JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Variable Selection, Outlier Detection, and Figures of Merit Estimation in a Partial Least-Squares Regression Multivariate Calibration Model. A Case Study for the Determination of Quality Parameters in the Alcohol Industry by Near-Infrared Spectroscopy

PATRÍCIA VALDERRAMA, JEZ WILLIAN B. BRAGA, AND RONEI JESUS POPPI*

Universidade Estadual de Campinas, Instituto de Química, C.P. 6154, 13084-971, Campinas SP, Brazil

Practical implementation of multivariate calibration models has been limited in several areas due to the requirement of appropriate development and validation to prove their performance to standardization agencies. Herein, a detailed description of the application of multivariate calibration based on partial least-squares regression models (PLSR) for the determination of soluble solids (BRIX), polarizable sugars (POL), and reducing sugars (RS) in sugar cane juice, based on near infrared spectroscopy (NIR), for the alcohol industries is presented. The development of the models, including variable selection and outlier elimination, and their validation by determination of figures of merit, such as accuracy, precision, sensitivity, analytical sensitivity, prediction intervals, and limits of detection and quantification, are described for a representative data set of 1381 sugar cane samples. Values estimated by PLSR are compared with appropriate reference methods, where the results indicated that the PLSR models can be used in the alcohol industry as an alternative to refractometry and lead clarification before polarization measurements (standard methods for BRIX and POL, respectively). For RS, the results of a titration reference method were compared with the PLSR estimates and also with an estimate based on BRIX and POL values, as actually used in the alcohol industry. The PLSR method presented a better agreement with the titration method. However, the results indicated that the RS estimates from both PLSR and those based on the BRIX and POL values, actually used, should be improved to a safe determination of RS.

KEYWORDS: Validation; near-infrared spectroscopy; PLSR; outliers; alcohol industry

INTRODUCTION

The industrial production of alcohol (ethanol) in Brazil can be considered an important and strategic area due its applicability as an alternative and less polluting fuel (*1*, *2*). Its production is based on cane sugar as the raw material, which presents the advantage of constituting a renewable source of energy. The main parameter for the calculation of sugar cane costs in the industry is the concentration of the recoverable total sugars (RTS), which is a function of the soluble solids (BRIX) and the polarizable (POL) and reducing sugars (RS) (*3*). BRIX can be defined as the percentage, in weight or in volume, of soluble solids expressed as sucrose. In sugar cane juice it is a quantitative measurement of the total solids (including all sugars), not giving any qualitative information about which sugars are present (*3*). Different from BRIX, POL is a measurement of the amount of sucrose in the mixture of sugars, because sucrose diverts the plane of polarized light. RS are considered to be glucose and fructose, since these sugars have the property of reducing copper from the Cu^{2+} state to Cu^{+} (*3*–*5*). The standard methods usually employed for determinations of the parameters mentioned above are densitometers, saccharimeters, and oxidation–reduction titration for BRIX, POL, and RS, respectively (*3*). However, due to the huge amount of sampling that is necessary to be carried out for each specific grower, according to the harvested area, to establish the payment, the determination of RS is not feasible and, in practice, it is just estimated using an equation that takes into consideration the BRIX and POL parameters (*3*, *4*).

Alternative methods for cane juice analysis have been investigated and tested with the aim to increase the reliability, uniformity of the method, and also the accuracy of the determinations (*3*, *5*–*7*), which would make possible a better evaluation of the raw material and the sugar cane growers' payment. For RS determination, flow injection analysis has already been described and presented good agreement with the reference methods. However, these methods were not validated with a representative number of samples in a test set (*6*, *7*). Recently, the implementation of near-infrared spectroscopy (NIR) in some alcohol industries with the aim of simultaneously determining these three parameters with only one spectrum has been accomplished by applying partial least-squares regression

10.1021/jf071538s CCC: $37.00
Published on Web 10/17/2007

(PLSR). PLSR methods have the very useful property of possibility the determination of an analyte in situations where a selective signal is not possible to be obtained (*8*). The performance of methods applying these models has already been confirmed in several applications in diverse areas, such as pharmaceutical (*9–11*), food (*12–14*), environmental (*15–17*), and biological (*18–20*) analysis, and recently their acceptance by official organs such as the American Society for Testing Materials (ASTM) (*21*) and the United States Pharmacopoeia (*22*) was achieved. However, for successful application for routine analysis, an appropriate validation is necessary to certify the prediction ability of the model, which is based on a determination of figures of merit.

On the basis of the official regulations of the sugar cane industry (*3*), the PLSR method should be validated. However, only the requirements for the number of samples used in a test set, which should be larger than 300, and the accuracy obtained by the model are mentioned, and no other information of how the calibration model should be developed is given. On the basis of the great importance of these properties in the sugar cane industries and the lack information in of the regulation to indicate the appropriate development and validation of a multivariate calibration model, this work presents a detailed description of the development and validation of PLSR models for determination of these properties based on NIR measurements. A previous paper (*23*) has already reported the determination of these parameters based on the full NIR spectra. In this paper, the procedures used for variable selection, identification of outlier samples in both calibration and test samples, the validation of the model based on figures of merit such as sensitivity, analytical sensitivity, selectivity, confidence intervals, precision at the level of repeatability, accuracy, limit of detection, and quantification are described and the model results are compared with reference values obtained by the standard methods to confirm the applicability of the proposed methodologies. For comparison, the results obtained using the RS estimated based on the equation that uses BRIX and POL values is presented, and this estimate is compared with the titration and NIR results, showing that the PLSR NIR model is better than the equation actually used for the industry for determination of this parameter.

MATERIALS AND METHODS

Experimental Measures. The experimental measurements of this work were carried out at an alcohol plant, Cocamar-Cooperativa Agroindustrial, located in São Tomé in the state of Paraná in Brazil. Ripe sugar cane arrives at the production unit transported by trucks and is sampled by a horizontal probe, ground, and taken to the laboratory. If the samples are of green sugar cane, for preharvest analysis, they are collected from the field by specialized technicians, ground at the factory, and taken to the laboratory. In the laboratory, the samples were pressed to 250 kgf/cm^2 in a hydraulic press for a period of 1 min, resulting in the cane juice for subsequent analyses.

NIR spectra were collected by using a NIRSystem spectrometer, model 5000, equipped with a monochromator, a tungsten filament source, a quartz cuvette having a 1 mm of optical path, polystyrene plate as internal reference, and a PbS detector using 32 scans. Acquisition of the spectra was accomplished in the range of 1100–2500 nm by using ISIScan software. Before spectra acquisition, the samples were filtered through cotton to eliminate suspended particles.

A total of 1381 samples of sugar cane juice were used in this work. Each sample was submitted to conventional analysis and the results were used as reference values for model development. The BRIX values were obtained directly, using a digital densimeter with a precision of 0.01 °Brix. For POL measurements, the cane juice was initially cleared with lead sub-acetate ($Pb(CH_3COO)_2 \cdot Pb(OH)_2$) and filtered through paper. The sample was analyzed in a digital saccharimeter with a precision of 0.01 units. The degree of polarization of the sample, expressed as percent juice, was calculated based on the saccharimetric reading (SR) and equation 1 (*3*).

$$\mathrm{POL_{ref}} = \mathrm{SR}[0.2605 - 0.0009882(\mathrm{BRIX})] \qquad (1)$$

where BRIX in eq 1 is expressed as percent juice and the others terms are appropriate scaling factors.

For RS determination, the standard methodology used for quantification via analysis was proposed by Eynon and Lane (*3*, *4*), which consists of the oxidation–reduction titration of the Fehling liqueur by the filtered cane juice. The titration reaction consists of the reduction of Cu^{2+} to Cu$_2$O from the Fehling solution by the glucose and fructose present in the juice. The RS quantity, also expressed as percent juice, present in each sample was obtained by eq 2, taking into consideration the standard volume used in the titration of the Felhling liqueur solution of 1.0% inverted sugar and the BRIX measurement (*3*):

$$\mathrm{RS_{\%juice}} = \frac{\left[\dfrac{5.2096 - \left(1.74993\ \mathrm{SR}\dfrac{\mathrm{VS}}{\mathrm{Vs}}\right)}{500}\right]}{25.64\dfrac{\mathrm{VS}}{\mathrm{Vs}}(0.00398\ \mathrm{BRIX} + 0.99692)} \qquad (2)$$

where VS is the volume (in milliliters) of cane juice used in the titration, Vs is the standard volume (in milliliters) of 1.0% invert sugar solution used in the titration, and the other terms are appropriate scaling factors.

As already mentioned, by reasons of time and cost for calculation of the sugar cane growers' payment, RS is not determined via analysis in practice but is just estimated (also expressed in percent juice) by an equation (eq 3) that takes into consideration the BRIX and POL parameters (*3*):

$$\mathrm{RS_{est}} = 9.9408 - 0.1049\left(\frac{\mathrm{POL}}{\mathrm{BRIX}}\right)100 \qquad (3)$$

For analysis, the 1381 samples were split into calibration and validation sets by the Kennard–Stone algorithm (*24*). The calibration set was composed of 1003 samples and the validation set was composed of 378 samples. Mean centered spectra were used for data preprocessing, followed by the elimination of an intense band in the region of 1900 nm (1890–2046 nm), due the water absorption (*25*). All calibration models for BRIX, POL, and RS parameters were developed using Matlab software version 6.5 and the PLS-Toolbox version 4.0 from Eigenvector Research, Inc. (*26*). Variables selection was accomplished through an interval PLSR (iPLSR) program, version 2.1, for Matlab, developed by Jesper Madsen Wagner, from the Royal Veterinary and Agricultural University of Denmark (*27*). The figures of merit of the models developed were calculated by homemade routines written in the Matlab environment.

Multivariate Calibration. Partial Least-Squares Regression (PLSR). The PLSR model has been discussed in detail in relevant references (*28–30*), thus only a brief description is presented here. The data matrix **X** is formed by the near-infrared spectra of the sugar cane juice and the vector **y** contains the reference values for each property of interest. One PLSR model for each property was built, and the outliers were identified and eliminated for each model.

In standard PLSR the relationship between the data matrix **X** and **y** is represented as a linear algebraic relation between their scores. The scores are obtained by decomposing the data matrices into a sum of rank one-component matrices (*28–30*).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\mathbf{T} + \mathbf{E} = \sum_{i=1}^{A} \mathbf{t_i}\mathbf{p_i^T} + \mathbf{E} \qquad (4)$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^\mathbf{T} + \mathbf{f} = \sum_{i=1}^{A} \mathbf{t_i} q_i^{\mathrm{T}} + \mathbf{f} \qquad (5)$$

where the **E** and **f** contain those parts of **X** and **y**, respectively, which are not explained by the model. Vector $\mathbf{t_i}$, which comprises the columns of **T**, is called the score vector, $\mathbf{p_i}$ and q_i are called the loading, and A

is the number of latent variables used for model development. Estimates for the interest property ($\hat{y}$) for a set of samples are obtained by multiplication of the NIR spectra by an appropriate regression vector (**b**), expressed as (*28–30*):

$$\hat{y} = \mathbf{Tq}^{\mathrm{T}} = \mathbf{XW}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{q}^{\mathrm{T}} = \mathbf{Xb} \qquad (6)$$

where **W** is the weight matrix determined in the PLS algorithm.

Interval partial least-squares regression (iPLSR) is an interactive extension of PLSR, which develops local PLSR models based on equidistant subintervals of the full-spectrum region. Its main use is providing an overall picture of the relevant information in different spectral subdivisions, thereby focusing on important spectral regions and removing interferences from other ones. The choice of the best iPLSR model is done by comparing the prediction performance of these local models with the global model built with the full spectrum (*27*). The comparison is mainly based on the validation parameter RMSECV (root mean squared error of cross-validation). Models based on the various intervals usually need a different number of PLSR components than do full-spectrum models to catch the relevant variation in **y**. This condition is caused by the varying amount of **y**-correlated information carried by the interval variables and is also related to the noise/interference carried by the variables. To ensure a fair comparison of the global and local models, it is necessary that the global and local model dimensions be selected separately (*27*).

Outlier Detection. Outliers can be defined as observations showing some type of departure from the bulk of the data. They may occur for many different reasons, such as, laboratory error, objects from another population, instrument error, etc. (*30*). Methods for their detection have already been described in several papers (*21, 30–32*). The three simplest forms to identify abnormal samples, usually recommended (*30*), are based on data with extreme leverage, unmodeled residuals in spectral data, and unmodeled residuals in the dependent variable.

Extreme Leverages. Leverage represents how much one sample is distant from the center of the data and can be defined as (*21, 30*):

$$h_i = \mathbf{t}_{A,\mathbf{i}}^{\mathbf{T}}\left(\mathbf{T}_A^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{t}_{A,\mathbf{i}} \qquad (7)$$

where **T** represents the scores of all calibration samples, $\mathbf{t_i}$ is the score vector of a particular sample, and A is the number of latent variables.

According to ASTM E1655-00 (*21*), samples with h_i larger than a limit value (h_{limit}), given by eq 8, should be removed from the calibration set and the model rebuilt.

$$h_{\mathrm{limit}} = 3\frac{A+1}{I_{\mathrm{c}}} \qquad (8)$$

where I_{c} is the number of calibration samples. Note that for models not centered in their mean the factor 1 in eq 8 is omitted.

It is not uncommon, when extreme leverages are eliminated in a first model and the model is rebuilt, to find new spectra with $h_i > h_{\mathrm{limit}}$. When repetitive application of the $h_i > h_{\mathrm{limit}}$ rule continues to identify outliers, the outlier test is said to “snowball”. If “snowballing” occurs, it may indicate some problem with the structure of the spectral data set. In these situations, the outlier test can be relaxed (*21*):

(1) The first model is built on an initial calibration set.

(2) Calibration spectra with $h_i > h_{\mathrm{limit}}$ are eliminated from the calibration set.

(3) A second model using the same number, A, of variables is built on the subset of calibration spectra and the calibration spectra with $h_i > h_{\mathrm{limit}}$ are identified for the second model. The second model could be used providing that no calibration samples have h_i greater than 0.5 (*21*).

Unmodeled Residuals in Spectra. Identification of outliers based on unmodeled residuals in spectral data are obtained by comparison of the standard deviation total residuals ($s(e)$) with the standard deviation of a particular sample ($s(e_i)$), defined as (*30, 32*):

$$s(e)^2 = \frac{1}{I_{\mathrm{c}}J - J - A_{\max}(I_{\mathrm{c}}, J)} \sum_{i=1}^{I_{\mathrm{c}}} \left(\sum_{j=1}^{J} (x_{i,j} - \hat{x}_{i,j})^2 \right) \qquad (9)$$

$$s(e_i)^2 = \frac{I_{\mathrm{c}}}{I_{\mathrm{c}}J - J - A_{\max}(I_{\mathrm{c}}, J)} \sum_{j=1}^{J} (x_{i,j} - \hat{x}_{i,j})^2 \qquad (10)$$

where J is the number of spectral variables, $x_{i,j}$ is absorbance value of the sample i at wavelength j and $\hat{x}_{i,j}$ is its estimated value with A latent variables. If a sample presents $s(e_i) > ns(e)$, where n is a constant that can vary from 2 to 3 (*30*), the sample should be removed from the calibration set. In this work, the constant was optimized as 2, which provides a good limit that could identify the samples presenting spectral residuals significantly larger than those observed from the other samples.

All tests described above can be applied to both calibration and validation sets. A further test appropriate to identify outliers in the validation set, when the calibration set was already optimized, is described in ASTM E1655-00 (*21*). It is based on the unmodeled residuals of samples measured at three different levels of concentration with seven replicates at each level and it can be used in substitution of the test for spectral residuals.

Unmodeled Residuals in Dependent Variables. Outliers are identified through comparison of the root mean square error of calibration (RMSEC) with the absolute error of that sample. If a sample presents a difference between its reference value (y_i) and its estimate ($\hat{y}_i$) larger than a constant that can vary from two to three times the RMSEC, it is identified as an outlier (*30*). In this work this, the constant was optimized as 3 and the RMSEC was determined as

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{I_{\mathrm{c}}} (y_i - \hat{y}_i)^2}{\nu}} \qquad (11)$$

where ν is the number of degrees of freedom, determined as proposed by Van der Voet, estimated as (*33*):

$$\nu = I_{\mathrm{c}} - I_{\mathrm{c}}\left(1 - \sqrt{\frac{\mathrm{MSEC}}{\mathrm{MSECV}}}\right) \qquad (12)$$

where MSECV is mean square error of cross validation estimated with the calibration samples and MSEC is the square of RMSEC.

Analytical Figures of Merit. *Accuracy.* This parameter reports the closeness of agreement between the reference value and the value found by the calibration model. In chemometrics, this is generally expressed as the root mean square error of prediction (RMSEP), which is an approximation of an average prediction error for the validation samples, obtained as (*30*)

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{I_\nu} (y_i - \hat{y}_i)^2}{I_\nu}} \qquad (13)$$

where I_ν is the number of prediction samples. However, RMSEP is a global parameter that incorporates both systematic and random errors. Hence, an F-test with the RMSEP of two methods is not appropriate to compare the accuracy, a better indicator is the regression of found versus nominal concentrations values and estimation of the linear regression slope and intercept, including the consideration of the elliptical joint confidence regions (*34*).

Sensitivity (SEN). This parameter is the fraction of analytical signal that is due to the increase of the concentration of a particular analyte at unit concentration. In inverse multivariate calibration models (such as PLSR), it is defined as (*35, 36*):

$$\mathrm{SEN} = \frac{1}{\mathbf{b}_A} \qquad (14)$$

where $\mathbf{b}_A$ is the vector of the regression coefficients in eq 6 with A latent variables.

Analytical Sensitivity (γ). The inverse of this parameter reports the minimum concentration difference between two samples that can be determined by the model. Considering that the spectral noise represents the larger source of error, it can be approximated determined as (*36–38*)

$$\gamma = \frac{\mathrm{SEN}}{\delta x} \qquad (15)$$

where δx is an estimation of the noise level in the data, which can be obtained by replicate measurements of a blank sample. In this work 28 replicates were used, and δx was estimated as the square root of the mean variance in each wavelength.

Prediction Intervals. This parameter can be defined as a range within which we may assume, with a given degree of confidence, that is, a certain probability, that the true value for that concentration of the analyte of interest is included. It can be determined from the application of *t* statistics and the estimated standard error of prediction ($s(\hat{y} - y_{\mathrm{ref}})$), expressed as (*36, 39*)

$$\mathrm{PI}(y_{\mathrm{ref}}) = \hat{y} \pm t_{\nu,1-\alpha/2}\, s(\hat{y} - y_{\mathrm{ref}}) \qquad (16)$$

$$PI(y_{\mathrm{ref}}) = \hat{y} \pm t_{\nu,1-\alpha/2}\sqrt{s^2(1 + h + 1/I_c)} \qquad (17)$$

where α is the significance level required for the prediction interval, $t_{\nu,1\text{-}\alpha/2}$ is the corresponding critical level for Student's *t* distribution with ν degrees of freedom, determined as proposed by Van der Voet (*33*), I_c is the number of calibration samples, *h* is the leverage (eq 7), and s^2 is an estimate of the standard deviation of the fit error for the training set, determined as

$$s^2 = \mathrm{MSEC} = \frac{\sum_{i=1}^{I_{\mathrm{cal}}} (y_i - \hat{y}_i)^2}{\nu} \qquad (18)$$

In eq 17 it is considered that mean centering is employed in the data, when this is not applied the term $1/I_c$ should be removed from eq 17.

Detection Limit (LOD). Following IUPAC recommendations, the LOD can be defined as the minimum detectable value of net signal (or concentration) for which the probabilities of false negative (α) and false positive (β) are 0.05 (*40*). It can be determined in multivariate calibration analogously to univariate calibration (*41, 42*):

$$\mathrm{LOD} = 3.3\ \delta x \|b_k\| = 3.3\ \delta x \frac{1}{\mathrm{SEN}} \qquad (19)$$

Equation 19 is the most simple and employed for determination LOD, it considers that the spectral noise represents the larger source of error. Therefore, eq 19 provides an overoptimistic value of the LOD. Other approximations taking into account the leverage and other sources of errors lead to a more specific LOD sample (*43*).

Quantification Limit (LOQ). The ability of quantification is generally expressed in terms of the signal or analyte concentration value that will produce estimates having a specified standard deviation, usually 10% (*40*). Following the same consideration of the LOD, the LOQ can be determined as (*41*)

$$\mathrm{LOQ} = 10\ \delta x \|b_k\| = 10\ \delta x \frac{1}{\mathrm{SEN}} \qquad (20)$$

RESULTS AND DISCUSSION

The calibration and validation data sets were composed by 1003 and 378 samples, respectively, selected by the Kennard–Stone algorithm (*24*). In this algorithm, the first sample selected is that one with the largest distance from the center of the data, and the next sample again presents the largest distance from the last point, and so on, until completing the number of samples for the calibration set. The optimum model dimension was determined by the minimum RMSECV for the calibration samples, obtained by 10 continuous blocks of cross-validation, which preset as results five, seven, and five latent variables for BRIX, POL, and RS, respectively.

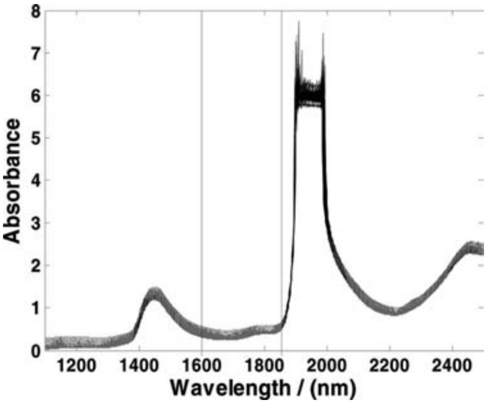

Figure 1. NIR spectra of sugar cane juice. The interval between 1600 and 1850 nm corresponds to the variables selected by iPLS.

Figure 1 shows the variables selected by iPLSR, which resulted in an interval of 1600–1850 nm. For this variable selection, the whole spectra was divided in five equally spaced intervals. The observed RMSECV with the full spectra and the iPLSR selection were 0.64, 0.85, and 0.34 and 0.63, 0.79, and 0.31 for BRIX, POL, and RS, respectively. However, the PLSR models built with the selected region presented a lower RMSEP (**Table 1**) than that observed with the full spectra 0.28, 0.42, and 0.26 (*23*) for BRIX, POL, and RS, respectively, mainly for POL. Moreover, with this selected region a lower time for spectra acquisition is necessary. The dominant absorption band observed in the selected region can be attributed to the first overtone of C–H stretching, from the sugars. This broad band is a composite of bands due to different sugars and, hence, it is not possible to assign it solely to sucrose (*25*). Nevertheless, the bands display the absorption in the region where the long chain C–H molecules should absorb; therefore, they can be used or selected as an indicator of the sugar content in the juice samples analyzed.

The calibration set was optimized by elimination of the samples that presented extreme leverages, extreme unmodeled residuals in the concentrations (*Y*) or spectral data (*X*) (as described in Outlier Detection in Materials and Methods). For validation samples, the outlier tests were used based on extreme leverages and spectral residuals. The outlier test with spectral residuals based on replicates, described in the ASTM E1655-00 (*21*), was also performed. However, for this data set this outlier test presented the same results as the test using spectral residuals (eqs 9 and 10). **Figures 2** and **3** present the plots observed using these tests in the calibration and validation samples, respectively.

Table 2 presents the results for outlier elimination and the variation of the RMSEC and RMSEP values for the three properties of interest. It is observed that, for all parameters, when the outliers identified in the first model were eliminated and the model was rebuilt, new outliers were identified. ASTM E1655-00 classifies these occurrences as a "snowballing effect" (*21*). In these cases it is advisable to detect and remove the outliers until the second model. Thus the third model for each analyte was built and considered optimized with 893, 914, and 891 calibration samples for BRIX, POL, and RS, respectively. Therefore, the outliers identified in the third models for the three parameters were not eliminated. After the optimization of the calibration set, the outlier tests were applied to the validation set and the outliers eliminated. In practice, the outlier test based on the residuals in the concentration values (*Y*) could not be

Table 1. Analytical Figures of Merit for PLSR Models for the Properties of Interest

figures of merit		BRIX	POL	RS
accuracy[a]	RMSEC	0.27	0.29	0.27
	RMSEP	0.28	0.27	0.25
precision[a]		0.04	0.04	0.02
sensitivity[b]		2.3×10^{-3}	5.8×10^{-4}	5.1×10^{-3}
analytical sensitivity^{-1a}		6.2×10^{-2}	2.4×10^{-1}	2.8×10^{-2}
fit	slope	0.99 ± 0.01[c]	0.99 ± 0.01[c]	0.76 ± 0.04[c]
	intercept	0.15 ± 0.15[c]	0.10 ± 0.10[c]	0.19 ± 0.04[c]
	corr coef (R^2)	0.992	0.994	0.758
LOD[a]		0.19	0.73	8.40×10^{-2}
LOQ[a]		0.62	2.45	0.28

[a] % juice. [b] % juice^{-1}. [c] 99% confidence interval.

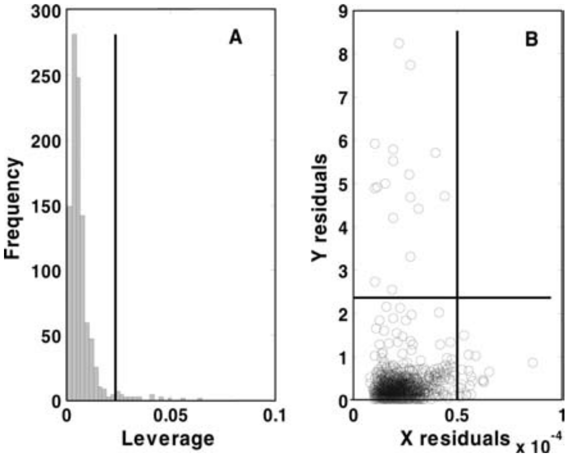

Figure 2. Visualization of outlier detection in the calibration set in the first model for POL. (**A**) Histogram of leverage values. (**B**) Plot of spectral residuals against concentration residuals. The lines show the limits for outlier detection.

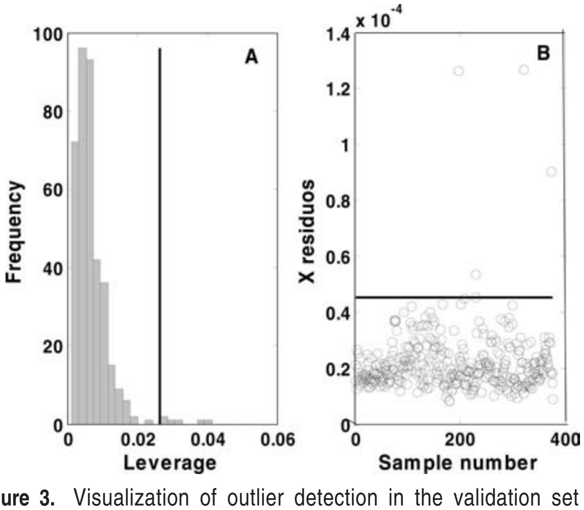

Figure 3. Visualization of outlier detection in the validation set for POL, after the calibration was optimized. (**A**) Histogram of leverage values. (**B**) Plot of spectral residuals. The line shows the limits for outlier detection.

applied, since in practice the reference value in a real prediction sample is unknown. However, as these reference values are available, the test was performed, where 18, 21, and 7 outliers in *Y* were identified in the 378 validation samples for BRIX, POL, and RS, respectively. Cane juice can be considered a complex sample and, in this sense, there is a probability of occurrence of errors, which can be due to errors in the estimated reference values or the PLSR estimated values. On the basis of the occurrence of these outliers, it is possible to estimate the probability of occurrence of outliers in *Y*, which were approximately equal to 4.8, 5.6, and 1.8%. These estimates can be considered as an acceptable result, since sugar cane juice is susceptible to several variations. An alternative to reduce these probabilities is the measurement of replicates of the samples, which can increase confidence in the result, but would increase both time and cost.

Results for the figures of merit are shown in the **Table 1**. The RMSEC and RMSEP showed that the estimated values for the BRIX and POL presented a good agreement with reference methods. Precision, at the level of repeatability, was assessed by analysis of three samples with six replicates each, in measurements made on the same day. The results for BRIX and POL showed that the repeatability of the multivariate models was 0.04% juice for both models. These results are better than those required by regulations for evaluation of the quality of the cane sugar (0.30 and 0.60% juice, respectively, for BRIX and POL) (*3*). For RS, a good result was also observed for precision, but the RMSEP value suggests the presence of a relative uncertainty in the model. Considering that the RS values occur approximately between 0.10 and 3.00, a mean relative error of 8.6% is observed for the validation samples. For RS there is no regulation that specifies precision or accuracy and, as already mentioned, it is not determined in practice, just estimated by eq 3. It is important to note that the true RMSEP and RMSEC values for the PLSR analysis are probably better than their values present in **Table 1**. This can be explained because RMSEP and RMSEC values incorporate two uncertainty sources: the one from the PLSR analysis and the error arising from the method employed to establish the nominal reference values, which is unknown in this particular application. If the uncertainty of the nominal reference values is known, the RMSEP and RMSEC could be corrected by the approximation proposed by Faber and Kowalski (*44*), providing only the average prediction error due the PLSR model.

The sensitivity and analytical sensitivity parameters showed good results for the three analytes estimated, taking into account the analytical range of the models. Analytical sensitivity is simpler and more informative for comparison and to judge the sensitivity of an analytical method. The inverse of this parameter permits establishment of a minimum concentration difference, which is discernible by the analytical method in the range of concentrations where it was applied, considering a perfect fit of the model. On the basis of this result, for example, in POL it is possible to distinguish samples with concentration differences of 0.24% juice. However, this value is an optimistic estimate that considers the spectral noise representing the larger source of error and does not take into account the lack of fit of the model.

Figure 4 shows the goodness of fit of the models, presented by plotting the reference values against the estimates for BRIX, POL, and RS, respectively. The slope and intercept for these

Table 2. Results for the Number of Outliers Identified in Each Test for Each Property of Interest and the Variation of the rmsEC and rmsEP Values Observed

		no. of outliers detected in each test					
models[a]	samples	leverage	*X* residuals	*Y* residuals	total	RMSEC	RMSEP
Brix M1	1003	15	42	14	66	0.63	0.77
Brix M2	937	14	22	13	44	0.31	0.78
Brix M3 Opt	893	5	12	13	28	0.27	0.78
Brix_{Val}	378	2	19	18	39	0.27	0.78
Brix_{Val} Opt	339	0	0	0	0	0.27	0.28
Pol M1	1003	26	21	17	53	0.78	0.91
Pol M2	950	13	10	15	36	0.37	0.92
Pol M3 Opt	914	12	2	3	16	0.29	0.92
Pol_{Val}	378	6	4	21	26	0.29	0.92
Pol_{Val} Opt	352	0	0	0	0	0.29	0.27
RS M1	1003	21	45	11	69	0.31	0.30
RS M2	934	14	23	6	43	0.28	0.31
RS M3 Opt	891	5	8	1	14	0.27	0.32
RS_{Val}	378	2	22	7	28	0.27	0.32
RS_{Val} Opt	350	0	0	0	0	0.27	0.25

[a] M, model; Opt, optimized; Val, validation set.

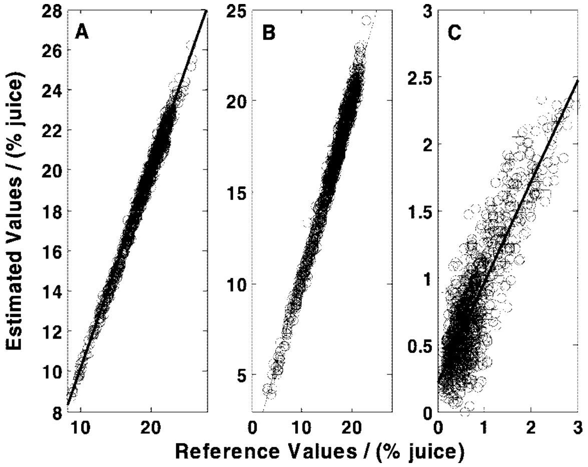

Figure 4. Reference values against the values estimated by the PLSR model for BRIX (**A**), POL (**B**), and RS (**C**), respectively: calibration samples (○) and validation samples (+).

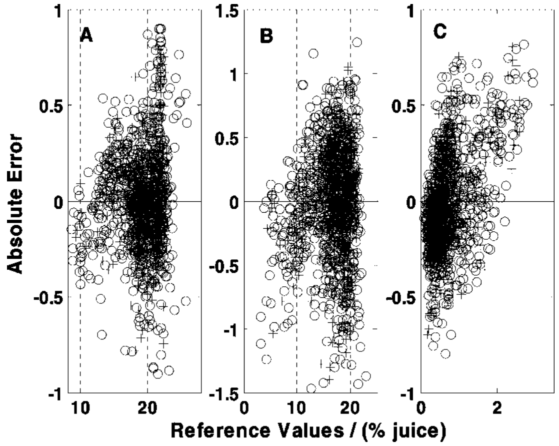

Figure 5. Reference values for BRIX (**A**), POL (**B**) and RS (**C**), respectively, against the absolute error: calibration samples (○) and validation samples (+).

linear fits are also shown in **Table 1**. On the basis of 99% confidence intervals, it can be concluded that no constant or proportional systematic errors were observed for BRIX and POL, since the intervals contain the expected values of 1 and 0 for slope and intercept, respectively. Also for BRIX and POL a similar and clearly better goodness of fit was observed than for RS. For RS, based on the slope and intercept, a significant systematic error and inferior fit were observed, which might be caused by a larger variance in the reference values (since it is difficult to determine the end point of the titration method). Alternatively it might indicate that this parameter presents some nonlinear behavior within the spectral data.

Figure 5 shows the plot of the residuals of the calibration and validation samples for BRIX, POL, and RS parameters, respectively. The distribution of the errors for BRIX and POL present an approximately random behavior, while for the RS parameter some tendency can be observed, which reinforces the suspicion of some nonlinearity in this parameter.

Limits of detection (LOD) and quantification (LOQ) for the models show result coherent with the measured quantities and the RMSEP obtained. On the basis of these values, the PLSR models are appropriate for BRIX and POL, since the ranges of these parameters are approximately 8.00–26.00% juice and 3.00–24.00% juice, respectively. For RS, the LOQ obtained show that the PLSR model is not able to quantify samples with RS values below 0.28, although the expected range of RS is 0.10–3.00% juice.

Table 3. Percentage of Coverage and Mean Prediction Intervals (PIs) Obtained for Each Parameter for Three Different Probability Levels

	BRIX (%)		POL (%)		RS (%)	
probability level (%)	recovery	PI	recovery	PI	recovery	PI
99.0	96.4	±0.70	98.6	±0.76	98.6	±0.70
95.0	90.5	±0.53	96.0	±0.58	95.1	±0.53
90.0	86.9	±0.45	93.5	±0.49	91.9	±0.44

Table 3 shows the results of percent of coverage of the prediction intervals (PIs), which represent the percentage of samples that have their true value inside the range estimated by the confidence intervals at probabilities of 99.0, 95.0, and 90.0%. Results showed that coverage for the PIs were appropriately close to those expected theoretically, where the largest difference observed was 4.5% for BRIX. **Table 3** also shows the mean PIs estimated for the PLSR models, which presents an acceptable uncertainty for BRIX and POL. For example, considering the 95% PIs for POL, the value 0.58 indicates that for a sample with a concentration of 20.0% juice, this value must be between 19.42 and 20.58% juice. **Figure 6** shows the error bars of the PIs for the validation samples for POL, which illustrates the PIs obtained. For the RS parameter, the calculated PIs were incompatible with the concentration

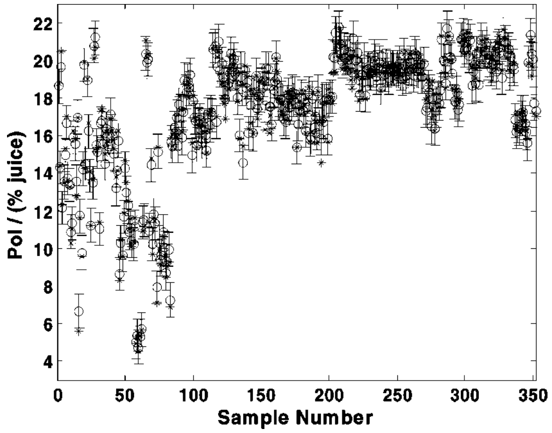

Figure 6. Error bars for the validation samples illustrating the prediction intervals obtained for POL: (○) reference values and (∗) estimated values for PLSR.

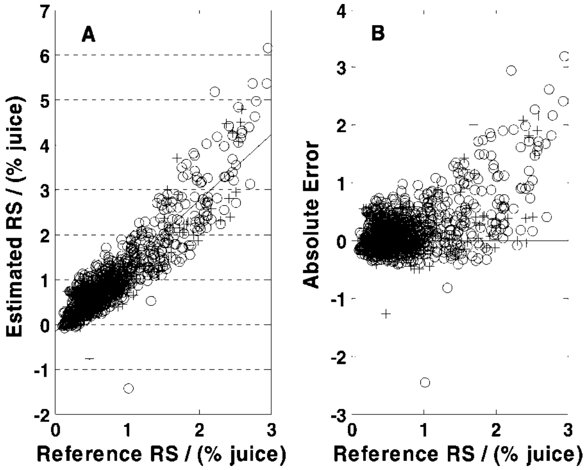

Figure 7. Plot of the RS reference values obtained by the titration method against the (**A**) RS_{est} values obtained by eq 3 and (**B**) the absolute errors.

range studied, which agrees with the results obtained for LOQ, and indicates that the proposed methodology based on NIR is not suitable for RS determination.

Figure 7 shows the results obtained for the RS_{est}, based on eq 3. A significant dispersion can be observed and a tendency in the absolute errors, similar to the results obtained with the PLSR model. The RMSEP calculated based on the RS_{est} and the reference values (obtained by the titration method) present a value of 0.38, which is significantly larger than that obtained with the PLSR model. **Figure 8** shows the elliptical joint confidence regions; for BRIX and POL it was observed that the ellipses contain the ideal point (1, 0), for slope and intercept respectively, showing that the reference method and PLSR results do not present a significant difference with 99% of confidence. For RS, it was observed that both PLSR and RS_{est} results do not contain the ideal point; hence both differ significantly from results of the reference method. However, the ellipse for the PLSR results presents a lower size and is nearest of the ideal point, showing that the PLSR results are in a better agreement than the RS_{est}.

Determinations of BRIX, POL, and RS parameters based on NIR spectra and multivariate calibration were built and validated by determination of the figures of merit, using a representative number of samples, where feasible, and acceptable results for BRIX and POL were obtained, which can be considered validated according the regulations followed by the sugar cane industry (*3*) and ASTM E1655-00 (*21*).

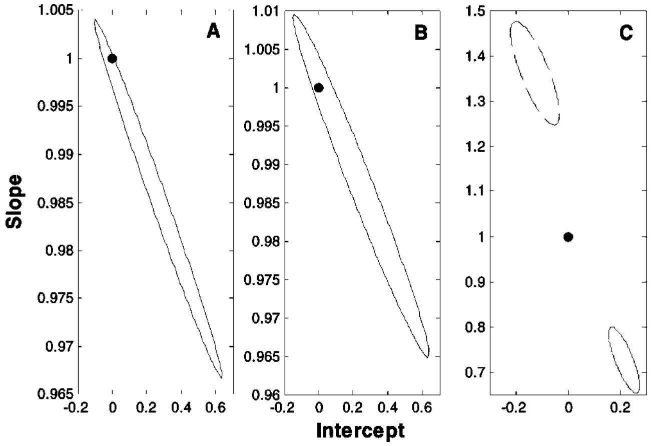

Figure 8. Elliptical joint confidence regions for slope and intercept of the regression of predicted concentrations vs reference values using ordinary least squares: (●) point of intercept equal to 0 and slope equal to 1; (**A**) BRIX; (**B**) POL; (**C**) RS, dashed line RS_{est} and solid line PLSR results.

The prediction errors obtained for BRIX and POL were smaller than those required by the regulations (*3*) and the estimated confidence limits for prediction samples showed good agreement with the expected probability of coverage. The models showed a large sensitivity capacity, differentiating samples with a small difference of concentration. The values for accuracy, precision, and other figures of merit presented promising results, indicating that the models developed for near-infrared spectroscopy for BRIX and POL can be safely used in the sugar cane industry as an alternative to refractometry and lead clarification for polarization measurements (standard methods for BRIX and POL, respectively). For the RS parameter, compared to the standard titration method, the absolute errors obtained for the NIR method were smaller than the errors found using eq 3. However, despite of the lack of regulations for this parameter, these errors and the estimated figures of merit for the multivariate model demonstrate that neither the PLSR model based on the NIR spectra nor eq 3 can be indicated for realistic determinations of RS in sugar cane juice. For this parameter, a viability study must be performed to optimize both the reference method and the NIR methodology.

ACKNOWLEDGMENT

The authors thank COCAMAR (Cooperativa Agroindustrial, São Tomé, Paraná, Brazil) for allowing the development of the experimental setup of this work in its industrial unit. We also thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and UNICAMP (Universidade Estadual de Campinas) for fellowships and financial support.

LITERATURE CITED

(1) Shreve, R. N.; Brink, J. A., Jr. *Chemical Process Industries*; McGraw-Hill International: Auckland, New Zealand, 1977.

(2) Payne, J. H. *Unit Operations in Cane Sugar Production*; Elsevier: Amsterdam, The Netherlands, 1982.

(3) Operational Norms for Sugar Cane Quality Evaluation - CONSECANA - PR, FAEP: Curitiba, Brazil, 2000.

(4) George, P. M. *Cane Sugar Handbook–a manual for canesugar manufacturers and their chemists*; John Wiley & Sons: New York, 1963.

(5) Johnson, T. P. Cane juice analysis by near infrared (NIR) to determine grower payment. *Int. Sugar J.* **2000**, *102*, 603–609.

(6) Oliveira, A. F.; Fatibello-Filho, O. Flow injection spectrophotometric determination of reducing sugars using a focalized coiled reactor in a domestic microwave oven. *Talanta* **1999**, *50*, 899–904.

(7) Alves, E. R.; Fortes, P. R.; Borges, E. P.; Zagatto, E. A. G. Spectrophotometric flow-injection determination of total reducing sugars exploiting their alkaline degradation. *Anal. Chim. Acta* **2006**, *564*, 231–235.

(8) Booksh, K. S; Kowalski, B. R. Theory of analytical chemistry. *Anal. Chem.* **1994**, *66*, 782A–791A.

(9) Sena, M. M.; Chaudhry, Z. F.; Collins, C. H.; Poppi, R. J. Direct determination of diclofenac in pharmaceutical formulations containing B vitamins by using UV spectrophotometry and partial least squares regression. *J. Pharm. Biomed. Anal.* **2004**, *36*, 743–749.

(10) Laasonen, M.; Harmia-Pulkkinen, T.; Simard, C.; Räsänen, R.; Vuorela, H. Development and validation of a near-infrared method for the quantitation of caffeine in intact single tablets. *Anal. Chem.* **2003**, *75*, 754–760.

(11) Braga, J. W. B.; Poppi, R. J. Figures of merit for the determination of the polymorphic purity of carbamazepine by infrared spectroscopy and multivariate calibration. *J. Pharm. Sci.* **2004**, *96*, 2124–2134.

(12) Lachenmeier, D. W. Rapid screening for ethyl carbamate in stone-fruit spirits using FTIR spectroscopy and chemometrics. *Anal. Bioanal. Chem.* **2005**, *382*, 1407–1412.

(13) Kays, S. E.; Barton, F. E. Near-infrared analysis of soluble and insoluble dietary fiber fractions of cereal food products. *J. Agric. Food Chem.* **2002**, *50*, 324–329.

(14) Heise, H. M.; Damm, U.; Lampen, P.; Davies, A. N.; McIntyre, P. S. Spectral variable selection for partial least squares calibration applied to authentication and quantification of extra virgin olive oils using Fourier transform Raman spectroscopy. *Appl. Spectrosc.* **2005**, *59*, 1286–1294.

(15) Morimoto, S.; McClure, W. F.; Crowell, B.; Stanfield, D. L. Near infrared technology for precision environmental measurements: Part 2. Determination of carbon in green grass tissue. *J. Near Infrared Spectrosc.* **2003**, *11*, 257–267.

(16) Smolders, R.; De Coen, W.; Blust, R. An ecologically relevant exposure assessment for a polluted river using an integrated multivariate PLS approach. *Environ. Pollut.* **2003**, *132*, 245–263.

(17) Mecozzi, M. Estimation of total carbohydrate amount in environmental samples by the phenol-sulphuric acid method assisted by multivariate calibration. *Chemom. Intell. Lab. Syst.* **2005**, *79*, 84–90.

(18) Rodriguez, A. M. G.; Torres, A. G.; Pavon, J. M. C.; Ojeda, C. B. Simultaneous determination of iron, cobalt, nickel and copper by UV-visible spectrophotometry with multivariate calibration. *Talanta* **1998**, *47*, 463–470.

(19) Lewis, C. B.; McNichols, R. J.; Gowda, A.; Cotez, G. L. Investigation of near-infrared spectroscopy for periodic determination of glucose in cell culture media in situ. *Appl. Spectrosc.* **2000**, *54*, 1453–1457.

(20) Escandar, G. M.; Damiani, P. C.; Goicoechea, H. C.; Olivieri, A. C. A review of multivariate calibration methods applied to biomedical analysis. *Microchem. J.* **2006**, *82*, 29–42.

(21) The American Society for Testing and Materials (ASTM), Practice E1655-00, ASTM Annual Book of Standards, West Conshohocken, PA, 2000.

(22) Near-infrared Spectrophotometer, Chapter 1119, United States Pharmacopoeia USP28NF23, 2005, 2691–2695.

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

(24) Kennard, R. W.; Stone, L. A. Computer aided design experiments. *Technometrics* **1969**, *11*, 137–148.

(25) Burns, D. A.; Ciurczak, E. W. *NIR Analysis of Polymers. In Handbook of Near-Infrared Analysis*; Marcel Dekker: New York, 2001; 659.

(26) Wise, B. M.; Gallagher, N. B. *PLS Toolbox 4.0 for Use with MATLAB*; Eigenvector Research: Wenatchee, WA, 2006.

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

(28) Geladi, P.; Kowalski, B. R. Partial Least-squares regression–A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.

(29) Brereton, R. G. Introduction to multivariate calibration in analytical chemistry. *Analyst* **2000**, *125*, 2125–2154.

(30) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York, 1989.

(31) Walczak, B.; Massart, D. L. Multiple outlier detection revisited. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 1–15.

(32) Fernández Pierna, J. A.; Wahl, F.; Noord, O. E.; Massart, D. L. Methods for outlier detection in prediction. *Chemom. Intell. Lab. Syst.* **2002**, *63*, 27–39.

(33) Van Der Voet, H. Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *J. Chemom.* **1999**, *13*, 195–208.

(34) Riu, J.; Rius, F. X. Assessing the accuracy of analytical methods using linear regression with errors in both axes. *Anal. Chem.* **1996**, *68*, 1851–1857.

(35) Lorber, A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* **1986**, *58*, 1167–1172.

(36) Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué, R.; Kalivas, J. H.; Mark, H. Uncertainty estimation and figures of merit for multivariate calibration. *Pure Appl. Chem.* **2006**, *78*, 633–661.

(37) Mandel, J.; Stiehler, R. D. Sensitivy–A criterion for the comparison of methods of test. *J. Res. Natl. Bur. Stand.* **1954**, *53*, 155–159.

(38) Rodriguez, L. C.; Campaña, A. M. G.; Linares, C. G.; Ceba, M. R. Estimation of performance characteristics of an analytical method using the data set of the calibration experiment. *Anal. Lett.* **1993**, *26*, 1243–1258.

(39) Faber, N. M.; Song, X.-H.; Hopke, P. K. Sample-specific standard error of prediction for partial least squares regression. *Trends Anal. Chem.* **2003**, *22*, 330–334.

(40) Currie, L. A. Nomenclature in evaluation of analytical methods including detection and quantification capabilities. *Pure Appl. Chem.* **1995**, *67*, 1699–1723.

(41) Boqué, R.; Rius, F. X. Multivariate detection limits estimators. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 11–23.

(42) Boqué, R.; Faber, N. M.; Rius, F. X. Detection limits in classical multivariate calibration models. *Anal. Chim. Acta* **2000**, *431*, 41–49.

(43) Boqué, R.; Larrechi, M. S.; Rius, F. X. Multivariate detection limits with fixed probabilities of error. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 397–408.

(44) Faber, K.; Kowalski, B. R. Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values. *Appl. Spectrosc.* **1997**, *51*, 660–665.

Received for review May 25, 2007. Revised manuscript received August 8, 2007. Accepted August 9, 2007.

JF071538S