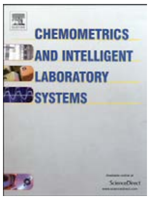
Chemometrics and Intelligent Laboratory Systems 102 (2010) 100–109

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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


CHEMOMETRICS
AND INTELLIGENT
LABORATORY
SYSTEMS

Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice A test field for variable selection methods

Natalia Sorol [a], Eleuterio Arancibia [b], Santiago A. Bortolato [c], Alejandro C. Olivieri [c,*]

[a] *Estación Experimental Agroindustrial Obispo Colombres, Av. William Cross 3150 (T4101XAC) Las Talitas, Tucumán, Argentina*
[b] *INQUINOA-CONICET, Departamento de Ingeniería de Procesos y Gestión Industrial, Facultad de Ciencias Exactas y Tecnología, Universidad Nacional de Tucumán, Avda. Independencia 1800, San Miguel de Tucumán (4000), Argentina*
[c] *Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario e Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (2000), Argentina*

ARTICLE INFO

Article history:
Received 9 November 2009
Received in revised form 26 March 2010
Accepted 9 April 2010
Available online 14 May 2010

Keywords:
Partial least-squares
Variable selection
Vis–NIR spectroscopy
Brix degrees
Sugar cane juice analysis

ABSTRACT

Several variable selection algorithms were applied in order to sort informative wavelengths for building a partial least-squares (PLS) model relating visible/near infrared spectra to Brix degrees in samples of sugar cane juice. Two types of selection methods were explored. A first group was based on the PLS regression coefficients, such as the selection of coefficients significantly larger than their uncertainties, the estimation of the variable importance in projection (VIP), and uninformative variable elimination (UVE). The second group involves minimum error searches conducted through interval PLS (i-PLS), variable-size moving-window (VS-MW), genetic algorithms (GA) and particle swarm optimization (PSO). The best results were obtained using the latter two methodologies, both based on applications of natural computation. The results furnished by inspection of the spectrum of regression coefficients may be dangerous, in general, for selecting informative variables. This important fact has been confirmed by analysis of a set of simulated data mimicking the experimental sugar cane juice spectra.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In multivariate near infrared (NIR) spectroscopic analysis, one of the main objectives is the prediction of a certain property (e.g., octane number in gasolines, glucose content in blood, oil concentration in seeds, Brix degrees in sugar cane, etc.) from the spectrum of a given sample. For this purpose, a multivariate model is built which mathematically relates the spectra for a group of reference samples with their known property values. If the spectra are collected in a matrix **X** (size $I \times J$, where I is the number of samples and J the number of wavelengths) and the reference property values in the vector **y** (size $I \times 1$), the usual multivariate model is expressed by:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where **b** is known as the vector of regression coefficients (size $J \times 1$) and **e** collects the model errors. Eq. (1) represents a calibration model known as 'inverse', because it expresses Beer's law in an inverse fashion. It relates the property of interest to the information contained in **X**, which consists of a superposition of many different signals. The latter ones may or may not be connected to the target property **y**.

The vector **b** can be estimated in various ways, one of the most popular being partial least-squares regression (PLS, see below) [1]. Once estimated, **b** can be employed to predict the property of a new sample (y) through:

$$y = \mathbf{x}^{\mathrm{T}}\mathbf{b} \tag{2}$$

where **x** is the spectrum for a test specimen (size $J \times 1$) and the subscript 'T' implies transposition.

Variable selection is usually performed in PLS analysis, implying that a limited number of signals is employed for building the multivariate model, discarding the remaining ones. The main purpose of this selection is the building of models with spectral data having a richer information content regarding the analyte or property of interest, as well as less spectral overlapping with potential interferences [2]. Both theory and practice shows that improved analytical performance is achieved in PLS upon variable selection, hence the interest in this chemometric activity through the years [3–8]. Other multivariate techniques such as principal component analysis, ridge regression, etc. may also benefit from variable selection, either for quantitative or classification purposes. In this sense, all the selection procedures to be discussed in the present paper are applicable to the latter methods, and therefore they are not restricted to PLS regression.

Two general types of variable selection method exist, namely: 1) the inspection of regression coefficients or latent variables for the full

* Corresponding author. Tel./fax: +54 341 4372704.
E-mail address: olivieri@iquir-conicet.gov.ar (A.C. Olivieri).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.04.009

spectral PLS model, and 2) the search for sensor ranges where the prediction error is minimum. The former are appealing because they are considerably faster than the latter. Perhaps the simplest one, which is still being advocated by many authors, is based on visual inspection of the vector of regression coefficients [2,9]. Regions with significant values of the regression vector (either positive or negative) are suggested to be included in the model, while spectral windows where the regression vector is noisy or low-intensity are discarded. Several modifications of this simple strategy are known, including: 1) the setting of critical limits to the values of the regression coefficients at each wavelength based on uncertainty considerations [10], 2) the elimination of uniformative variables (UVE) based on the addition of noise [11], and 3) the concept of variable importance in projection (VIP) [12]. These more elaborate versions of the inspection of regression coefficients intend to provide automatic, analyst-independent, variable selection alternatives.

It should be noticed that the intuitive power of regression coefficients to aid in variable selection has been challenged [13–15]. Recently, Brown and Green have shown in detail which are the dangers of performing variable selection based on regression coefficients, concluding that they strongly depend on the specific data under analysis and their noise structure [16]. Regions where the regression coefficients are low or near zero may actually correspond to spectral windows where the analyte of interest is highly responsive. Conversely, regions with significant values of the regression coefficients may arise from spectral ranges where the analyte does not respond [16]. Due to the extremely complex behaviour of the vector of regression coefficients in inverse calibration, the relevant conclusion of this recent work is that “direct comparison of the regression vector to the pure-component spectrum of the analyte is simply not meaningful in either the negative or the affirmative” [16].

Other alternatives exist for selecting variables for PLS regression which avoid carrying out extensive searches. Lindgren et al. described, for example, interactive variable selection (IVS)-PLS, in which elements of the PLS latent vectors are selectively modified during the modelling phase, under the guide of an estimate for predictive quality [17]. Teófilo et al. have recently discussed the inspection of several vectors in search for informative spectral regions, such as: 1) the correlation vector between variables and concentrations, 2) the vector of residuals of the reconstruction of the original variables, 3) the vector of net analyte signals, or 4) the signal-to-noise vector [18]. However, the consideration of the regression coefficients seemed to perform better than any of the remaining alternatives [18].

Another important group of variable selection tools includes the search for sensor ranges where the predictive indicators are optimum. They assume that sensor ranges with improved analytical ability are related to spectral windows with maximum information content regarding the analyte of interest. The simplest of these methods is the so-called interval-PLS (i-PLS), where a multivariate model is constructed in each of the spectral windows provided by a moving-window strategy with a fixed window size [19]. The minimum prediction error corresponds, ideally, to the best spectral region for regression. A somewhat more elaborate method involves variable-size moving-window: the error indicator is now a matrix with two indexes, the first sensor and the sensor width [20,21]. This method allows one to find regions with a width which can be considerably larger than the minimum spectral window. However, it is unable to find regions which combine separate spectral sub-regions. An interesting derivation of i-PLS and window search has been recently described [22].

Since a fully comprehensive search may be prohibitively time consuming when the full spectral range includes a large number of sensors, such as those employed in visible/near infrared (Vis–NIR) spectroscopy, alternative strategies have been proposed, based on algorithms for global searches inspired in natural processes. Genetic algorithms (GA) [23–27], particle swarm optimization (PSO) [28], simulated annealing [29] and ant colony optimization (ACO) [30] are pertinent examples. A potential problem with these methods is the time required to complete the calculations, especially when leave-one-out cross-validation is employed at each algorithmic step to set the optimum number of PLS latent variables in each studied spectral region. Alternatives have been proposed based on penalized errors or generalized cross-validation errors [28]. However, they easily tend to overfit the data, pointing to excessive number of PLS components. A good alternative is the random division of the calibration set, with judicious selection of the number of latent variables by analyzing the prediction error on a monitoring sub-set of samples. Repeated calculations with different random divisions of the calibration set make the method as close as possible to full cross-validation.

In the present report, several of the above variable selection methodologies were applied to a simulated data set and to the determination of Brix in sugar cane juice from Vis/NIR data. The best results were obtained by wavelength searches conducted with the GA version described below. The connections with the method of inspection of the regression coefficients are also discussed. Brix analysis is a relevant industrial parameter characterizing sugar cane juice, which is conveniently measured by combining Vis/NIR spectroscopy and multivariate calibration [31–36]. In this context, Lima et al. have recently reported the use of PLS pruning based on the Hessian matrix of errors for discrete wavelength selection [37], reaching an average error of 0.4 units in the analysis of Brix. However, this approach employs a very limited number of wavelengths (eleven in the latter case), which may compromise the sensitivity of the determination. We report on wavelength ranges including a significantly larger number of wavelengths, including sensitivity estimates in each of the analyzed cases.

2. Theory

2.1. PLS regression

In PLS regression analysis, a model is constructed relating the calibration spectral data matrix $\mathbf{X}$ (size $J \times I$, J is the number of sensors, I the number of samples) with the vector of calibration concentrations of the analyte or property of interest $\mathbf{y}$ (size $I \times 1$). The basic underlying assumption of the PLS model is that the studied system is driven by a small number of latent variables, which are linear combinations of the observed variables, and are defined in order to maximize the covariance of the signal matrix $\mathbf{X}$ to the vector of properties $\mathbf{y}$. Particularly important are the latent variables known as scores, because they replace the observed variables in Eq. (1), in order to predict the $\mathbf{y}$ values by the following inverse least-squares model:

$$\mathbf{y} = \mathbf{T}\mathbf{v} + \mathbf{e} \tag{3}$$

where $\mathbf{v}$ is the vector of regression coefficients in the latent space. If the number of significant latent variables is A, then the sizes of $\mathbf{T}$ and $\mathbf{v}$ are $I \times A$ and $A \times 1$ respectively. The matrix $\mathbf{T}$ is related to the original data matrix $\mathbf{X}$ through additional latent variables provided by the PLS model: the matrix $\mathbf{P}$ ($J \times A$) of loading vectors, and the matrix $\mathbf{W}$ ($J \times A$) of weight loading vectors:

$$\mathbf{T} = \mathbf{X}\mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \tag{4}$$

A new sample vector of signals ($\mathbf{x}$) is first projected onto the latent variables, producing a score vector $\mathbf{t}$ ($A \times 1$):

$$\mathbf{t} = \left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{x} \tag{5}$$

which renders the predicted concentration (or property) y through:

$$y = \mathbf{t}^{\mathrm{T}}\mathbf{v} \tag{6}$$

Prediction can also proceed from the original spectrum **x** by Eq. (2), where **b** is given in terms of the latent variables as:

$$\mathbf{b} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{v} \tag{7}$$

The sensitivity of the analysis is provided by the length of the vector of regression coefficients, i.e. [38,39]:

$$SEN = 1 / \|\mathbf{b}\| \tag{8}$$

where || || indicates the Euclidean norm. The SEN is a relevant figure of merit which directly affects the uncertainty in predicted values [38,39].

2.2. Inspection of regression coefficients

One simple way to select relevant spectral region from the vector **b** of Eq. (7) is to compare the individual values of b_j at each of the J wavelengths with its associated uncertainty $s(b_j)$. Coefficient uncertainties are easily estimated in principal component regression (PCR) analysis [10], but this is not straightforward in PLS [40]. Recently, Faber has analyzed several methods for estimating the variance in PLS regression coefficients [41], concluding that the following approximate expressions are useful for this purpose:

$$s\left(b_j\right) = \left[\mathbf{V}(\mathbf{b})_{jj}\right]^{1/2} \tag{9}$$

$$\mathbf{V}(\mathbf{b}) = MSEC \times \left(\mathbf{R}\mathbf{R}^{\mathrm{T}}\right) \tag{10}$$

where **V**(**b**) is the covariance matrix of the vector **b**, MSEC is the mean squared error of the concentration or property of interest (i.e., $||y_{\mathrm{nom}} - y_{\mathrm{pred}}||^2/I$, where y_{nom} and y_{pred} are the nominal and predicted property values), and **R** is a $J \times A$ matrix of weights relating the calibration matrix **X** and the matrix scores **T** ($\mathbf{T}=\mathbf{X}^{\mathrm{T}}\mathbf{R}$) in the PLS formalism known as SIMPLS [42]. Once **b** and s(**b**) are available, data points are simply selected at wavelengths where the modulus $|b_j|$ is larger than $3 \times s(b_j)$. Eqs. (9) and (10) constitute one of the several available approximations for obtaining the variance in PLS regression coefficients [43].

2.3. Uninformative variable elimination

This method of variable selection intends to set an alternative critical limit to the value of the regression coefficients [11]. First the (unscaled) calibration data matrix **X** is augmented with a matrix of the same size containing Gaussian random noise (Ref. [11] recommends that the standard deviation for this noise should be very small, even smaller than the estimated instrumental noise, such as for example 1×10^{-10}), so that the augmented matrix $\mathbf{X}_{\mathrm{augm}}$ has a size $2J \times I$. Samples are then left out from $\mathbf{X}_{\mathrm{augm}}$ one at a time, and a regression vector $\mathbf{b}_{\mathrm{augm}}$ $(2J \times 1)$ is estimated from a PLS model relating the remaining data with the corresponding property values. The I regression vectors are averaged over the sample population, and the standard deviation is estimated at each of the $2J$ sensors. A critical limit is established as the maximum value of the ratio of coefficient to standard deviation in the noisy region (which ranges from $J+1$ to $2J$ sensors). Regression coefficients are then considered as significant in the experimental region from 1 to J when they exceed this limit. New cut-offs have also been recently discussed and applied to the NIR analysis of illicit drugs [44].

2.4. Variable importance in projection

The concept of variable importance in projection (VIP) uses the regression coefficients in the variable space [i.e., the vector **v** in Eq. (3)], the scores **T** and the weight loading factors contained in **W**, in order to define an importance parameter for each intervening sensor j [12]:

$$VIP_j = \sqrt{\frac{J \sum_{a=1}^{A} W_{ja} v_a^2 \mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}{\sum_{a=1}^{A} v_a^2 \mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a}} \tag{11}$$

where W_{ja} is an element of the matrix **W**, $\mathbf{t}_a$ is the ath column of **T**, and v_a the ath element of **v**. Eq. (11) implies that larger contributions from the signal at sensor j in predicting the target y are expected when the following parameters are significant: 1) the weight loadings (W_{ja}, computed for the ath factor at sensor j), 2) the ath component of the regression vector **v** (v_a), and 3) the scores for the ath factor ($\mathbf{t}_a$). This is understandable in view that all these parameters contribute to the final regression vector **b** [see Eqs. (5) and (7)].

The average value of the squared VIP over all sensors is 1, hence usually VIPs larger than 1 are considered to correspond to informative regions and included in the model, although different cut-off values have been proposed [12].

2.5. Moving window strategies

The simplest moving window strategy is interval-PLS (i-PLS), which adopts a fixed window size. In each of the regions defined by this moving window, a statistical indicator of the quality of the model is estimated. We have employed as statistical indicator the leave-one-out cross-validation mean square error (RMSECV), which is obtained as follows: each training samples is systematically removed, and the remaining ones are used for construction of the latent factors and regression. The predicted concentrations are then compared with the actual ones for each of the calibration samples, and the predicted error sum of squares [$\mathrm{PRESS} = \Sigma(y_{\mathrm{nom}} - y_{\mathrm{pred}})^2$] is calculated as a function of a trial number of factors. The RMSECV in each region is given by:

$$RMSECV = \sqrt{\frac{PRESS}{I}} \tag{12}$$

After finishing the cross-validation procedure, the optimum number of factors is suggested using the Haaland and Thomas criterion [45], which involves the following operations: 1) compute the ratios $F(A) = \mathrm{PRESS}(A{<}A^*)/\mathrm{PRESS}(A)$ [where A is a trial number of factors and A^* corresponds to the minimum PRESS], and 2) select the value of A leading to a probability of less than 75% that $F{>}1$. This procedure chooses the least complex model that is statistically indistinguishable from the optimal cross-validation model. In this way, less complex models with similar prediction accuracy are obtained. The sensor region with minimum cross-validation error is subsequently employed for model building and prediction on the independent test sample set.

When the window size is variable, two parameters control the spectral range employed for model building: the first sensor and the sensor size. They are varied in order to cover all possible pairs of values, using a certain minimum window. In each spectral region, cross-validation is carried out, and the minimum RMSECV indicates: 1) where the best spectral region starts and 2) which is the recommended spectral width for model building and prediction. We call this method VS-MW (variable-size moving-window).

2.6. Genetic algorithms

In this strategy, natural selection is algorithmically mimicked. The version employed in the present case is the so-called ranked regions genetic algorithm (RRGA) already described [27], except that the calibration set is randomly divided into two subsets only once in each computation cycle (70% for training and 30% for monitoring). During the program execution, the PLS model is built with the signals of the training sub-set of samples, only at the sub-region wavelengths for which the GA string contains a 1, discarding those containing a 0. Prediction proceeds on the monitoring sub-set for a number of latent variables from 1 to a certain maximum, and at the same wavelengths selected for calibration. The squared prediction errors are then analyzed in a manner similar to the cross-validation PRESS values discussed above, in order to estimate the number of factors. The objective function to be minimized is the root mean square error of prediction (RMSEP) for the monitoring sub-set using the optimum number of latent variables. This version is faster than the already published version of the algorithm.

The final string of 1s and 0s is then weighted inversely to the prediction error on the monitoring set, estimated by an i-PLS model for each of the wavelength sub-regions. In order to avoid chance predictions, which are common in GA analysis, the calculations are repeated ten times, and the averages at each sub-region are stored in a histogram. Wavelengths are then selected as those corresponding to values of histogram averages exceeding a certain tolerance, for example, 30 (on a scale where the maximum value is 100). The corresponding standard errors at each sub-region are also stored, in order to investigate the robustness of the procedure. For further details, see Ref. [27] and the material provided as Supplementary information.

2.7. Particle swarm optimization

PSO is designed to mimic the process of bird flocking [46,47]. This algorithm provides the relative importance of a given sensor (or sensor window) for building the PLS model. Notice that in the GA the inclusion of sub-regions encoded in the chromosomes is discrete, i.e., they are either included or excluded, whereas in PSO all sub-regions are included with a certain weigh.

In the present work the algorithm was implemented as described by Clerc and Kennedy [47], which differs from a previous work where PSO was applied to PLS variable selection [28]. In the present version, the objective function to be minimized is defined as in the case of the GA. We have also repeated the calculations 10 times, registering a histogram of the relative sub-region weight in each cycle (along with the corresponding uncertainties), in the same manner described above for GA. This tends to compensate for the random selection of the monitoring sub-set of samples, which is different for each algorithmic cycle. The Supplementary information collects additional specific details.

3. Data

3.1. Experimental data

3.1.1. Apparatus

Vis/NIR spectra were measured with a NIRSystems 6500 spectrometer, equipped with a cell with 1.0 mm optical path. Spectra were acquired using the spectrometer software ISISCAN, and then converted to ASCII files for further data processing.

Reference Brix data were measured with a Leica AR600 refractometer.

3.1.2. Samples

Sugar cane juices were analyzed at the quality control laboratory of the Estación Experimental Obispo Colombres, Tucumán, Argentina. The laboratory receives samples from several different cane processing units of the sugar-producing province of Tucumán. Cane samples are first processed in the sugar mills, where juice (65% of the cane) is extracted, and are then sent to the laboratory. For the calibration set, 59 samples were randomly selected, having Brix values in the range 11.76–23.15, as measured with the refractometer. The test set was composed of 46 samples with Brix values in the range 12.26–23.79, different than those employed for calibration. Vis/NIR spectra were measured in random order, in the wavelength range 400–2498 nm each 2 nm (i.e., 1050 data points). Absorbance spectra were pre-processed by applying standard normal variate, detrending and mean-centering (see below).

3.2. Simulated data

A synthetic data set was created with the purpose of illustrating the dangers in selecting informative wavelengths by inspection of the PLS regression coefficients. In the simulated data set, three components occur, with component 1 being the analyte of interest. All constituents are present in ten calibration samples and 100 test samples, at randomly chosen concentrations ranging from 0 to 20 units for components 1 and 2, and from 0 to 100 units for component 3. Fig. 1A shows the pure component spectra, all at concentrations of 20 units. From these noiseless profiles, calibration and test spectra were built. Specifically, each spectrum **x**, whether belonging to the calibration or to the test set, was created using the following expression:

$$\mathbf{x} = y_1\mathbf{S}_1 + y_2\mathbf{S}_2 + y_3\mathbf{S}_3 \tag{13}$$

where $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ are the pure component spectra at unit concentration, and y_1, y_2 and y_3 are the component concentrations in a specific sample. Gaussian noise with a standard deviation of 0.1 units was added to all concentrations, before inserting them in Eq. (13). Finally, a vector of signal noise (size $J \times 1$, standard deviation = 0.003 units) was added to each **x** vector after applying Eq. (13).

3.3. Data pre-processing

Pre-processing is usually performed in order to furnish more parsimonious PLS models, by avoiding the presence of certain spectral trends before regression. In the present work, it was only applied to the experimental Vis/NIR data set, where scattering corrections are frequently applied. Multiplicative scattering correction (MSC) [48] is used to correct for light scattering variations in reflectance spectroscopy. The standard normal variate (SNV) transformation, first introduced by Barnes et al. [49], is used to remove interferences due to light scattering and path length variations. Barnes et al. also described the use of detrending with a quadratic polynomial, together with the SNV transformation, in order to correct for curvilinear trends and linear baseline shifts in the spectra.

Spectral derivatives may also be employed to improve resolution and to highlight the selectivity towards a particular analyte when strong multicollinearity is present. Mean centering is almost universally applied, consisting of subtracting the mean calibration spectrum from both calibration and test spectra, and also the mean calibration concentration from the calibration values. After prediction, the value of y [see Eq. (6)] should be de-centered.

In our case, several alternatives were checked, with the best results obtained by applying detrending with SNV, no derivatives and mean-centering.

4. Software

PLS regression and variable-size moving window were applied using the software MVC1, already described for first-order multivariate calibration [50]. The variable selection algorithms based on regression coefficients, UVE, VIP, i-PLS and PSO were all implemented in MATLAB [51] according to the literature description of these methods. GA was applied using the RRGA algorithm, as already

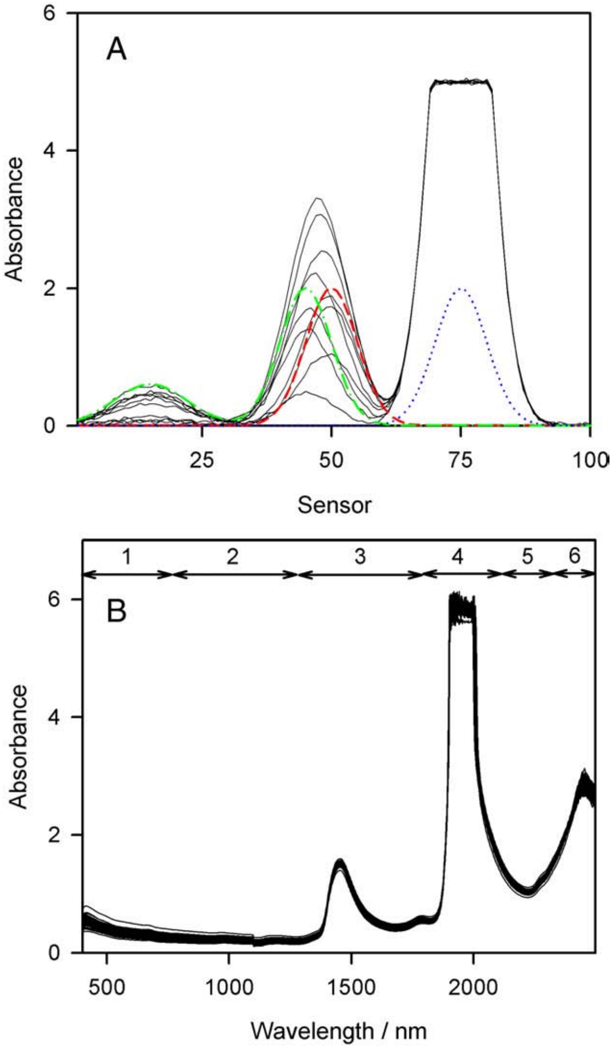


Fig. 1. A) Synthetic data set: thin black lines, the ten calibration spectra, thick red dashed line, pure component 1, thick green dashed-dotted line, pure component 2, thick blue dotted line, pure component 3. The pure components are at a concentration of 20 units. B) Vis/NIR calibration spectra for sugar cane juices, recorded in the range 400–2498 nm. The six spectral regions described in the text are shown on the top.

described [27]. All simulations were done with suitable in-house MATLAB routines.

5. Results and discussion

5.1. General considerations

It is first important to discuss the figures of merit which are relevant to the application of the strategies discussed in the present work. Some of them are: 1) the number of calibration latent variables, 2) the number of selected wavelength regions, 3) the speed of program operation, and 4) the prediction accuracy on an independent test sample set. Among the latter, the most relevant one is the accuracy in prediction, which is of paramount importance from the analytical perspective. The remaining issues have a comparably lower importance: the number of latent variables and selected sensors do not pose limitations on the calibration/prediction process. Finally, the time consumed by variable selection is small for most algorithms, and this activity is performed only once before the calibration phase, with the selected regions remaining constant during the prediction of all future samples.

5.2. Full spectra

5.2.1. Simulated data

In the synthetic calibration spectra (Fig. 1A), three major regions can be distinguished: sensors 1–30, dominated by component 2, sensors 30–60, where strong overlapping between 1 and 2 occurs, and sensors 60–100, where the main feature is a saturated signal due to component 3. These spectra intend to mimic the experimental Vis/NIR data set to be described below, and will be employed to test the ability of the different variable selection algorithms in choosing the analyte spectral features occurring at sensors 40–60. The full-spectrum PLS model requires 3 latent variables, and renders an RMSEP value of 0.16 units (Table 1), 1.6% with respect to the mean calibration concentration. As with all ideal data sets, PLS with the proper number of factors (three in this case) is able to provide a satisfactory answer upon employing all of the available spectral sensors. However, this may not be the case with the complex experimental data do be discussed below.

5.2.2. Experimental Vis/NIR data

Fig. 1B shows the calibration Vis/NIR spectra employed to train the PLS models for the determination of Brix in sugar cane juices. In order to assess the number of PLS latent variables, adequately pre-processed full spectral data (see above) were submitted to leave-one-out cross validation, with the result that a rather large number of factors (12) was required to model the data. Regression analysis of the independent test samples by PLS furnished the results for Brix which are collected in Table 2: a disappointing RMSEP value of 0.52 units is obtained. In comparison, refractometric measurements have errors which are typically less than 0.3 units.

These results can be explained by inspecting Fig. 1B, where six major spectral windows are apparent in the calibration spectra: 1) a small signal below ca. 800 nm, 2) a noisy region extending from 800 to 1350 nm, 3) a significant signal in the range 1350–1850 nm, 4) a high-absorbance noisy region (dominated by the intense NIR absorption by water) from 1850 to 2100 nm, 5) another significant signal in the range 2100–2350 nm, and 6) a noisy region at wavelengths longer than 2350 nm. The inclusion of noisy regions (especially the one corresponding to water absorption) is likely to be responsible for the low analytical performance of the full spectral model.

The spectral properties of the experimental calibration data provides the opportunity of checking the performance of variable selection techniques on an interesting test field, having noisy regions with both low and high absorbance, and also several separate regions with potentially useful signals.

5.3. Inspection of regression coefficients

5.3.1. Simulated data

The full spectrum of regression coefficients for the synthetic data, estimated with 3 latent variables, is shown in Fig. 2A, along with the corresponding uncertainties. The latter are plotted in Fig. 2A as thin

Table 1
Analytical results for the simulated data set on the 100 independent test samples.[a]

Method	Selected regions/sensors	A	RMSEP
None	1–100	3	0.16
b and s(**b**)	1–65	3	0.16
VIP	65–85	1	10
UVE	1–60	2	0.16
i-PLS	45–55	2	0.20
VS-MW	10–60	2	0.17
GA	40–55	2	0.17
PSO	45–55	2	0.20

[a] A = number of PLS latent variables, RMSEP = root mean square error of prediction.

Table 2
Analytical results for the determination of Brix in sugar cane juice samples using different methods of variable selection.[a]

Method	Selected regions/nm	*A*	RMSEP	REP%	SEN[b]
None	400–2498	12	0.52	2.9	0.056
b and *s*(**b**)	400–1896, 2064–2360	10	0.37	2.0	0.054
VIP	1878–2026	5	1.6	8.8	0.083
UVE	700–960, 1278–1894, 2072–2348	4	0.37	2.0	0.055
i-PLS	2260–2320	2	0.39	5.8	0.013
	1420–1480, 1540–1600, 1690–1750	3	1.0	2.2	0.012
VS-MW	1480–1778	4	0.30	1.7	0.010
GA	1300–1840, 2080–2320	4	0.28	1.6	0.052
PSO	1300–1540, 1660–1780, 2020–2320	4	0.34	1.8	0.044

[a] A = number of PLS latent variables, RMSEP = root mean square error of prediction, REP% = relative prediction error for the independent test sample set.
[b] In Absorbance units × Brix^{-1}.

red lines at $\pm 3 \times s(b_j)$, estimated from Eq. (10) with MSEC = 0.01 squared units (recall that the noise introduced in calibration concentrations was 0.1 units). As can be seen, the high-intensity band at 60–100 sensors is avoided by this strategy. However, not only the relevant spectral region 35–65 sensors where the analyte 1 responds is selected, but also the region 1–30 sensors, where only component 2 is responsive. This is the type of dangers described by Brown and Green [16].

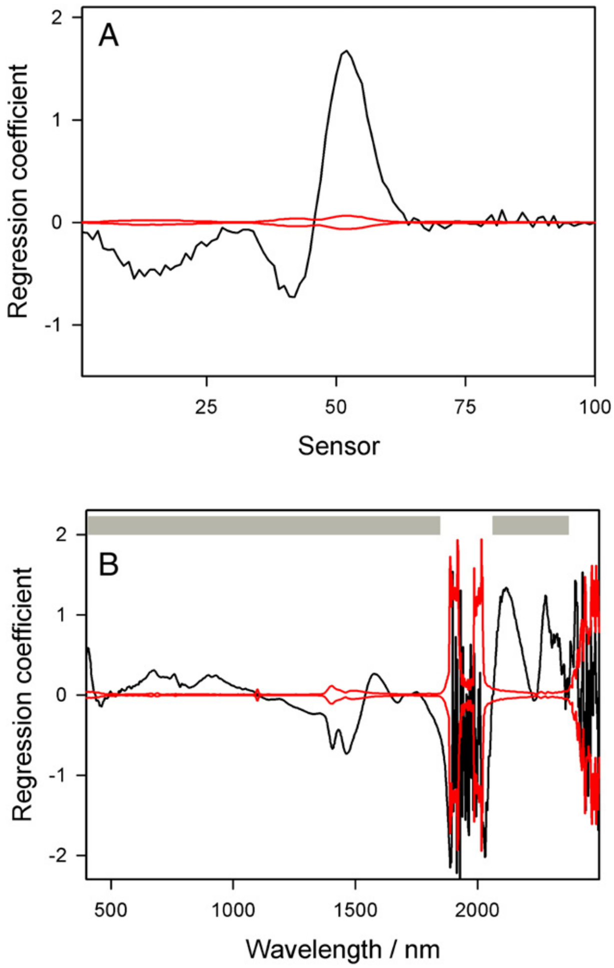


Fig. 2. Spectra of PLS regression coefficients corresponding to models built in the full spectral ranges (black solid lines). The thin red lines mark the uncertainties of the regression coefficients (see text). A) Simulated data set. B) Experimental Vis/NIR data set, with the gray boxes indicating the wavelength ranges selected for PLS regression.

In any case, Table 2 shows that the analysis in the selected region is similar to the full spectral study, as expected in ideally synthetic data.

5.3.2. Experimental Vis/NIR data

The PLS regression coefficients for the experimental data set in the full spectral range are shown in Fig. 2B, computed with the optimum number of 12 latent variables, as suggested by cross-validation. The importance of these coefficients may in principle be gathered from the comparison of each value of the vector **b** with its associated uncertainty. Critical limits for the regression coefficients, estimated from Eq. (10) using an MSEC of $(0.3)^2$ Brix squared units, are plotted in Fig. 2B as thin red lines at $\pm 3 \times s(b_j)$.

Inspection of Fig. 2B reveals that the relevant spectral windows where coefficients are significant involve: 1) all wavelengths below 1896 nm and 2) the region 2064–2360 nm. The noisy region 1896–2064 nm and also the wavelengths above 2360 nm should be discarded. Provided the criterion of the regression coefficients is followed, Table 2 collects the prediction of Brix in the independent test set by PLS regression, using data in these latter three regions. They appear to be satisfactory for Brix analysis in terms of the RMSEP value, but the number of latent variables remains rather large. In any case, the use of regression coefficients should always be checked with independent techniques in order to establish its validity for variable selection.

5.4. VIP

5.4.1. Simulated data

The importance of variables in projection, measured through the value of their VIPs, is observed in Fig. 3A in the full spectral range for the simulated data. Surprisingly, sensors having VIP values larger than 1 do only occur in the high-absorbance noisy region 65–85 sensors, which is unlikely to produce good analytical results. The result is considerably worse than that obtained in the previous section from the comparison of regression coefficients and their associated errors (Table 1).

5.4.2. Experimental Vis/NIR data

The trend discussed in the previous section is also observed on inspection of the VIP values for the experimental data set (Fig. 3B), because values larger than 1 do only occur in the high-absorbance region 1850–2100 nm. Indeed, Table 2 points to high prediction errors and low sensitivity towards Brix determination.

The largest VIP values correspond to the edges of the high water absorption at 1878 and 2026 nm. This may be due to the fact that the weight loading factors have large values at these wavelengths, and this provides unreasonably large importance to the VIPs. Since the average VIP value is 1, large VIPs for some regions implies low VIPs for other wavelength ranges, leading to the wrong conclusion that the potentially useful signals at 1350–1850 nm and 2100–2350 nm are almost unimportant (Fig. 3B).

5.5. UVE

5.5.1. Simulated data

Regarding the simulated data set, uninformative variable elimination renders the results shown in Fig. 4A when full spectral information is included in the model. Again, as in the case of the analysis of regression coefficients, two spectral regions appear to be important for this analysis: one including the analyte peak at 35–65 sensors, but an additional one including the irrelevant region 1–30 sensors, where only component 2 responds (Fig. 4A). As expected, the analytical results are similar than when selecting variables directly from the regression coefficients (Table 1).

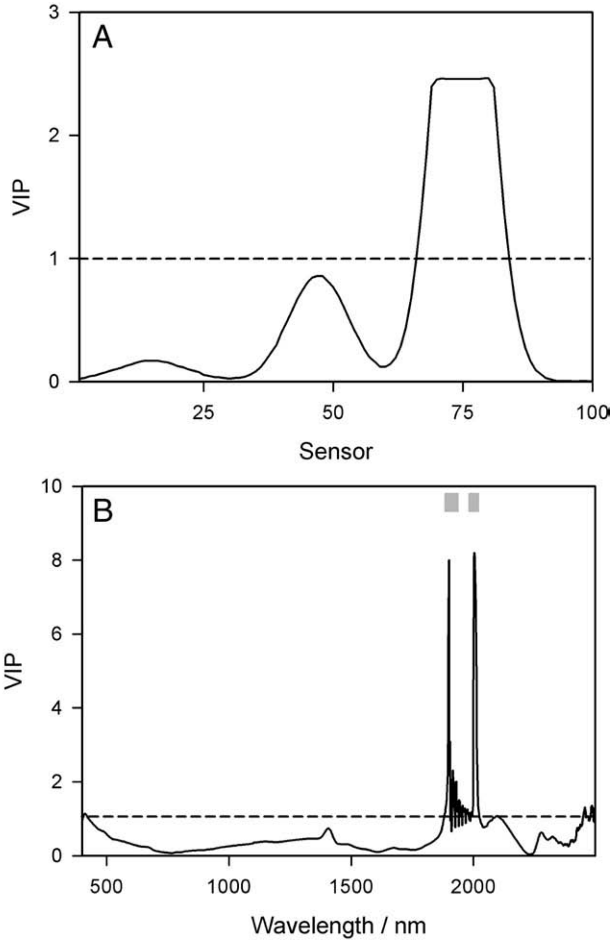


Fig. 3. Values of VIP estimated in the full spectral ranges for A) the synthetic data set, and B) the experimental Vis/NIR data set. In both plots the dashed lines indicate the critical limits of VIP = 1. The gray boxes in plot B) correspond to the wavelength ranges selected for PLS regression.

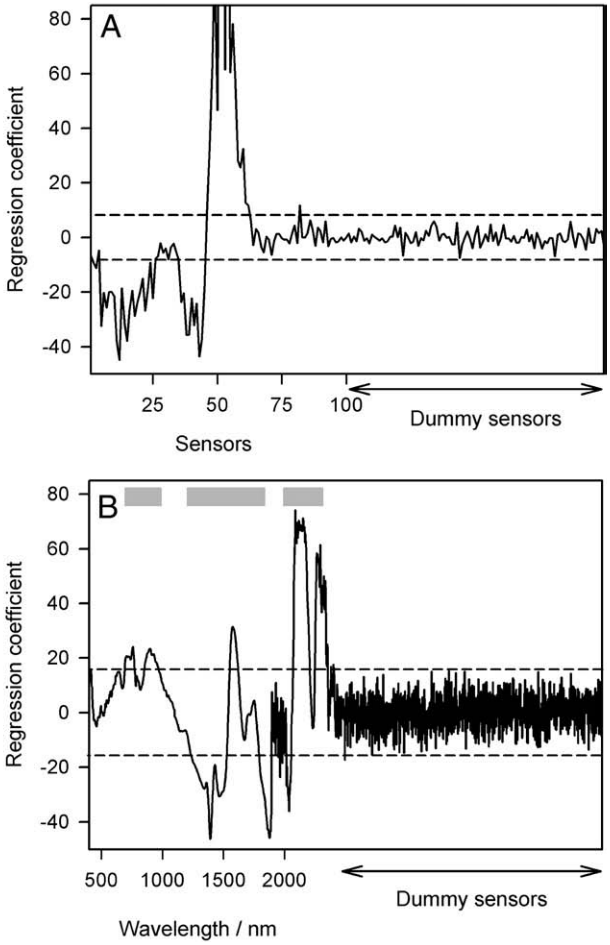


Fig. 4. Regression coefficients found by uninformative variable elimination (UVE) for A) the synthetic data set, and B) the experimental Vis/NIR data set. In both plots the dashed lines indicate the critical limits; coefficients with absolute values larger than the limits are considered significant. The gray boxes in plot B) indicate the wavelength ranges selected for PLS regression.

5.5.2. Experimental Vis/NIR data

When a similar UVE analysis is performed on the experimental data set, Fig. 4B is obtained, suggesting the following spectral zones for Brix prediction: 700–960, 1278–1894 and 2072–2348 nm. The analytical performance of the PLS model using these regions is shown in Table 2: the prediction error is similar to that furnished by analysis of the regression coefficients and their uncertainties, but the PLS model built after UVE requires a smaller number of factors.

Overall, judging from the analytical figures of merit, the study of the regression coefficients to the experimental data set seems to favour the strategy of uninformative variable elimination for variable selection. However, simulations indicate that this strategy tends to include regions where analytes may not be responsive, prompting to the application of alternative methods which do not rely on the regression coefficients.

5.6. Moving-window

5.6.1. Simulated data

The i-PLS method was applied to the simulated data set using a window of 5 sensors. In each sub-region, leave-one-out cross-validation allowed to estimate the optimum number of factors, and the RMSECV was computed in order to provide a guide for the selection of informative regions. Fig. 5A shows the results, with bars proportional to the RMSECV values. A clear window in the region 40–60 sensors appears, corresponding to the known spectral maximum of the analyte of interest. Indeed, using two consecutive regions, one corresponding to the minimum error in Fig. 5A (45–50 sensors), and another one having almost identical RMSECV (50–55 sensors), provides satisfactory prediction results (Table 1). This encouraging result, however, is typical of synthetic data with perfect behaviour, and may not always be encountered when dealing with complex experimental data (see below).

A more elaborate version of i-PLS, i.e., variable-size moving-window, renders the results shown in Fig. 5B for this data set. A landscape is obtained of RMSECV values as a function of first sensor and window width of all possible variable-size windows. For the synthetic data, the minimum RMSECV occurs at sensors 10–60 (Fig. 5B). In this region, cross-validation results appear to be satisfactory even when including a spectral zone where only component 2 responds (10–30, see Fig. 1A). Predictions proceed with a quality similar to previous analysis (Table 1).

5.6.2. Experimental Vis/NIR data

For the experimental data set, the simple i-PLS strategy was applied with a fixed interval of 30 sensors. It was found that using a smaller number of sensors in each sub-region led to poor predictions. Fig. 6A shows the corresponding RMSECV results. If they were blindly applied, then the best region for PLS model building would lie at 2288–2348 nm. The predictions on the independent test set in this latter region, collected in Table 2, shows otherwise. This is probably a spurious result obtained by chance.

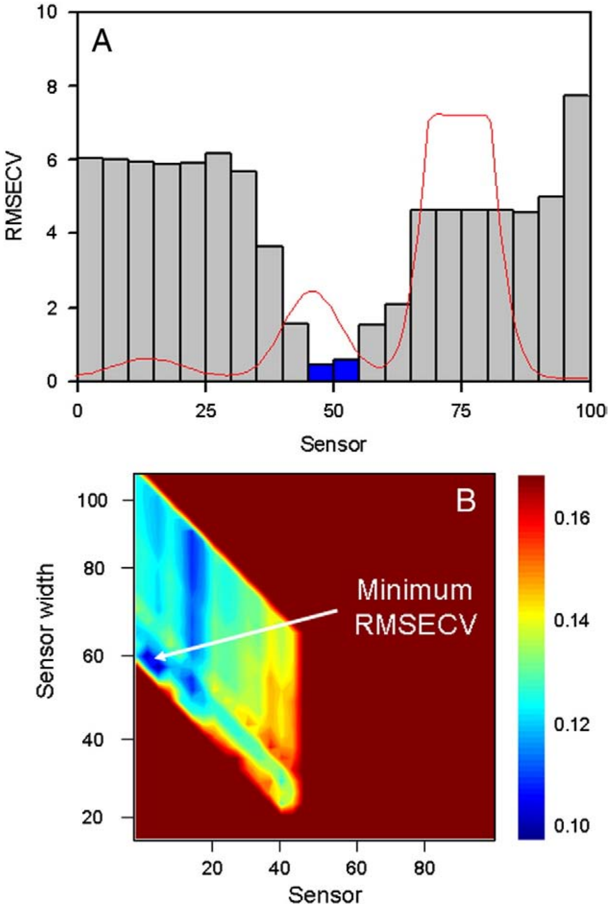


Fig. 5. A) Results from interval-PLS for the synthetic data set. The bars indicate the cross-validation root mean square error (RMSECV) in each of the sub-regions. The blue bars indicate the minimum RMSECV and a region a value close to the minimum. Superimposed is the mean calibration spectrum in red. B) Results from the variable-size moving window strategy. The contours correspond to values of the RMSECV as a function of first sensor and sensor width. The optimum region is indicated.

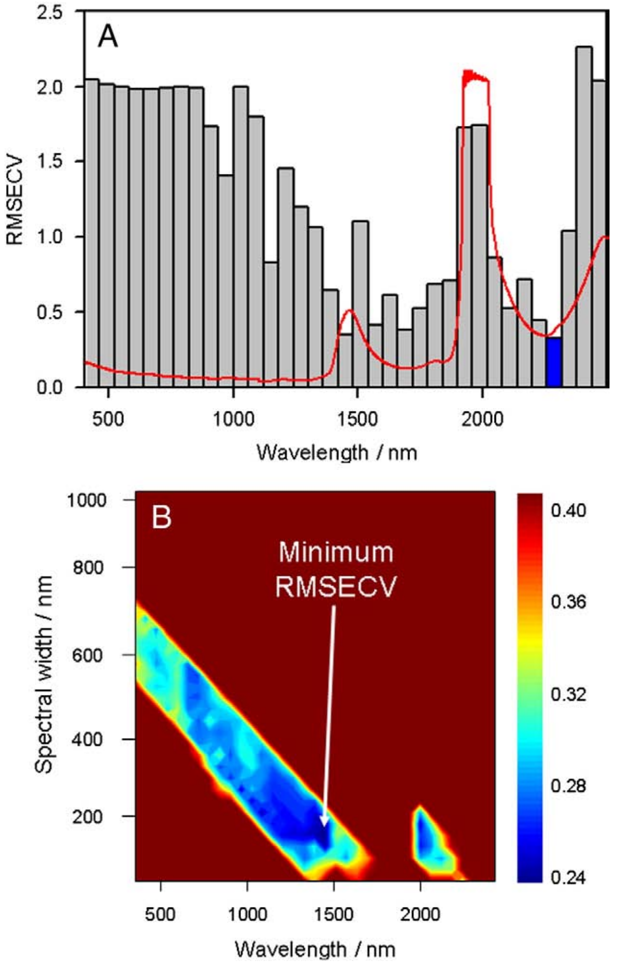


Fig. 6. A) Results from interval-PLS for the experimental Vis/NIR data set. The bars indicate the cross-validation root mean square error (RMSECV) in each of the sub-regions. The blue bar corresponds to the minimum RMSECV. Superimposed is the mean calibration spectrum in red. B) Results from the variable-size moving window strategy. The contours correspond to values of the RMSECV as a function of first sensor and sensor width. The optimum region is indicated.

The variable-size moving-window alternative, on the other hand, rendered the results shown in Fig. 6B concerning the Brix determination in the experimental samples. The best region lies at 1480–1778 nm (Fig. 6B), which provides the results of Table 2 on the independent test set. Overall, this latter strategy leads to good results, comparable to those obtained on visual inspection of the regression coefficients, although with considerably lower sensitivity.

5.7. GA

5.7.1. Simulated data

The application of the above described genetic algorithm allowed us to obtain the relative weight for each sub-region encoding 5 sensors. The weights are pictorially represented in a bar graph in Fig. 7A, with the corresponding uncertainties. Superimposed to this Figure is the mean calibration spectrum, for better appreciation of the selected wavelengths. The results is a net selection of the region where the analyte of interest is known to respond, avoiding both the high-intensity region due to component 3 at sensors 60–80, and the irrelevant region due to component 2 at sensors 1–30. This is in line with previous applications describing the success of RRGA in selecting regions for improving the analytical figures of merit in other complex systems [27]. Table 1 implies good prediction ability towards new test samples in the GA selected region. The final RMSEP is close to those obtained with other strategies, as is usual with ideally synthetic data.

5.7.2. Experimental Vis/NIR data

The application of the genetic algorithm to the experimental set proceeded by encoding 30 sensors (60 nm) in each sub-region. The final weights (including uncertainties) are shown in Fig. 7B, with the mean calibration spectrum superimposed. The first conclusion is that the GA avoids the potentially harmful region with high water absorbance. Comparatively larger weight is given to the important regions 1330–1870 and 2110–2350 nm (blue bars in Fig. 7B), implying that they are informative for the determination of Brix in these studied samples. When a final PLS model is built with all calibration samples in these spectral regions, the RMSEP and REP values are seen to be satisfactory, and better than those achieved on simple inspection of the regression coefficients (Table 2). The associated sensitivity using these ranges does also appear to be reasonable.

An important outcome of this window search strategy is the avoidance of the spectral wavelengths below 800 nm, in contrast to the analysis based on regression coefficients. It is likely that the spectral features observed in Fig. 1 at these wavelengths are unrelated to the sugar cane juice property being analyzed. This implies that the use of regression coefficients (see above) should be taken with some

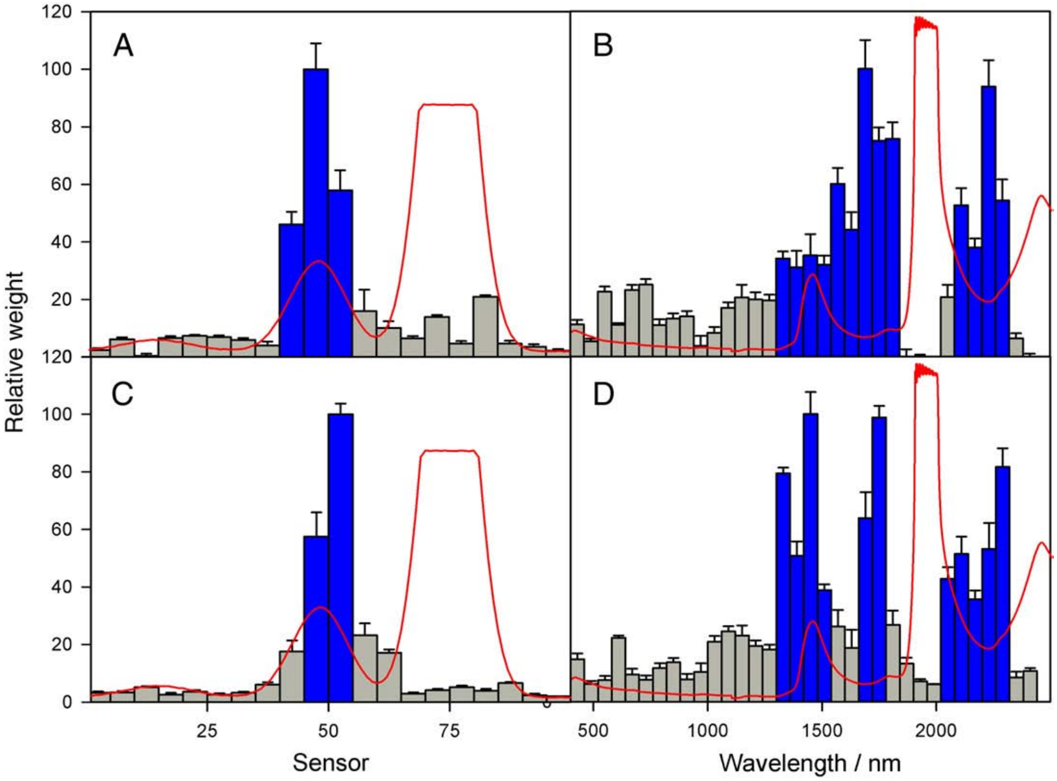


Fig. 7. Variable selection results from: genetic algorithms applied to the synthetic data set (A) and to the experimental data set (B), and particle swarm optimization applied to the synthetic data set (C) and to the experimental data set (D). In all cases, the bars indicate the relative weight of each of the sub-regions, with uncertainties indicated on top of each bar. The mean calibration spectrum superimposed in red. Blue bars correspond to sub-regions included in the final model, while gray bars correspond to the excluded sub-regions.

caution, and that conclusions based on the latter analysis should be confirmed by independent sources.

5.8. PSO

5.8.1. Simulated data

Particle swarm optimization was finally applied to the calibration data set. The final histogram registering the relative weight and uncertainties assigned to each spectral sub-region is shown in Fig. 7C. In the case of the application of this algorithm, as well as with the genetic algorithm discussed above, the inclusion of a final weighting of the sub-regions by i-PLS was needed in order to better discriminate among the different regions [27]. Otherwise, these algorithms tend to given unreasonably large weight to spectral ranges carrying low signal intensities.

Inspection of Fig. 7C provides support to PSO as a variable selection technique, although the analytical results are somewhat poorer in comparison with GA (Table 1).

5.8.2. Experimental Vis/NIR data

The PSO histogram of spectral sub-region weights for the experimental sugar cane juices is shown in Fig. 7D. As can be seen, PSO is highly efficient in removing the high-absorbance uninformative region due to intense water absorption. Moreover, as in the case of the application of GA, PSO avoids the region below 800 nm.

Using the wavelength ranges suggested by PSO, Table 2 shows reasonable figures of merit, although somewhat poorer than from GA, as was the case with the synthetic data set.

6. Conclusion

The analysis of both simulated and experimental data shows that regression coefficients should always be complemented with some sort of window search in order to properly select sensor ranges for successful PLS regression. In this regard, algorithms based on natural computation appear to be highly useful, because they are able to interrogate a large variable space in search of the best combination of wavelength regions, defined by a suitable statistical indicator. In the present case, genetic algorithms provided the most adequate answer to variable selection in multivariate calibration.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemolab.2010.04.009.

References

[1] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.

[2] R.K.H. Galvao, M.C.U. Araujo, Variable selection, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, 2009, p. 233.

[3] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Genetic algorithms as a method for variable selection in PLS regression, with applications to pyrolysis mass spectrometry, Anal. Chim. Acta 348 (1997) 71–86.

[4] Q. Ding, G.W. Small, M.A. Arnold, Genetic algorithm-based wavelength selection for the near-infrared determination of glucose in biological matrixes: initialization strategies and effects of spectral resolution, Anal. Chem. 70 (1998) 4472–4479.

[5] K. Hasegawa, T. Kimura, K. Funatsu, GA strategy for variable selection in QSAR studies: enhancement of comparative molecular binding energy analysis by GA-based PLS method, Quant. Struct. Act. Relat. 18 (1999) 262–272.

[6] A.S. Bangalore, R.E. Shaffer, G.W. Small, Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy, Anal. Chem. 68 (1996) 4200–4212.

[7] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm, Anal. Chem. 70 (1998) 35–44.

[8] J.-P. Gauchi, P. Chagnon, Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, Chemometr. Intell. Lab. Syst. 58 (2001) 171–193.

[9] United States Pharmacopoeia (USP), Chapter 1119, Near-Infrared Spectrophotometry, USP, Baltimore, MD, USA, 2008.

[10] P.J. Gemperline, J.R. Long, V.G. Gregoriou, Nonlinear multivariate calibration using principal components regression and artificial neural networks, Anal. Chem. 63 (1991) 2313–2323.

[11] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.

[12] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemometr. Intell. Lab. Syst. 78 (2005) 103–112.

[13] M.B. Seasholtz, B.R. Kowalski, Qualitative information from multivariate calibration models, Appl. Spectrosc. 44 (1990) 1337–1348.

[14] O.M. Kvalheim, T.V. Karstang, Interpretation of latent-variable regression models, Chemometr. Intell. Lab. Syst. 7 (1989) 39–51.

[15] A.J. Burnham, J.F. MacGregor, R. Viveros, Interpretation of regression coefficients under a latent variable regression model, J. Chemometr. 15 (2001) 265–284.

[16] C.D. Brown, R.L. Green, Critical factors limiting the interpretation of regression vectors in multivariate calibration, Trends Anal. Chem. 28 (2009) 506–514.

[17] F. Lindgren, P. Geladi, S. Rännar, S. Wold, Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms, J. Chemometr. 8 (1994) 349–363.

[18] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemometr. 23 (2009) 32–48.

[19] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[20] H.C. Goicoechea, A.C. Olivieri, Wavelength selection by net analyte signals calculated with the multivariate factor-based hybrid linear analysis (HLA). A theoretical and experimental comparison with partial least-squares (PLS), Analyst 124 (1999) 725–731.

[21] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2004) 183–191.

[22] J.A. Cramer, K.E. Kramer, K.J. Johnson, R.E. Morris, S.L. Rose-Pehrsson, Automated wavelength selection for spectroscopic fuel models by symmetrically contracting repeated unmoving window partial least squares, Chemometr. Intell. Lab. Syst. 92 (2008) 13–21.

[23] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Anal. Chim. Acta 461 (2002) 189–200.

[24] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometr. Intell. Lab. Syst. 41 (1998) 195–207.

[25] H.C. Goicoechea, A.C. Olivieri, Wavelength selection for multivariate calibration using a genetic algorithm: a novel initialization strategy, J. Chem. Inf. Comp. Sci. 42 (2002) 1146–1153.

[26] C.E. Boschetti, A.C. Olivieri, A new genetic algorithm applied to the near-infrared analysis of gasolines, J. NIR Spectrosc. 12 (2004) 85–91.

[27] H.C. Goicoechea, A.C. Olivieri, A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy, J. Chemometr. 17 (2003) 338–345.

[28] L. Xu, J.-H. Jiang, H.-L. Wu, G.-L. Shen, R.-Q. Yu, Variable-weighted PLS, Chemometr. Intell. Lab. Syst. 85 (2007) 140–143.

[29] J.H. Kalivas, N. Roberts, J.M. Sutter, Global optimization by simulated annealing with wavelength selection for ultraviolet–visible spectrophotometry, Anal. Chem. 61 (1989) 2024–2030.

[30] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimisation: a powerful tool for wavelength selection, J. Chemometr. 20 (2006) 146–157.

[31] W.H. Chang, S. Chen, C.C. Tsai, Development of a universal algorithm for use of NIR in estimation of soluble solids in fruit juices, Trans. ASAE 41 (1988) 1739–1745.

[32] P. Valderrama, J.W.B. Braga, R.J. Poppi, Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy, J. Braz. Chem. Soc. 18 (2007) 259–266.

[33] J. Irudayaraj, F. Xu, J.J. Tewari, Rapid determination of invert cane sugar adulteration in honey using FTIR spectroscopy and multivariate analysis, J. Food Sci. 68 (2003) 2040–2045.

[34] S. Sivakesava, J. Irudayaraj, Fourier transform infrared spectroscopy for Kona coffee authentication, J. Food Sci. 66 (2001) 972–978.

[35] J. Tewari, R. Mehrotra, J. Irudayaraj, Direct near infrared analysis of sugar cane clear juice using a fibre-optic transmittance probe, J. Near Infrared Spectrosc. 11 (2003) 351–356.

[36] A. Salgo, J. Nagy, É. Mikó, Application of near infrared spectroscopy in the sugar industry, J. Near Infrared Spectrosc. 6 (1998) A101–A106.

[37] S.L.T. Lima, C. Mello, R.J. Poppi, PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors, Chemometr. Intell. Lab. Syst. 76 (2005) 73–78.

[38] A.C. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Uncertainty estimation in spectroscopic multivariate calibration, Pure Appl. Chem. 78 (2006) 633–661.

[39] A.C. Olivieri, N.M. Faber, Validation and error, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, 2009, pp. 91–120.

[40] K. Faber, B.R. Kowalski, Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares, J. Chemometr. 11 (1997) 181–238.

[41] N.M. Faber, Uncertainty estimation for multivariate regression coefficients, Chemometr. Intell. Lab. Syst. 64 (2002) 169–179.

[42] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (1993) 251–263.

[43] A. Phatak, P.M. Reilly, A. Penlidis, The asymptotic variance of the univariate PLS estimator, Linear Algebra Appl. 354 (2002) 245–253.

[44] J. Moros, J. Kuligowski, G. Quintás, S. Garrigues, M. de la Guardia, New cut-off criterion for uninformative variable elimination in multivariate calibration of near-infrared spectra for the determination of heroin in illicit street drugs, Anal. Chim. Acta 630 (2008) 150–160.

[45] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202.

[46] J. Kennedy, R.C. Eberhart, Particle swarm optimization, Proc. IEEE Int. Conf. Neural Networks, Perth, Australia, Nov. 1995, pp. 1942–1948.

[47] M. Clerc, J. Kennedy, The particle swarm — explosion, stability, and convergence in a multidimensional complex space, IEEE Trans. Evol. Comput. 6 (2002) 58–73.

[48] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491–500.

[49] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.

[50] A.C. Olivieri, H.C. Goicoechea, F.A. Iñón, MVC1: an integrated Matlab toolbox for first-order multivariate calibration, Chemometr. Intell. Lab. Syst. 73 (2004) 189–197.

[51] MATLAB 7.0, The Mathworks, Natick, Massachusetts, USA, 2007.