Prediction of Lignin Content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS)

Camila Assis[1], Rachel S. Ramos[2], Lidiane A. Silva[2], Volmir Kist[2], Márcio H.P. Barbosa[2], and Reinaldo F. Teófilo[1]

Abstract

The building of multivariate calibration models using near-infrared spectroscopy (NIR) and partial least squares (PLS) to estimate the lignin content in different parts of sugarcane genotypes is presented. Laboratory analyses were performed to determine the lignin content using the Klason method. The independent variables were obtained from different materials: dry bagasse, bagasse-with-juice, leaf, and stalk. The NIR spectra in the range of 10 000–4000 cm^{-1} were obtained directly for each material. The models were built using PLS regression, and different algorithms for variable selection were tested and compared: *i*PLS, *bi*PLS, genetic algorithm (GA), and the ordered predictors selection method (OPS). The best models were obtained by feature selection with the OPS algorithm. The values of the root mean square error prediction (RMSEP), correlation of prediction (R_P), and ratio of performance to deviation (RPD) were, respectively, for dry bagasse equal to 0.85, 0.97, and 2.87; for bagasse-with-juice equal to 0.65, 0.94, and 2.77; for leaf equal to 0.58, 0.96, and 2.56; for the middle stalk equal to 0.61, 0.95, and 3.24; and for the top stalk equal to 0.58, 0.96, and 2.34. The OPS algorithm selected fewer variables, with greater predictive capacity. All the models are reliable, with high accuracy for predicting lignin in sugarcane, and significantly reduce the time to perform the analysis, the cost and the chemical reagent consumption, thus optimizing the entire process. In general, the future application of these models will have a positive impact on the biofuels industry, where there is a need for rapid decision-making regarding clone production and genetic breeding program.

Introduction

In the last few decades, the world has seen a gradual increase in interest in the development of biofuels due to greater concern about the environmental problems caused by the burning of fossil fuels. Environmental damage caused by increasing the concentration of the gases responsible for the greenhouse effect, and the depletion of the easily extracted oil reserves, have encouraged the use of renewable inputs.[1,2] The goal is to reduce the consumption of fossil fuels such as oil, coal, and natural gas. In this context, economic, social, and environmental sustainability across the world will depend heavily on research into alternative energy sources. In Brazil, research into biofuels is crucial to maintain the current position of Brazilian energy autonomy, beyond the interest in producing surpluses for export.[3] Renewable biofuels must be understood as a way to reduce the dependence on oil, reduce emissions

[1]Department of Chemistry, Universidade Federal de Viçosa. 36570-900, Viçosa, Minas Gerais, Brazil
[2]Department of Plant Science, Universidade Federal de Viçosa, 36570-900, Viçosa, Minas Gerais, Brazil

Corresponding author:
Reinaldo F. Teófilo, Department of Chemistry, Universidade Federal de Viçosa. 36570-900, Viçosa, MG, Brazil.
Email: rteofilo@gmail.com

of greenhouse gases, and encourage development in the agricultural sector.[4]

Considering this need, lignocellulosic biomass is an important alternative energy source to oil. Sugarcane (*Saccharum* spp.) has about two-thirds of its mass in lignocellulosic material, a complex polymer consisting of cellulose, hemicellulose, and lignin.[5,6] Besides sugarcane juice, which is a raw material for the production of sugar and ethanol, the industrialization process also produces bagasse and straw, previously classified as waste agricultural production. This highly energy-dense biomass has the potential to be converted into chemicals and biofuels, liquid or gaseous, through biochemical and thermochemical processes.[1,7] This material contains abundant energy that is not used efficiently by the current technologies, therefore representing a huge potential for energy production.[8–11]

The possible applications of biomass from sugarcane bagasse as an energy source include either burning it directly or obtaining biofuels or chemicals. Therefore, clones with the highest-productivity lignocellulosic biomass are necessary for this purpose. The quantification of these biopolymers and sugars in the cane is needed to aid decision-making in relation to the production of clones and genetic breeding program.[5,12] Thus, a large amount of wet chemical analysis is necessary. Despite their accuracy, precision, and robustness, these analyses cannot be applied in an industrial or commercial setting or in extensive research, since they are expensive, extremely time-consuming, and environmentally harmful due to the high consumption of reagents and the necessary disposal of the polluting products.

The use of chemometric methods to extract information from multivariate data, such as spectra, can significantly reduce the time, cost, and environmental impact of chemical analysis.[5,13–15] In this sense, the use of spectroscopy in the region of the near infrared (NIR), which covers the range of 12 800–4000 cm^{-1}, has been widely and successfully applied in the nondestructive determination of lignin composition in various agricultural products.[16–27] However, few studies have been conducted with spectra obtained directly from the leaf and/or stem, which require no sample preparation procedure. Most of the work involves the use of bagasse, which needs to be dried, ground, and sieved before obtaining the spectra. Rapid and nondestructive determination is highly desirable in situations with a large number of samples and a short period of time for decision making. Besides, to the best of our knowledge, this is the first work to perform this spectroscopic analysis directly on sugarcane material.

Near-infrared spectroscopy is a rapid technique (< 1 min to obtain the spectrum), non-invasive, suitable for use in production lines, and requires minimal sample preparation. In addition, NIR spectroscopy coupled with chemometric methods provides robust calibrations, i.e., the model parameters do not change significantly when new samples are added or removed from the calibration set.

The classical analysis of the lignin content is the Klason[28,29] method, which is an extremely time-consuming method, and requires considerable amounts of chemical reagents (mainly acidic), thus being impractical for implementation in large quantities of samples and especially for fast decision making. Hence, the aim of this study is to build and validate multivariate calibration models for predicting the lignin content directly in the cane or biomass, using NIR spectroscopy and chemometric methods, such as PLS regression. In this way, the models will be applied to perform estimates of the lignin content in different clones, to select the best crops for energy production. The contribution of this work is to use different parts (or forms of processing) of cane sugar, such as bagasse, bagasse-with-juice, and different parts of the stem and leaf. The main objective of the work is to build a multivariate model with high predictive capacity to lignin contents from sugarcane materials with minimum sample preparation. Thus, we aim to find simultaneously the best model and sampling approach to predict lignin. The goal is to reduce substantially the time required for sample preparation, in order to obtain quick results.

Experimental

Samples

Three hundred experimental sugarcane (*Saccharum spp.*) genotypes were supplied by the germplasm bank of the Sugarcane Genetic Breeding Program (PMGCA) from Universidade Federal de Viçosa, Viçosa, Minas Gerais (MG), Brazil. The plantation was set in May 2014 using rows 5 m in length in an experimental field in Viçosa, MG, Brazil (20°44′37″ latitude south, 42°50′38″ longitude west). Ten stalks per genotype, both with and without juice, were ground. Thus, a homogeneous and representative sample of bagasse was crushed, using a Willy type grinder with 0.5 mm sieves, and ground, of each genotype and brought to the laboratory. Different stalks of the same genotype may not have the same lignin content, but a sampling pull from ten stalks would be representative. The aim, therefore, was to work with a representative sampling, considering variability of the fiber content and featuring morphoanatomic plants. Thus, it is assured that the calibration set was representative of the population for which the future predictions will be performed.

Chemical Analysis

The stalks were submitted to disintegration and homogenization. An aliquot of 500 g was submitted to a hydraulic press to obtain the extracted juice. After the juice extraction, the residue was dried in an oven for 24 h at 45 °C. It was then crushed and separated through a 0.4 mm mesh. The extractive samples (2 g dry weight) were extracted successively in water and ethanol in a Soxhlet extraction unit for a period of 10 h and 5 h, respectively. Finally, the

samples were taken to a greenhouse (65 °C) for drying. This procedure is necessary to remove compounds that are not part of the structural biomass and may interfere in analysis. According to the Klason method,[28,29] approximately 0.3 g of the material was placed in 100 mL tubes and 3.00 ± 0.01 mL of 72% sulfuric acid (H_2SO_4) added and shaken until the sample reached complete homogeneity. The tubes were sampled in a bath of water at 30 ± 3 °C for 2 h, stirring the samples with a glass stick every 10 min, without withdrawing the sample from the bath. The sample was then diluted with 84.00 ± 0.04 mL of deionized water to reduce the concentration of sulfuric acid. Samples were autoclaved at 121°C, 15 psi for 1 h. Then, the acid-insoluble fraction was allowed to stand before being filtered. Vacuum filtration of the hydrolysis solutions after autoclaving was carried out through the previously weighed sintered glass crucibles (Laborglas 2D 50 mL). The crucibles with lignin and acid-insoluble residue pellets were placed in the stove for a period of 5 h (65 °C), until weight constant, and the Klason lignin content determined gravimetrically. The mass of the insoluble dry residue represents the lignin content. Although the National Renewable Energy Laboratory (NREL) procedure recommended the ash subtraction step, this procedure was not performed because we considered that the amount of ash was small and would not have large variations between the samples.

Spectrophotometric Analysis

For spectrophotometric analysis, the samples used for the determination of the lignin content were selected. The spectra were obtained in different locations and conditions of the sample: dry bagasse, bagasse-with-juice, stalk, and leaf. The goal at this stage was to find the best material to perform the prediction quickly and with high accuracy. For the bagasse-with-juice, the stalks of the genotypes were submitted to disintegration and homogenization, and then they were ground together with the juice and stored. The dry bagasse spectra were obtained with the same material used for the determination of lignin (Figure 1: D, C1). In addition, the spectra were taken from the leaf blade sheet +3 of each genotype, according to the Kuijper system.[30] To obtain the spectra, the middle third leaves were used, excluding the midrib (Figure 1: D, D1). The leaves were removed, individually conditioned in plastic bags and frozen at −80°C. At the time of obtaining the spectra, the leaves were thawed. Prior to the process of obtaining the spectra, spectra were drawn on different parts of the leaves and it was realized that there was no great variation. Thus, a region close to the groove was chosen and all samples followed this pattern (Figure 1: D, D1). The spectra were taken from one to six months after harvest.

For the stalk, the samples were cut into two different parts, i.e., the upper part (third upper stalk) and the middle region (third middle stalk), and stored in a freezer at −80 °C. To obtain the spectra, a longitudinal cut was made and the spectrum was obtained from the inner part (Figure 1: A, B, B1, and B2).

A total of 378 analyses were performed on dry bagasse, 232 on bagasse-with-juice, 256 on leaf, 221 on middle stalk, and 223 on top stalk. These differences in the number of
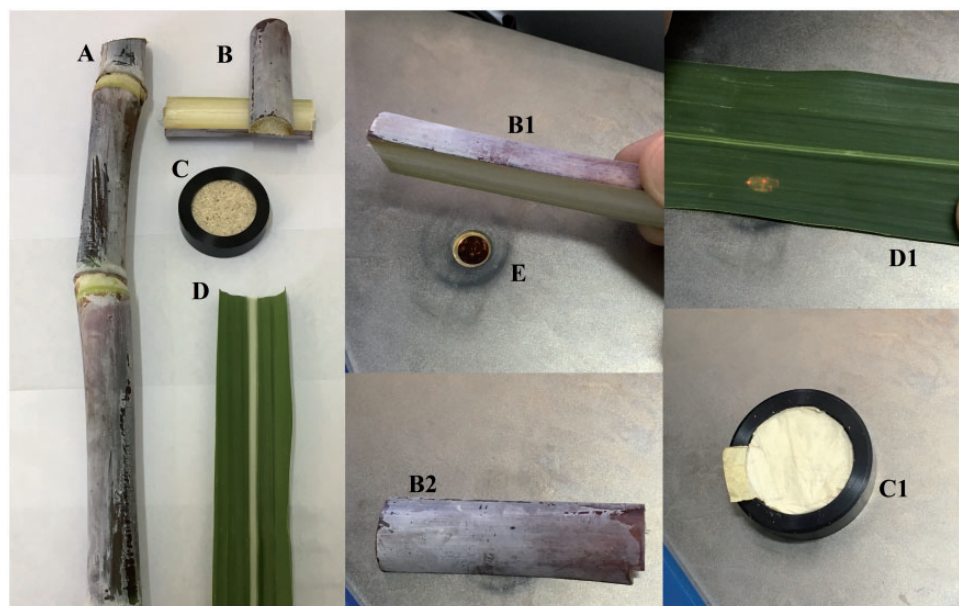

Figure 1. Samples and way of obtaining the spectra. (A, B) Stalk and stalk cut in longitudinal section, (C) accessory used to contain the bagasse dry and with broth, (D) cut leaf, (E) input of reflectance-integrating sphere, (B1, B2) way of obtaining the stalk spectra, (C1) way of obtaining the bagasse spectra, (D1) way of obtaining the leaf spectra.

samples occurred because some samples were lost in collection, preparation, and storage.

The y-vector (dependent variable) was built with the same samples for all experimental conditions. The X matrix (independent variables) was different for each condition (leaf, stalk, dry bagasse, and bagasse-with-juice), but the y vector was the same.

The NIR spectra were obtained using a Fourier transform NIR (FT-NIR) 660 spectrometer (Agilent Technologies), with a reflectance-integrating sphere accessory, from PIKE Technologies. The range investigated was 10 000–4000 cm^{-1} with an increment of 4 cm^{-1}. The spectra were obtained using the software Pro Resolutions Version 5.1, storing information such as log (1/R), where R is the reflectance collected. For each sample, a total of 64 scans were performed and the average was stored.

Multivariate Analytical Calibration

The spectra were imported by the software Matlab 2016a (The Mathworks Inc.). For model building, the PLS regression was used. Algorithms for feature selection, such as genetic algorithm (GA),[31,32] iPLS,[33] and OPS,[34] were applied to improve the models and verify the more important spectral regions for regression. The algorithms for the variable selection using OPS, model-building, and validation are homemade, available at http://www.deq.ufv.br/chemometrics. All the calculations were performed in Matlab 2016a. The computational package used for the calculations of iPLS was iToolbox.[33] PLS-Toolbox 6.7 for Matlab was used to perform the GA. The GA was carried out using the following optimized parameters: population (54); generations (300); mutation rate (0.008); window width (1); convergence (80); startup (50); and cross-over (2).

The Kennard and Stone algorithm[35] was used to separate the sets for calibration and prediction. Thus, it sought samples representative of the population distributed throughout the set space. For all models, 40 samples were selected for the prediction set.

Models Evaluation and Figures of Merit Calculation

The quality of the models was evaluated by the root mean square error (RMSE), which was calculated according to Eq. 1. R, the correlation coefficient, was calculated using Eq. 2:

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

$$R = \frac{\sum_{i=1}^{I_m}(\hat{y}_i - \hat{\mathbf{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I_m}(\hat{y}_i - \hat{\mathbf{y}})^2 \sum_{i=1}^{I_m}(y_i - \bar{y})^2}} \tag{2}$$

where y is the experimental value, $\hat{y}$ and $\hat{\mathbf{y}}$ are the scalar and vector of the estimated values, respectively, $\bar{y}$ is a scalar of the mean values in **y** and I_m is the number of samples. The number of factors in the model was determined by internal validation (cross-validation [CV]), applying the random method with splits equal to 15 and iterations equal to 5. In cross-validation, when I_m is the number of samples in the calibration set (training), the error and correlation coefficients are denoted RMSECV and R_{cv}, respectively. When I_m is the number of predicting samples (P), the error and correlation coefficients are denoted RMSEP and R_p, respectively.[34]

The performance of model was also evaluated by calculating the ratio of performance to deviation (RPD), calculated as the standard deviation (SD) of the cross-validation set divided by the RMSECV.[36] According to Chang et al.,[36] the prediction results can be divided into three classes, with (1) good predictions having RPD > 2; (2) predictions with potential having RPD between 1.4 and 2.0; and (3) unreliable predictions having RPD < 1.4. The percentage of the relative error (%RE), ratio between the absolute error and the measured value, was also calculated and it is an important parameter to verify the accuracy of the built models.

Bias is defined as the arithmetic mean of the prediction errors, and must be near to zero. This parameter can be obtained by Eq. 3:

$$bias = \frac{1}{I_m}\sum_{i=1}^{I_m}(y_i - \hat{y}_i) \tag{3}$$

where y is the true value, $\hat{y}$ is the estimated value and I_m is the number of predictions. A t-test for evaluating the statistical significance of the bias was performed.

In addition, figures of merit were calculated for the best model.

1. Selectivity: this is a measure of the degree of overlap between the signal of the species of interest and interferents in the sample.[5,37] The selectivity was determined according to Eq. 4:

$$SEL = \frac{nas}{\|\mathbf{x}\|} \tag{4}$$

where nas is the scalar value of the net analytical signal for a given sample and $\|\mathbf{x}\|$ is the norm of the instrumental signal vector.

2. Sensitivity: the sensitivity, in the inverse calibration model, is proportional to the regression vector $\|\mathbf{b}\|$,[5,38,39] according to Eq. 5:

$$SEN = \frac{1}{\|\mathbf{b}\|} \tag{5}$$

3. Analytical sensitivity (γ): this is defined as the ratio between the sensitivity and the standard deviation of the reference signal, according to Eq. 6:[5,37]

$$\gamma = \frac{\text{SEN}}{\|\delta_x\|} \tag{6}$$

The inverse of this parameter (γ^{-1}) represents the smallest difference in the concentration of samples that can be distinguished by the method.[38,40]

4. Limit of detection (LOD) and limit of quantification (LOQ): in general, LOD and LOQ are calculated according to Eqs. 7–8. [39]

$$LOD = 3 \cdot \|\delta_r\| \cdot \|\mathbf{b}\| \tag{7}$$

$$LOQ = 10 \cdot \|\delta_r\| \cdot \|\mathbf{b}\| \tag{8}$$

where $\|\delta_r\|$ is the norm of the standard deviation vector of the columns of the noise matrix and $\|\mathbf{b}\|$ is the norm of the regression vector. In this work, an alternative method for determining the LOD, based on one of the proposed IUPAC methods[41] was applied, according to Eq. 9:

$$LOD = \frac{\hat{a} + e_a}{\hat{b}} \tag{9}$$

where $\hat{a}$ is the intercept of the regression equation of the measured and predicted values, $\hat{b}$ is the regression coefficient, and e_a is the variance of the intercept.[42]

Results and Discussion

Near-Infrared Spectra

The NIR spectra of the samples dry bagasse, bagasse-with-juice, leaf, middle stalk, and top stalk are presented in Figure 2.

The interpretation of the spectra presented in Figure 2 is facilitated by using the pure spectra of the main components present in biomass sugarcane, i.e., lignin, cellulose, and xylose. The spectra of the pure components are shown in Figure 3.
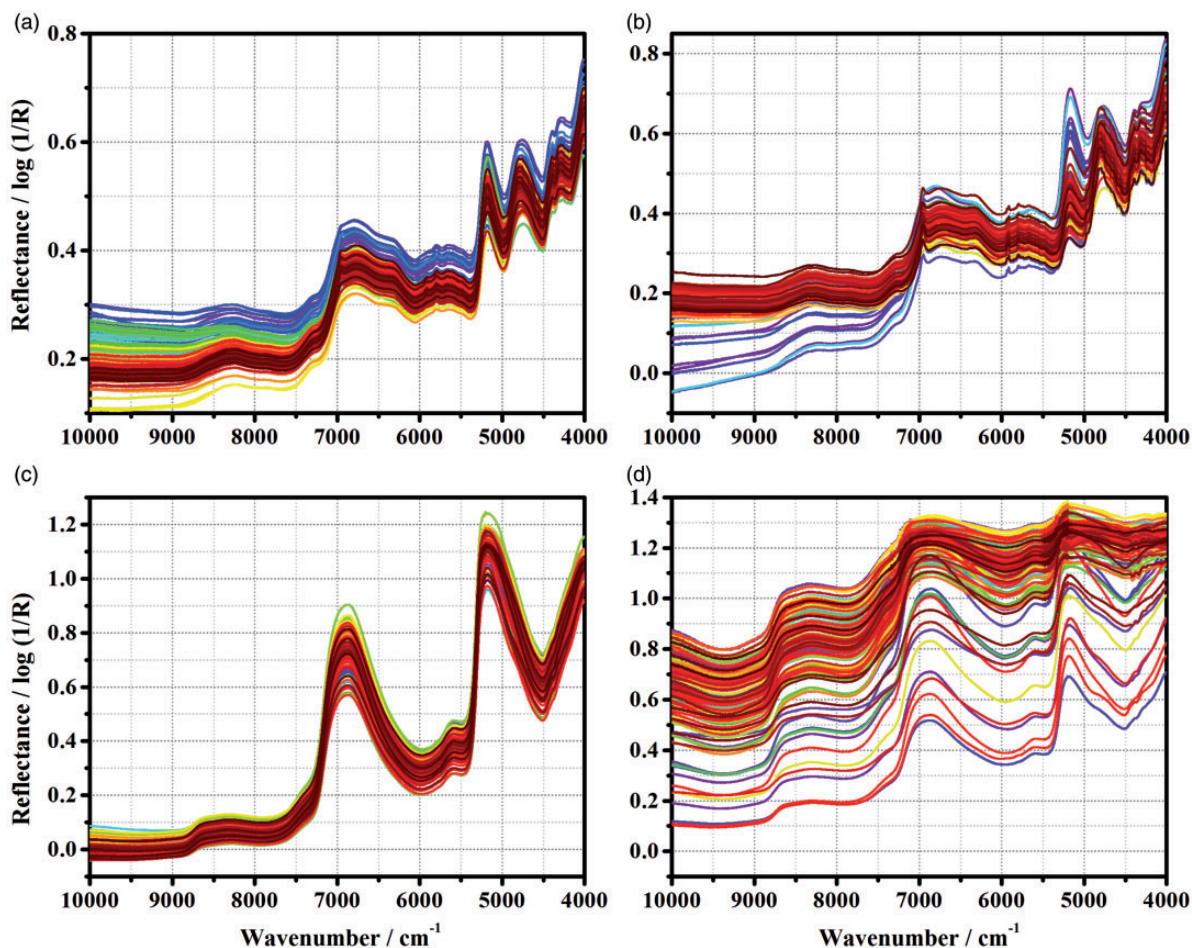

Figure 2. Near-infrared spectra for (a) dry bagasse, (b) bagasse-with-juice, (c) leaf, and (d) stalks.

Although lignin is a polymer that does not have a standard structure, the NIR spectra of the extracted and purified compound can be used as a spectral reference. This is possible because most functional groups present in the lignin structures are common, such as phenols, hydroxyl groups, carbonyl groups, carbon–carbon bonds, and carbon–hydrogen bonds. Note that the NIR spectrum of pure lignin has information on these groups. The region at 4000–6000 cm^{-1} contains C–H and C=O stretching, and stretching combinations of C–C and C–O–C. The region at 5000–5500 cm^{-1} contains OH stretching, and stretching combinations of C–O (3υ). For water, O–H stretches occur at 5100 cm^{-1} and 7000 cm^{-1}. These is differentiated from the phenolic O–H stretch, 10 500 cm^{-1}, outside your bandpass, 7000 cm^{-1}, obscured by water and 5200 cm^{-1}, which can see in the extracted sample, but not the spectra. Note that aromatic C–H stretch is present at 5995 cm^{-1} (Figure 3). The continuous increase above 8000 to 10 000 cm^{-1} is characteristic of polymeric OH combinations (2υ).[43]

For cellulose, the most important regions are CH stretching and the stretching combinations of C–C and C–O–C at 4000 cm^{-1}. Between 4019 and 4386 cm^{-1}, the C–H and C–C stretching combinations are found.

The O–H and C–H stretching for cellulose are found at 4405 cm^{-1}. The stretching combinations of O–H and C–O bonds are found at 4762 cm^{-1}. The polymeric stretching modes of OH (2υ) are found at 4785 cm^{-1}. OH stretching and stretching combinations of C=O (3υ) occur at 5495 cm^{-1}. The polymeric OH stretching combinations (2υ) are found at 6897 cm^{-1}.

Xylose is the main hemicellulose monomer in the bagasse of sugarcane. It is observed from Figure 3 that xylose has regions in common with cellulose.[43]
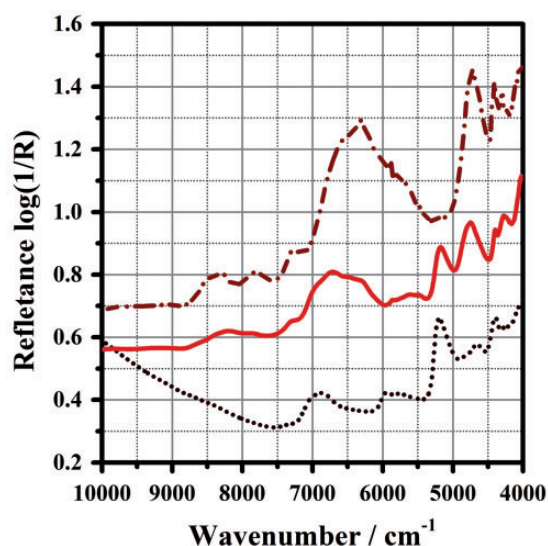

Figure 3. Spectra of pure solid compounds: D-(+)-xylose (dash dot line), alkaline lignin (dot line), and cellulose (solid line).

From Figure 2, note that the spectra of dry bagasse and dry bagasse-with-juice include the spectral information contained in the spectra of cellulose, lignin, and xylose.

From the leaf spectra, information is observed on phenolic groups around 5950 cm^{-1} and the presence of water in the strong peak at 5220 cm^{-1}, which is a combination of the asymmetric stretching and bonds of water molecules. At 7020 cm^{-1} another strong peak of the first overtone occurs, and at 8640 cm^{-1} the isosbetic point of water is observed. Therefore, the leaf spectra display marked interference from water.

The stalk spectra have a strong systematic shift; however, water, cellulose, and lignin information can be detected in the region 4000–6000 cm^{-1}.

Building of Models

For lignin, the obtained values were in the range of 16.23–33.85% (w/w). The mean value was 23.44% and the pooled standard deviation was 0.96; therefore, the relative standard deviation was 4.9%. This result confirms the high precision of the gravimetric method.

For all data sets, the best transformation method was smoothing, second derivative, followed by mean centering.

The statistical parameters calculated for all models are presented in Table 1.

Analyzing Table 1, it can be noticed that, in general, the number of latent variables required to build the models was 10, except for the middle part of the stalk. This number, although high, is justified since, in most cases, the spectra were obtained in biological samples where the complexity of the system is considerable high, implying in a larger number of latent variables to explain the model. However, the models in the literature, for the most part, have long processes of preparation of samples such as: drying, grinding, grinding, purification, among others, requiring a lower number of latent variables.

Furthermore, it can be observed that for all models, the bias was not significant, indicating that there is no trend in the residues and, therefore, the systematic error of the model can be disregarded.

Comparison and Interpretation of Models

The high *RMSEP* and *RMSECV* errors of the model containing all variables confirm the necessity of performing the selection of variables in the multivariate calibration. The worst *RMSEP* was obtained when performing the selection using *i*PLS and *bi*PLS. The reason for this poor performance can be linked to the fact that lignin is a complex molecule and, consequently, is active in different regions of the NIR (Figure 3). As the *i*PLS and *bi*PLS perform selection by intervals, they do not work well when the information is spread across the spectrum.

Table 1. Statistical parameters for dry bagasse, bagasse-with-juice, leaf and middle, and upper third of the stalk models.

	Models dry bagasse					Models bagasse-with-juice				
	Full	OPS	iPLS	biPLS	GA	Full	OPS	iPLS	biPLS	GA
nLV	10	10	10	10	10	10	10	10	10	10
hOPS	–	24	–	–	–	–	19	–	–	–
nVars	1038	445	346	519	387	1038	265	346	520	352
RPD	1.84	2.87	1.84	1.64	2.7	1.42	2.77	1.68	1.51	2.57
RMSECV	1.42	0.89	1.43	1.58	0.95	1.31	0.71	1.2	1.22	0.74
RMSEC	0.53	0.44	0.76	0.63	0.5	0.3	0.32	0.33	0.38	0.32
Rc	0.98	0.99	0.96	0.97	0.98	0.99	0.99	0.99	0.98	0.99
Rcv	0.87	0.94	0.86	0.83	0.94	0.79	0.94	0.83	0.81	0.93
RMSEP	0.98	0.85	1.39	1.08	0.88	1.16	0.65	1.48	1.37	0.67
Rp	0.93	0.97	0.87	0.93	0.96	0.86	0.94	0.8	0.81	0.95
Bias	−0.052	−0.0066	0.018	−0.0004	−0.0033	0.03	0.003	−0.093	−0.044	0.03
%RE	3.22	2.82	4.9	3.48	2.7	3.17	1.94	3.31	3.74	2.35

	Models – leaf					Models – middle stalk				
	Full	OPS	iPLS	biPLS	GA	Full	OPS	iPLS	biPLS	GA
nLV	10	10	10	10	10	8	8	8	8	8
hOPS		25					16			
nVars	1038	305	346	519	324	1038	205	346	519	338
RPD	1.41	2.56	1.15	1.17	2.72	1.77	3.24	1.38	1.31	2.16
RMSECV	1.32	0.76	1.84	1.55	0.72	1.21	0.63	1.76	1.57	0.87
RMSEC	0.43	0.35	0.54	0.69	0.35	0.3	0.31	0.33	0.38	0.35
Rc	0.98	0.98	0.97	0.94	0.99	0.99	0.99	0.99	0.98	0.99
Rcv	0.78	0.93	0.62	0.69	0.94	0.84	0.96	0.71	0.72	0.91
RMSEP	0.77	0.67	0.99	1.01	0.64	0.73	0.61	0.93	0.9	0.62
Rp	0.93	0.96	0.9	0.91	0.95	0.9	0.95	0.87	0.89	0.94
Bias	−0.015	0.017	0.003	−0.045	0.0043	0.017	0.018	−0.083	−0.0099	−0.039
%RE	2.62	2.47	3.47	3.27	2.3	2.52	1.97	2.82	2.83	2

	Models - upper third of the stalk				
	Full	OPS	iPLS	biPLS	GA
nLV	10	10	10	10	10
hOPS		15			
nVars	1038	300	346	519	357
RPD	1.47	2.33	1.21	1.43	2.55
RMSECV	1.43	0.89	2.1	1.58	0.83
RMSEC	0.29	0.3	0.37	0.38	0.33
Rc	0.99	0.99	0.98	0.98	0.99
Rcv	0.78	0.91	0.61	0.74	0.93
RMSEP	0.58	0.58	0.65	0.65	0.57
Rp	0.96	0.96	0.95	0.95	0.96
Bias	0.019	0.045	−0.088	−0.065	0.051
%RE	2.01	1.94	2.3	1.98	1.91

In contrast, a comparison between the GA and OPS algorithms shows that there was no significant difference in the *RMSEP* and *RMSECV* values. Additionally, all the models built with GA and OPS obtained *RPD* values greater than 2, indicating that the prediction is reliable. However, higher *RPD* values were observed for the OPS. The OPS algorithm also selected fewer variables, producing a more robust model which will last longer and be easier to maintain. In this context, it is important to highlight the computational time of both methods. While the OPS performed the calculations in minutes, the GA took hours. Additionally, significant time was consumed to optimize the parameters used in the GA. Thus, the model using the OPS algorithm can be considered more efficient and simpler. The results presented in Table 1 indicate that the OPS model shows a high ability to predict the lignin content in sugarcane bagasse-with-juice, with high accuracy in relation to the reference method. The variables selected for the OPS model are presented in Figure 5a.

For dry bagasse, the highest relative error found (in absolute values) was 5.57% and the lowest was 0.03% (calibration set). For the prediction set, the values were in the range (in absolute value) of 7.70–0.16%. For dry bagasse, as the error values are relatively small, it can be concluded that the model is able to perform lignin prediction in sugarcane dry bagasse with high accuracy. Measured versus predicted values plotted for the calibration and prediction sets are presented in Figure 4a.

For bagasse-with-juice, the biggest relative error found (in absolute value) was 4.28%, and the lowest 0.01% (calibration set). The prediction-set values were in the range (in absolute value) 9.02–0.002%. As the relative error values are small, it can be concluded that the model can predict lignin in sugarcane bagasse-with-juice with high accuracy. The measured versus predicted values are shown in Figure 4b.

Regarding the leaf, different studies are found in the literature relating the NIR spectra to various leaf properties.[44,45] In these studies, the leaves were dried and crushed, in contrast to the present work, in which the spectrum was obtained directly from the green leaf, with no sample preparation procedure. Menesatti et al.[46] evaluated the nutritional properties of oranges by obtaining the spectra directly from the dry leaf. In that case, there was a direct correlation between the properties of a leaf and its spectrum. The current work is dedicated to correlating the leaf spectrum (obtained without any sample preparation) with the bagasse lignin content, and therefore provides an indirect correlation.

This is possible because the leaf is known to contain compounds that may be indirectly related to the content of lignin in bagasse. The major advantage of this approach is that the spectra can be obtained on-site, directly from the leaf, without the need to harvest. This result suggests the possibility of monitoring the concentration of lignin during plant growth.

The relative error ranges obtained are (in absolute values) 8.94–0.04% (calibration set) and 9.06–0.07% (prediction set). This shows that it is feasible to realize the prediction of the lignin content in sugarcane through the leaf spectrum, reducing significantly the time spent on analysis and the consumption of chemical reagents.

Figure 4c contains the measured versus predicted values for set calibration and prediction.

The purpose of building a model using the stalk spectrum is the ease of obtaining the spectrum. In this work, we have prepared two different parts of the stalk (top and middle) to build the models. To perform the analysis, the sample preparation is minimal. Differently from the dry bagasse, which requires grinding, sieving and drying of the sample, the stalk is simply cut and taken to the laboratory for analysis. Thus, one can obtain hundreds of results in a single day, eliminating considerable time dedicated to sample preparation.

For the model obtained with the middle part of the stalk, using the OPS algorithm, the error range (in absolute values) was 4.66–0.002%. For the prediction set, the range is 12.05–0.08%. For the top of the stem, the relative error values found for the calibration set and prediction were, respectively, 0.009–6.18% and 6.53–0.02%. These figures confirm the model's ability to predict lignin in sugarcane with high accuracy.

Figure 4d and 4e shows the measured versus predicted values for these sets of data (calibration and prediction).

In general, comparing the models statistically (F test, 99%), for most cases there is no significant difference between the models using the different methods of variable selection (except for *i*PLS and *bi*PLS which worsened the performance of the models). In all cases, the OPS and GA algorithms do not present significant difference. Although RMSEC and R values are very close, this fact does not exclude the need to use variable selection: the models become more interpretive, the possibility of identifying variables with greater predictive capacity and a considerable decrease of relative error values (prediction set).

The variables selected for the OPS and GA models are presented in Figure 5.

From Figure 5, note that for dry bagasse, no significant differences can be observed in the variables selected by the various selection methods used.

The OPS and the GA selected 445 and 387 variables, respectively. The difference of 58 variables can be observed in the unselected regions shown in Figure 5f.

These results indicate that ground, sieved, and homogeneous dry bagasse facilitates access to lignin by NIR. The selection of variables across the spectrum is justified, since the lignin spectra (Figure 2) contain information throughout the region investigated.

For bagasse-with-juice, OPS and GA selected 265 and 352 variables, respectively. The difference of 87 variables can be observed in the regions not selected in Figure 4b.
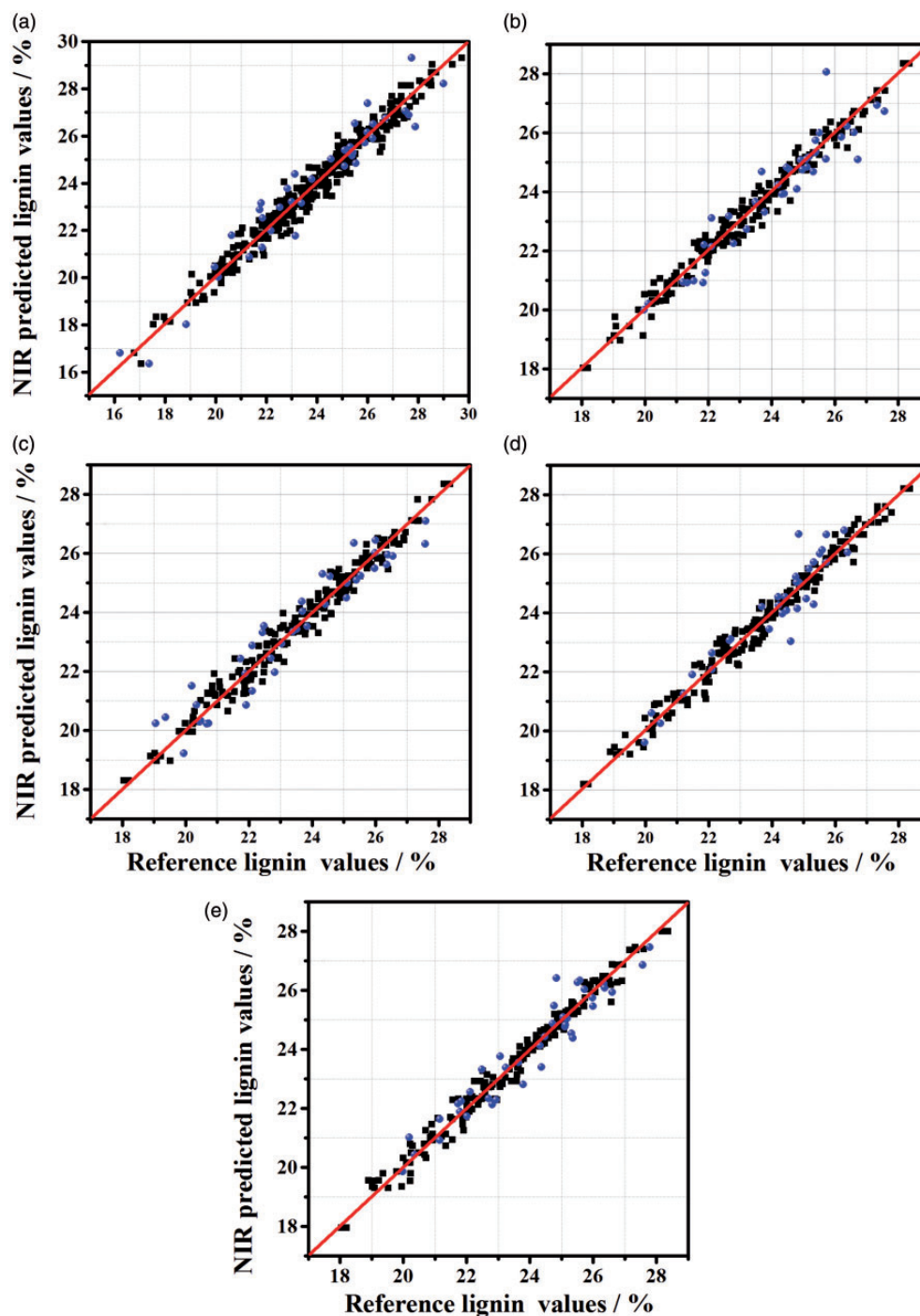
Figure 4. Reference lignin values versus NIR-predicted lignin values for (a) dry bagasse, (b) bagasse-with-juice, (c) leaf, (d) middle stalk, and (e) upper third of the stalk.

In this case, the OPS showed better results and chose a smaller number of variables than the GA.

From Figure 5b, it is observed that the regions at 4300–6300 cm^{-1} are those with large interference from cellulose and xylose, which justifies the selection of few variables by the OPS in this region. Similar behavior can be observed for the leaf and stalk.

On the other hand, the regions at 4000–4200 cm^{-1} and 8500–10000 cm^{-1} were always selected, since in these regions lignin has bands that differentiate it from the other compounds present.

In most cases, the OPS algorithm selected fewer variables than GA. Genetic algorithm, on the other hand, failed to select variables for bagasse-with-juice at
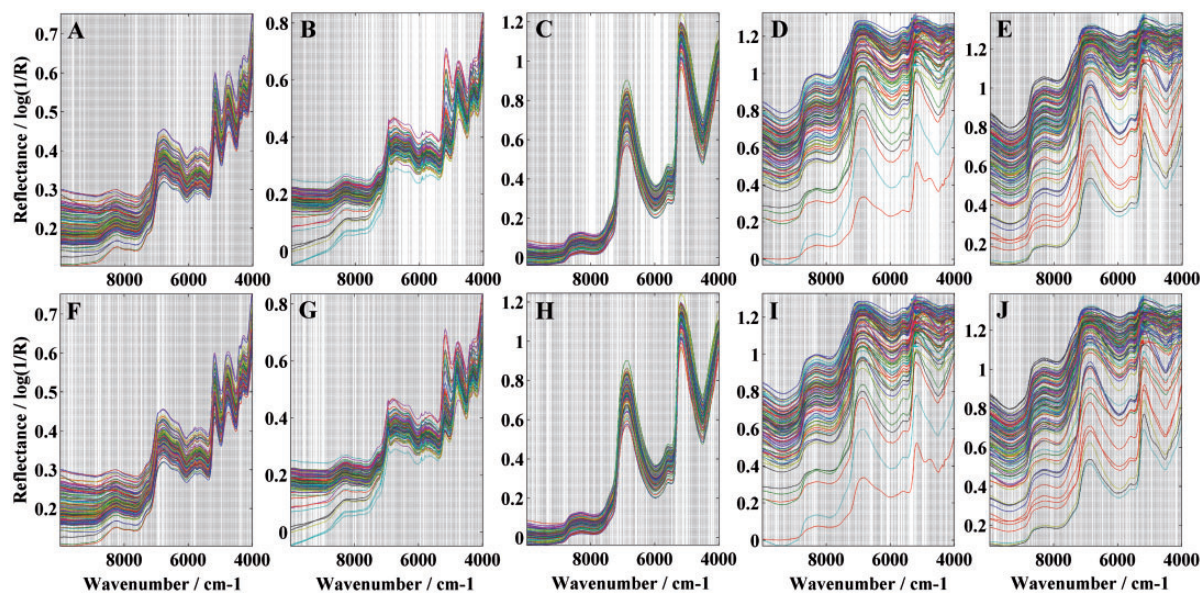
Figure 5. Feature selection using OPS (a–e) and GA (f–j). (a–f) Dry bagasse, (b–g) bagasse-with-juice, (c–h) leaf, (d–i) middle stalk, and (e–j) upper third of the stalk. Vertical lines indicate the selected variables.

Table 2. Figures of merit calculated for the OPS–PLS model and for the model with all variables (full model).

	Dry bagasse		Bagasse-with-juice		Leaf		Middle stalk		Upper stalk	
	OPS	Full	OPS	Full	OPS	Full	OPS	Full	OPS	Full
SEL	0.0503	0.0373	0.057	0.038	0.0571	0.0422	0.062	0.039	0.057	0.042
SEN.	1.22×10^{-5}	1.52×10^{-5}	1.80×10^{-5}	5.87×10^{-5}	3.78×10^{-5}	4.37×10^{-5}	2.00×10^{-4}	8.90×10^{-4}	7.6×10^{-5}	6.9×10^{-5}
γ	1.72	1.77	6.31	6.91	8.69	12.31	36.34	10.49	8.54	10.49
γ^{-1}	0.58	0.56	0.16	0.14	0.11	0.081	0.15	0.095	0.12	0.090
LOD	2.53	2.71	3.69	3.65	3.74	4.10	3.75	3.66	3.58	3.66
LOQ	7.67	8.22	11.18	11.07	11.33	12.43	11.38	11.1	10.84	11.1

$9000-10\,000\ \mathrm{cm}^{-1}$, which is a region in which the lignin has good selectivity.

Figures of Merit

Figures of merit for the OPS–PLS and full models are presented in Table 2. A comparison between the results shown in Table 2 indicates that the selectivity of the OPS model increased compared with the full model, which was expected given the variable selection.

The inverse of the analytical sensitivity, γ^{-1}, allows us to establish the smallest variation in concentration between samples that can be distinguished by the method. It is noted from Table 2 that this figure was higher for the OPS models.

The LOD and LOQ values did not change significantly from the OPS to the full model, indicating that for the data studied, OPS does not decrease the LOD and LOQ values.

Conclusion

The analysis of the statistical parameters indicates that both the OPS and GA algorithms decreased significantly the errors of cross-validation, prediction, and *RPD*. Note that the computational time of the OPS algorithm is considerably less than the GA. For this reason, the final models were all chosen using the algorithm OPS. The OPS method combined with PLS regression allows the building of more simple, interpretable, and predictive models. All the models showed good predictive capability and the methods are inexpensive, environmentally friendly, and extremely quick. As all models, after the selection of variables, presented satisfactory statistical parameters, a viable option for industrial application would be the use of stalks and leaves, which do not require any sample preparation.

Conflict of Interest

The authors report there are no conflicts of interest.

Funding

References

1. A. Demirbas. "Progress and Recent Trends in Biofuels". Prog. Energy and Combust. Sci. 2007. 33(1): 1–18.
2. A. Demirbas. "Biofuels Sources, Biofuel Policy, Biofuel Economy and Global Biofuel Projections". Energy Convers. Manage. 2008. 49(8): 2106–2116.
3. R.C.d. Cerqueira Leite, M.R.L. Verde Leal, L.A. Barbosa Cortez, W.M. Griffin, M.I. Gaya Scandiffio. "Can Brazil Replace 5% of the 2025 Gasoline World Demand with Ethanol?" Energy. 2009. 34(5): 655–661.
4. M.R. Schmer, K.P. Vogel, R.B. Mitchell, R.K. Perrin. "Net Energy of Cellulosic Ethanol from Switchgrass". Proc. Natl. Acad. Sci. 2008. 105(2): 464–469.
5. I.P. Caliari, M.H. Barbosa, S.O. Ferreira, R.F. Teofilo. "Estimation of Cellulose Crystallinity of Sugarcane Biomass Using Near Infrared Spectroscopy and Multivariate Analysis Methods". Carbohydr. Polym. 2017. 158: 20–28.
6. C.A. Cardona, J.A. Quintero, I.C. Paz. "Production of Bioethanol from Sugarcane Bagasse: Status and Perspectives". Bioresour. Technol. 2010. 101(13): 4754–4766.
7. C.S. Byrt, C.P.L. Grof, R.T. Furbank. "C 4 Plants as Biofuel Feedstocks: Optimising Biomass Production and Feedstock Quality from a Lignocellulosic Perspective". J. Integr. Plant Biol. 2011. 53(2): 120–135.
8. M. Balat, H. Balat. "Recent Trends in Global Production and Utilization of Bio-Ethanol Fuel". Applied Energy. 2009. 86(11): 2273–2282.
9. N. Berding, R.S. Pendrigh. "Breeding Implications of Diversifying and Uses of Sugarcane". P. Aus. Soc. Sugarcane Tech. 2009. 31: 24–38.
10. I.R. Lawler, L. Aragones, N. Berding, H. Marsh, W. Foley. "Near-Infrared Reflectance Spectroscopy is a Rapid, Cost-effective Predictor of Seagrass Nutrients". J. Chem. Ecol. 2006. 32(6): 1353–1365.
11. N. Berding, G.A. Brotherton, D.G. Lebrocq, J.C. Skinner. "Near-Infrared Reflectance Spectroscopy for Analysis of Sugarcane from Clonal Evaluation Trials .1. Fibrated cane". Crop Sci. 1991. 31(4): 1017–1023.
12. F. Masarin, D.B. Gurpilhares, D.C. Baffa, M.H. Barbosa, W. Carvalho, et al. "Chemical Composition and Enzymatic Digestibility of Sugarcane Clones Selected for Varied Lignin Content". Biotechnol. Biofuels. 2011. 4(1): 55–60.
13. D.L. Sills, J.M. Gossett. "Using FTIR to Predict Saccharification from Enzymatic Hydrolysis of Alkali-Pretreated Biomasses". Biotechnol. Bioeng. 2012. 109(2): 353–362.
14. G.E. Acquah, B.K. Via, O.O. Fasina, L.G. Eckhardt. "Non-destructive Prediction of the Properties of Forest Biomass for Chemical and Bioenergy Applications Using Near Infrared Spectroscopy". J. Near Infrared Spectrosc. 2015. 23(2): 93–102.
15. V.Q. Dang, N.K. Bhardwaj, V. Hoang, K.L. Nguyen. "Determination of Lignin Content in High-yield Kraft Pulps Using Photoacoustic Rapid Scan Fourier Transform Infrared Spectroscopy". Carbohydr. Polym. 2007. 68(3): 489–494.
16. H. An-min, J. Ze-hui, L. Gai-yun. "Determination of Holocellulose and Lignin Content in Chinese Fir by Near Infrared Spectroscopy". Spectrosc. Spect. Anal. 2007. 27(7): 1328–1331.
17. G.R. Hodge, W.C. Woodbridge. "Global Near Infrared Models to Predict Lignin and Cellulose Content of Pine Wood". J. Near Infrared Spectrosc. 2010. 18(6, SI): 367–380.
18. S. Jin, H. Chen. "Near-infrared Analysis of the Chemical Composition of Rice Straw". Ind. Crops Prod. 2007. 26(2): 207–211.
19. V. Kothiyal, A. Raturi, A. Kaler, S. Naithani. "Klason Lignin Estimation in Leucaena Leucocephala by Near Infrared Spectroscopy for Selection of Superior Material for Pulp and Paper". J. Indian. Acad. Wood Sci. 2012. 9(2): 105–114.
20. M.K.D. Rambo, E.P. Amorim, M.M.C. Ferreira. "Potential of Visible-Near Infrared Spectroscopy Combined with Chemometrics for Analysis of Some Constituents of Coffee and Banana Residues". Anal. Chim. Acta. 2013. 775: 41–49.
21. M.K.D. Rambo, F.L. Schmidt, M.M.C. Ferreira. "Analysis of the Lignocellulosic Components of Biomass Residues for Biorefinery Opportunities". Talanta. 2015. 144: 696–703.
22. U.F. Rodriguez-Zuniga, C.S. Farinas, R.L. Carneiro, G.M. da Silva, A.J. Goncalves Cruz, et al. "Fast Determination of the Composition of Pretreated Sugarcane Bagasse Using Near-Infrared Spectroscopy". Bioenergy Res. 2014. 7(4): 1441–1453.
23. S. Yao, M. Xing, S. Zhou, J. Pu, G. Wu, et al. "Determination of Lignin Content in Acacia spp. Using Near-Infrared Spectroscopy". BioResources. 2002. 5(2): 556–562.
24. C. Zhou, W. Jiang, B.K. Via, O. Fasina, G. Han. "Prediction of Mixed Hardwood Lignin and Carbohydrate Content Using ATR-FTIR and FT-NIR". Carbohydr. Polym. 2015. 121: 336–341.
25. M.A. Sanderson, F. Agblevor, M. Collins, D.K. Johnson. "Compositional Analysis of Biomass Feedstocks by Near Infrared Reflectance Spectroscopy". Biomass Bioenerg. 1996. 11(5): 365–370.
26. T.M. McLellan, J.D. Aber, M.E. Martin, J.M. Melillo, K.J. Nadelhoffer. "Determination of Nitrogen, Lignin, and Cellulose Content of Decomposing Leaf Material by Near-Infrared Reflectance Spectroscopy". Can. J. For. Res.-Rev. Can. Rech. For. 1991. 21(11): 1684–1688.
27. M. Martin, J. Aber. "Analyses of Forest Foliage III: Determining Nitrogen, Lignin and Cellulose in Fresh Leaves Using Near Infrared Reflectance Data". J. Near Infrared Spectros. 1994. 2(1): 8.
28. J.B. Sluiter, R.O. Ruiz, C.J. Scarlata, A.D. Sluiter, D.W. Templeton. "Compositional Analysis of Lignocellulosic Feedstocks. 1. Review and Description of Methods". J. Agric. Food Chem. 2010. 58(16): 9043–9053.
29. A. Sluiter, B. Hames, R. Ruiz, C. Scarlata, J. Sluiter, et al. "Determination of Structural Carbohydrates and Lignin in Biomass". Lab. Anal. Proc. NREL. 2008. 1617(1): 1–18.
30. A. Cheavegatti-Gianotto, H.M.C. de Abreu, P. Arruda, J.C. Bespalhok Filho, W.L. Burnquist, et al. "Sugarcane (Saccharum x Officinarum): A Reference Study for the Regulation of Genetically Modified Cultivars in Brazil". Trop. Plant Biol. 2011. 4(1): 62–89.
31. R. Leardi. "Application of Genetic Algorithm-PLS for Feature Selection in Spectral Data Sets". J. Chemometr. 2000. 14(5–6): 643–655.
32. J. Koljonen, T.E.M. Nordling, J.T. Alander. "A Review of Genetic Algorithms in Near Infrared Spectroscopy and Chemometrics: Past and Future". J. Near Infrared Spectros. 2008. 16(3): 189–197.
33. L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, et al. "Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy". Appl. Spectrosc. 2000. 54(3): 413–419.
34. 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. 2009. 23(1): 32–48.
35. R.W. Kennard, L.A. Stone. "Computer Aided Design of Experiments". Technometrics. 1996. 11(1): 137–148.
36. C.-W. Chang, D.A. Laird, M.J. Mausbach, C.R. Hurburgh. "Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties". Soil Sci. Soc. Am. J. 2001. 65(2): 480–490.
37. A.C. Olivieri, N. Faber, J. Ferre, R. Boque, J.H. Kalivas, et al. "Uncertainty Estimation and Figures of Merit for Multivariate Calibration". Pure Appl. Chem. 2006. 78(3): 633–661.
38. M.S. Collado, V.E. Mantovani, H.C. Goicoechea, A.C. Olivieri. "Simultaneous Spectrophotometric-Multivariate Calibration Determination of Several Components of Ophthalmic

Solutions: Phenylephrine, Chloramphenicol, Antipyrine, Methylparaben and Thimerosal". Talanta. 2000. 52(5): 909–920.

39. K.S. Booksh, B.R. Kowalski. "Theory of Analytical-Chemistry". Anal. Chem. 1994. 66(15): A782–A791.

40. H.C. Goicoechea, A.C. Olivieri. "Enhanced Synchronous Spectrofluorometric Determination of Tetracycline in Blood Serum by Chemometric Analysis. Comparison of Partial Least-Squares and Hybrid Linear Analysis Calibrations". Anal. Chem. 1999. 71(19): 4361–4368.

41. J.A.P. Cuadrado, M.P. Forn. "Validación de Métodos Analíticos". Barcelona: A.E.F.I., 2001.

42. R.F. Teófilo. Chemometric Methods in the Electrochemical Studies of Phenols on Boron-Doped Diamond Films. [PhD. Thesis]. Universidade Estadual de Campinas, Campinas, 2007.

43. W. Jerry Jr., L. Weyer. "Practical Guide to Interpretive Near-Infrared Spectroscopy". Boca Raton: CRC Press, 2008.

44. H.-J.G. Jung, J.F.S. Lamb. "Prediction of Cell Wall Polysaccharide and Lignin Concentrations of Alfalfa Stems from Detergent Fiber Analysis". Biomass Bioenerg. 2004. 27(4): 365–373.

45. L. Liu, X.P. Ye, A.R. Womac, S. Sokhansanj. "Variability of Biomass Chemical Composition and Rapid Analysis Using FT-NIR Techniques". Carbohydr. Polym. 2010. 81(4): 820–829.

46. P. Menesatti, F. Antonucci, F. Pallottino. "Estimation of Plant Nutritional Status by Vis–NIR Spectrophotometric Analysis on Orange Leaves". Biosyst. Eng. 2010. 105(4): 448–454.