.eigenvector.com/MATLAB/OSC.html.
[10] J.A. Fernández Pierna, D.L. Massart, O.E. de Noord, Ph. Ricoux, Direct orthogonalization: some case studies, Chemom. Intell. Lab. Syst. 55 (2001) 101–108.
[11] S. Wold, J. Trygg, A. Berglund, H. Antii, Some recent developments in PLS modelling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150.
[12] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. 16 (2002) 119–128.
[13] J.A. Westerhuis, S. de Jong, A.K. Smilde, Direct orthogonal signal correction, Chemom. Intell. Lab. Syst. 56 (2001) 13–25.
[14] K. Jetter, U. Depczynski, K. Molt, A. Niemöller, Principles and applications of wavelet transformation to chemometrics, Anal. Chim. Acta 420 (2000) 169–180.
[15] B. Walczak (Ed.), Wavelets in Chemistry, Elsevier, The Netherlands, 2000.
[16] A.K.M. Leung, F.T. Chau, J.B. Gao, A review on applications of wavelet transform techniques in chemical analysis: 1997, Chemom. Intell. Lab. Syst. 43 (1998) 165–184.
[17] B. Walczak, D.L. Massart, Wavelets—something for analytical chemistry? Trends Anal. Chem. 16 (1997) 451–463.
[18] B.K. Alsberg, D.B. Kell, J.J. Rowland, M.K. Winson, A.M. Woodward, Wavelet de-noising of infrared spectra, Analyst 122 (7) (1997) 645–652.
[19] J. Trygg, S. Wold, PLS regression on wavelet compressed NIR spectra, Chemom. Intell. Lab. Syst. 42 (1998) 209–220.
[20] B. Walczak, D.L. Massart, Noise suppression and signal compression using the wavelet packet transform, Chemom. Intell. Lab. Syst. 36 (1997) 81–94.

[21] V.J. Barclay, R.F. Bonner, I.P. Hamilton, Application of wavelet transforms to experimental spectra: smoothing, de-noising, and data set compression, Anal. Chem. 69 (1997) 78–90.
[22] C. Pizarro, I. Esteban-Díez, A.J. Nistal, J.M. González-Sáiz, 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 509 (2004) 217–227.
[23] M. Forina, S. Lanteri, C. Armanino, C. Cerrato Oliveros, C. Casolino, V-PARVUS 2004, an extendable package of programs for explorative data analysis, classification and regression analysis, Dip. Chimica e Tecnologie Farmaceutiche, University of Genova.
[24] MATLAB® 6.5, The MathWorks, Natick, USA 2002.
[25] M. Misiti, Y. Misiti, G. Oppenheim, J.M. Poggi, Wavelet Toolbox for Use with MATLAB®, The MathWorks, Natick, USA, 2000.
[26] B.M. Wise, N.B. Gallagher, PLS Toolbox 2.1, Eigenvector Research Inc., USA, 1998.