BioEnergy Research
https://doi.org/10.1007/s12155-024-10761-4

Assessing the Application of Near-Infrared Spectroscopy to Determine Saccharification Efficiency of Corn Biomass

Sonia Pereira-Crespo[1] · Noemi Gesteiro[2] · Ana López-Malvar[1] · Leonardo Gómez[3] · Rogelio Santiago[2]

Received: 25 October 2023 / Accepted: 16 April 2024

Abstract

Nowadays, in the bioethanol production process, improving the simplicity and yield of cell wall saccharification procedure represent the main technical hurdles to overcome. This work evaluated the application of a rapid and cost-effective technology such as near -infrared spectroscopy (NIRS) for easily predict saccharification efficiency from corn stover biomass. Calibration process focussing on the number of samples and the genetic background of the maize inbred lines were tested; while Modified Partial Least Squares Regression (MPLS) and Multiple Linear Regression (MLR) were assessed in predictions. The predictive capacity of the NIRS models was mainly determined by the coefficient of determination (r^2ev) and the index of prediction to deviation (RPDev) in external validation. Overall, we could check a better efficiency of the NIRS calibration process for saccharification using larger number of observations (1500 sample set) and genetic backgrounds; while MPLS regression provided better prediction statistics (r^2ev = 0.80; RPDev = 2.21) compared to MLR (r^2ev = 0.68; RPDev = 1.75). These results indicate that NIRS could be successfully implemented as a large-phenotyping tool in order to test the saccharification potential of corn biomass.

Highlights

- NIRS could be successfully implemented as a large-phenotyping tool in order to test the saccharification potential of corn biomass.
- NIRS wavelengths noted provide information about associated chemical components interfering in the saccharification potential.
- The best efficiency in the NIRS calibration process was obtained using larger number of observations (1500 samples) and genetic backgrounds.
- MPLS regression model is the most reliable for NIRS prediction of corn saccharification.

Keywords Corn stover · Breeding · Calibration process · Regression models · Bioethanol

Introduction

Dependence on fossil biofuels has led to a major energy crisis with environmental and economic consequences of global concern. This situation has led to the development of new methods to find sustainable energy alternatives to meet the environmental requirements [1, 2]. The production of ethanol from starch represents the most technically advanced option but gives rise to strong competition between energy and food supply. Second generation biofuels, such as corn lignocellulose, derived from plant residues has become one of the main sustainable alternatives, not only for its high availability and wide adaptability but also for not interrupting energy demand and food supply [3].

Corn is an important food and feed crop, used as processed food, oil, feed and by-products. In addition, it can

Sonia Pereira-Crespo and Noemi Gesteiro contributed equally to this work.

✉ Rogelio Santiago
rsantiago@mbg.csic.es

[1] Facultad de Biología, Dept. Biología Vegetal & Ciencias Suelo, Unidad Asociada MBG-CSIC, Universidad Vigo, Lagoas Marcosende, 36310 Vigo, Spain

[2] Misión Biológica de Galicia (CSIC), El Palacio – Salcedo, 36143 Pontevedra, Spain

[3] Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Heslington, York YO10 5DD, UK

Published online: 23 April 2024

be used as a bioenergy crop in two ways, (i) the starch in the seeds can be used to produce ethanol, and (ii) crop residues could potentially be used to produce lignocellulosic ethanol [3, 4]. The conversion of lignocellulosic biomass to ethanol is a three-step process: (i) a pretreatment step, followed by (ii) hydrolytic degradation of the carbohydrates to the constituent sugar monomers (saccharification) and (iii) final fermentation of the free sugars to ethanol [1, 5]. Nevertheless, the key obstacle to the production of second-generation biofuels is the complicated structure of the cell wall, which is naturally resistant to decomposition and sugars release [6].

Besides cell wall recalcitrance, evaluating and selecting the optimal feedstock from large-scale germplasm for saccharification efficiency is an indispensable strategy to improve lignocellulosic biofuel production [7, 8]. The analysis of large plant populations in breeding studies for cell wall digestibility is time-consuming, labor-intensive, and economically expensive; and it is still restricted to various physical and chemical pretreatments [9]. In this regard, near-infrared spectroscopy (NIRS) is a versatile, low-cost and non-destructive indirect analytical technology than can be assessed [10, 11].

NIRS uses electromagnetic radiations in the NIR region to rapidly measure the biochemical composition of samples; however, NIR is a secondary technique, meaning that a laboratory reference method is required to create a NIR calibration. Accurate NIRS predictions of unknown samples depend on a calibration set (i.e. a large database) that is representative of the spectral and chemical variation encountered in the target population [12]. Additionally, the samples should be representative of future unknown samples to be measured in all areas of potential variability including, origin, background, constituent range(s), seasonal variation, etc. [13]; collecting the right samples is often the most difficult step in creating an accurate calibration.

Once the reference laboratory data are obtained, they are added to raw sample spectra and these data are regressed against each other. Processed and standardized NIR spectra contain multiple variables in the form of reflectance that is regressed with targeted traits. Multivariate regression techniques such as multiple linear regression (MLR), partial least square (PLS) and principal component regression (PCR) are used to generate robust and effective models [12, 14]. Partial least squares (PLS) regression is a powerful multivariate technique that finds latent factors in the data to maximize the covariance between spectra and the target trait. To ensure that the underlying relationship is captured in a PLS model, researchers typically perform cross-validation on the calibration set. The newly developed calibration model is then tested using spectra from independent samples, the validation set, to ensure that the model is neither over-fitted nor under-fitted [15, 16]. Moreover, modified PLS (MPLS) is considered stable and less prone to over fitting due to the influence of intragroup variations [17].

The final output will be a linear equation that can be applied to future unknown samples in order to predict constituents or properties of interest. It allows the high-throughput screening of populations at both qualitative and semi-quantitative levels. In recent years, this technology has been applied to evaluate biomass digestibility in several species, such as miscanthus [18], Jerusalem artichoke [11], wheat [19], eucalyptus [20], sweet sorghum [21], rice [22], and sugarcane [23]. However, so far, nothing has been explored on the NIRS potential for the determination of corn stover saccharification.

In the current work, we investigate about the efficiency of the calibration process focussing on the number of samples and the genetic background of the maize inbred lines included. Moreover, we compared two common multivariate regression methods in the calibration development (MLR and MPLS). Overall, the main objective of the present work is to evaluate the capability of NIRS as a fast tool to predict the saccharification efficiency of lignocellulosic biomass of corn, in order to use this instrument for breeding purposes.

Materials and Methods

Field Trials/Sample Dataset

We used inbred lines from two different sources in order to explore maize genetic variability: Recombined inbred lines (RILs) from a Multi-parent Advanced Generation Inter-Crosses (MAGIC) population. We optimized the MAGIC using eight temperate maize inbred lines of diverse genetic origin, as five of them derive directly from different open pollinated varieties from Spain, Italy, and France, while two lines are from Northern North America; all the parental lines belong to the non-stiff stalk genetic group [24]. On the other hand, the USDA North Central Regional Plant Introduction Station in AMES, Iowa maintains over 3000 maize inbreds from around the world. When the inbreds were classified according to breeding program of origin, the different breeding programs tended to group together, with most of the USA programs in the two major germplasm groups recognized by temperate maize breeders (stiff stalk and non-stiff stalk). They also include other materials from international programs (for example, Spain, France, China, Argentina, or Australia) that seem to represent germplasm pools different from those commonly used in North American programs [25].

Field evaluations were carried out at Misión Biológica de Galicia in Pontevedra (42°24′ N, 8°38′ W, 20 m above sea level). The complete field trials consisted of (i) a subset of

408 lines from a MAGIC population together with the eight founders (EP17, EP43, EP53, EP86, PB130, F473, A509, and EP125) [2] in 2016 and 2017, and (ii) a reduced subset of 836 lines, belonging to the AMES Association Panel (North Central Regional Plant Introduction Station, USA), together with 6 controls (A619, A632, A662, A665, PH207, EP42) in 2018 and 2019 [26].

The subset of 408 lines from the MAGIC population was evaluated following a single augmented design with 10 blocks, 42 non-replicate lines were included in each block, along with the eight inbred founders. Each plot consisted of a single row, 2.4 m long and 13 plants per row, with the spacing between consecutive hills in a row being 0.18 m and 0.8 m between rows. Whereas, the subset of 836 lines from the AMES Panel was evaluated following an augmented 17-block design, each block consisting of 50 lines and the six testers. Each plot consisted of a single row, 2.4 m long and 13 plants per row, with the spacing between consecutive hills in a row being 0.21 m and 0.8 m between rows. From the 836 lines evaluated in the field, 300 lines with great genetic variability, as well as adapted to the growth conditions of Pontevedra and with sufficient material for saccharification analyses were included. Local agronomical practices were followed.

The global dataset included 1500 corn stover samples collected from both subsets (approximately 400 inbred lines from the MAGIC and 300 inbred lines from AMES during 2 years of evaluation, including replicated testers in the corresponding blocks)(Supplementary file 1). Each sample was composed of tissues from 2 to 10 plants collected at grain harvest starting from 55 days after flowering. The samples, once dried (60 °C, 7 days) were ground in a mill (Restch SM100, Germany) with a 0.75 mm mesh for subsequent saccharification determination.

Saccharification Efficiency Measurements

Saccharification efficiency was determined following the method described by Gómez and coauthors at the Centre for Novel Agricultural Products (CNAP) [9]. Ground material was weighed into 96-well plates, each well contained 4 mg of each sample using a custom-made robotic platform (Labman Automation, Stokesley, North Yorkshire, UK). Pretreatment, hydrolysis and sugar determination were performed automatically by a robotic platform (Tecan Evo 200; Tecan Group Ltd. Männedorf, Switzerland). Samples were pre-treated with sodium hydroxide (NaOH, 0.5 M, Fisher Scientific, UK) at 90 °C for 30 min, washed four times with 500 µl sodium acetate buffer ($C_2H_3NaO_2$, Sigma-Aldrich, UK) and finally subjected to enzymatic digestion (Celluclast 2, 7FPU/g, Novozymes, Bagsvaerd, Denmark) at 50 °C for 9 h. Samples were analyzed in duplicate/triplicate (SD mean from 10 to 15). The amount of released sugars was assessed against a glucose standard curve using the 3-methyl-2-benzothiazolinone hydrozone method (MTBH, Sigma-Aldrich, UK) [27]. This method was tested for detection of a range of sugars that are released from the cell wall, and showed sensitive detection of several monosaccharides.

NIR Spectra Acquisition

Every sample was allowed to stabilize at room temperature prior to spectral data acquisition. The determinations were carried out in duplicate in a temperature-controlled room (~ 24 °C), with the dry and ground samples (~ 30 g) loaded in a circular quartz cuvette for solids (internal diameter of 11 cm) [13]. NIR spectra were collected on an instrument FOSS NIRS D2500 spectrometer (FOSS, Hillerød, Denmark) in the visible and near infrared region (400–2498 nm) at 0.5 nm intervals, in reflectance mode [12]. The acquired spectra were processed with WinISI software (version 4.12, Infrasoft International, PA, USA). The average spectrum of each sample was used for calibration and validation procedures.

Statistical Analysis

Three different calibration and validation process were developed: a global approach including samples from both panels, and independent processes for MAGIC and AMES panels. Chemometric analysis was performed by both Modified Partial Least Square Regression (MPLS) and Multiple Linear Regression (MLR) methods. The MLR models were built with Stepwise selection of wavelength applies an F-test to identify the best-fitted model. The different datasets were randomly divided into two subsets using the SELECT algorithm included in the WinISI IV software.

A principal component analysis (PCA) on the first derivative of the absorbance was used to calculate the global Mahalanobis distance (GH) of each sample to the centre of the population in an n-dimensional space [28] using the CENTER algorithm included in the WinISI IV software. The samples with GH > 3 were identified as spectral outliers and removed, repeating the operation until all samples had a GH value lower than the recommended maximum [28]. During calibration process, three elimination passes of chemical outliers were applied, considering the critical T-statistic value set for chemical outliers detection was 2.5 [29]. Calibrations were developed after removing all outliers.

In order to develop the most accurate calibration models, different combinations of scatter corrections (NONE, no correction; D, detrending; SNV, standard normal variate; SNV + D, standard normal variate and detrending; WMSC, weighted multiplicative scatter correction; and SMSC, standard multiplicative scatter correction) and mathematical treatments (0, 0, 1, 1; 1, 4, 4, 1; 1, 5, 5, 1; 1, 6, 4, 1; 1, 8, 4, 1; 1, 10, 5, 1; 1, 10, 10, 1; 2, 4, 4, 1; 2, 5, 5, 1; 2, 6, 4, 1; 2, 8, 4, 1; 2, 10, 5, 1; 2, 10, 10, 1; where the first digit is the derivative order, the second is the gap over which the derivative is calculated, the third is the number of data points in the first smoothing, and the fourth is the second smoothing) were tested [12].

The prediction models were developed using a subset as the calibration set using ~ 75% of the samples (n = 1150 in the global approach, n = 536 MAGIC, n = 527 AMES) evaluated by leave-one-out cross-validation, and then tested on the remaining ~ 25% of the samples performing an external validation (n = 350 in the global approach, n = 195 MAGIC, n = 180 AMES).

The best-fit equation was considered qualified as prediction model on the basis of results for standard error of cross-validation (SEcv), the standard error of external validation (SEev), the coefficient of determination calculated in internal cross-validation (1 – VR) and external validation (r^2ev). In addition, in order to evaluate the accuracy of a calibration model and to allow standard comparison with other studies, we calculated the index of prediction to deviation (RPD), a non-dimensional statistic for the quick evaluation and classification of NIR spectroscopy calibration models which has been widely used in NIRS studies, and defined as the ratio of the standard deviation of the reference data for the samples to SEcv/SEev; and the range error index (RER), defined as the ratio of the range in the reference data for the samples to the SEcv/SEev [30, 31]. Finally, bias and slope were calculated with the external validation samples; the slope represents a change in predicted values with a unit change in reference values, and biasness is the average of residuals of laboratory and reference values, which account as well for prediction accuracy.

Results and Discussion

Lignocellulosic biomass consists of three main structural units: cellulose, hemicellulose and lignin. Cellulose is a crystalline polymer of glucose, hemicellulose is an amorphous polymer of xylose and arabinose, and lignin is a complex polymer of aromatic alcohols. Vibration bands associated with these chemical biomass components [32] can be observed in Fig. 1, which displays the average NIR spectra of 1500 analysed samples of corn stover. Five main absorption peaks at 1456, 1912, 2100, 2252 and 2310 nm, were in accordance with the spectral fingerprint showed by Guimarães and coauthors [33] for prediction of theoretical ethanol yield in sorghum biomass.

Regarding the wavelengths selected in MLR calibration, the results showed that two wavelengths were the most relevant characteristic absorption peaks, particular at 824 and 880 nm, which are associated to the third overtone band of C-H bond, related to sugars [34]. The wavelength region from 1600 to 1800 nm is associated to the absorption band of a C-H stretching first overtone corresponding to fiber components of cell wall [35, 36], peaks around 1780 nm being associated to the absorption band of a C-H stretching first overtone corresponding to carbohydrates, such as cellulose and hemicellulose [35, 37]. Other relevant coefficient appears in the region ~ 2332 nm, which assigned to cellulose and lignin absorption (C–H stretching/C–H deformation combination) [35]. Overall, both regression methods used for calibration showed similar trends in wavelength (or regions) associated/related to the chemical composition of corn stover biomass. The cell wall structure and composition governs bioethanol production [8]; therefore, the wavelengths defined in the current work provide useful information about associated chemical components interfering in the saccharification potential.

The range of variation for the saccharification efficiency of the complete dataset obtained by laboratory analysis at CNAP is shown in Fig. 2. Samples of the calibration set are reported after the removal of all outliers (spectral and chemical), where the means (and ranges) expressed as nmol mg^{-1} $material^{-1}$ h^{-1}were: 153.3 (min. 77.6 to max. 204.5) and 150.5 (min. 77.6 to max. 204.5) for MPLS and MLR model, respectively. The external validation set had similar mean and range values, with 153.6 (min. 77.6 to max. 204.5) nmol mg^{-1} $material^{-1}$ h^{-1} for both regression models. These means and ranges were higher and wider than previous studies using other crop species such as rice, barley, wheat, triticale, sorghum, miscanthus or brachypodium [38–40]. This range of enzyme-released glucose was expected due to the extensive background of the samples [2, 26], and suggest that many expected shifts will be represented in order to accomplish new germplasm phenotyping screenings.

The prediction models resulting from the second derivative (2, 4, 4, 1), and a combination of standard normal variate and detrend as scatter correction method, provided a more accurate and precise estimate for saccharification efficiency using the complete dataset. During calibration procedures, the number of samples removed as chemical T outliers, expressed as a percentage of the total initial samples in the set, ranged from 7.2 to 8.3% for both prediction models obtained, these values being lower than the maximum value (20%) annotated by Shenk and Westerhaus [12].

Attending to calibration and cross-validation statistics showed in the Table 1, we can define a better prediction model for the MPLS regression in comparison to the MLR. The coefficients of determination (1 – VR) and the standard errors of prediction in cross-validation (SEcv) were 0.84 and 10.80 nmol mg^{-1} $material^{-1}$ h^{-1} for MPLS model, and 0.68 and 14.85 nmol mg^{-1} $material^{-1}$ h^{-1} for MLR model, respectively. In this sense, and according

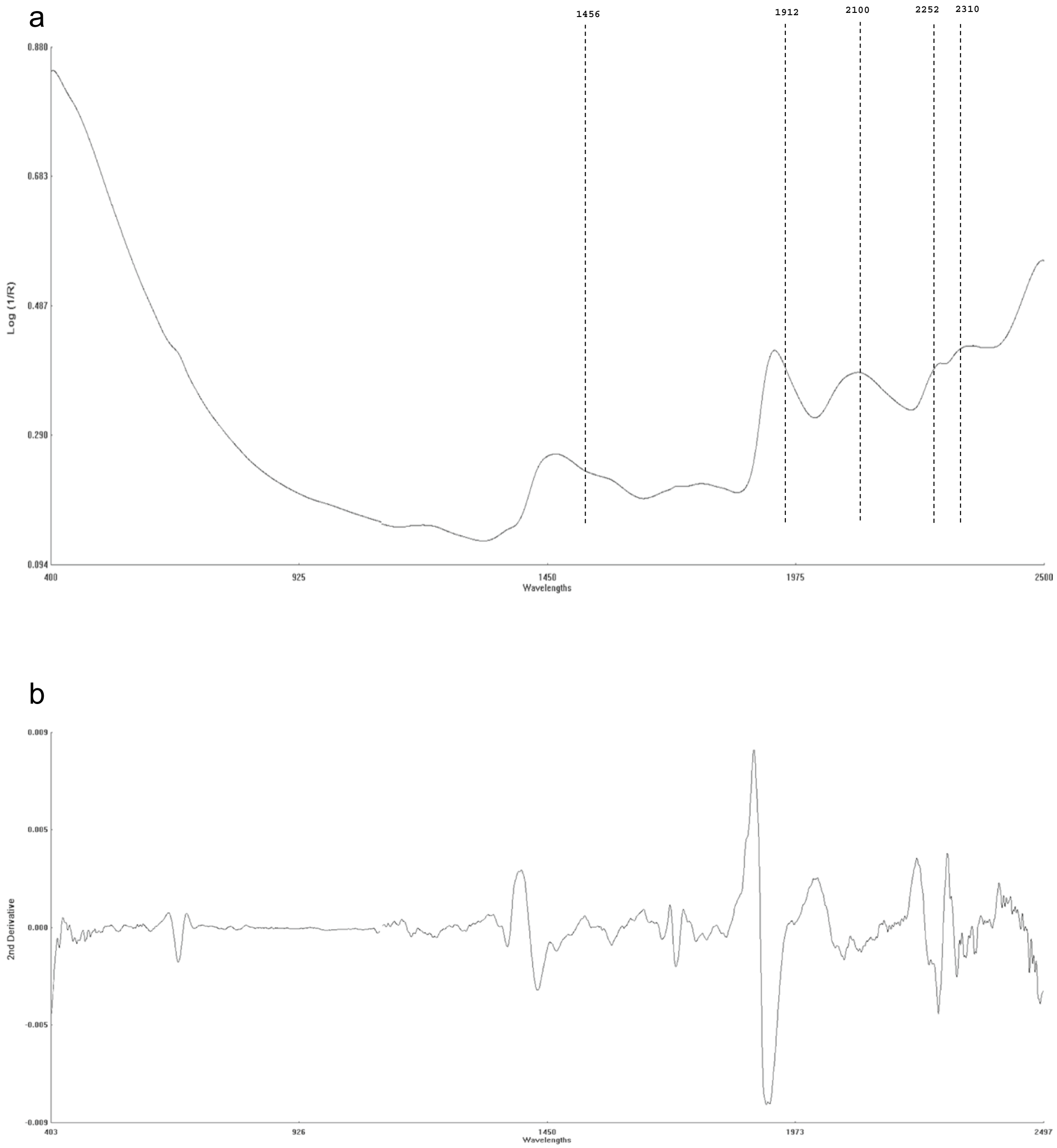

Fig. 1 Average raw (**a**) and second derivative spectra (**b**) of a total set ($n = 1500$) of corn biomass samples using near-infrared spectroscopy in reflectance mode. Dotted lines indicate five main absorption peaks related to the main components of corn stover spectra, in accordance with the spectral fingerprint for prediction of theoretical ethanol yield in sorghum biomass [33]. NIR spectral absorbance values [log (1/R)], where R is the reflectance

to Shenk and coauthors [41], our NIRS prediction using MPLS model with an 1-VR value higher than 0.70 indicate a good predictive ability, while the use of MLR model with a 1-VR lower than this value could be just used to qualitative estimation purposes (separating groups with higher and lower analytical values).

On the other hand, RPD value governs the prediction accuracy of the models. RPD is defined as the ratio of prediction to standard deviation of reference values, wherein an RPD value < 1.5 indicates that the calibration is not reliable; a value between 1.5 and 2.0 indicates the capacity of a model to distinguish between high and low values; a value between 2.0 and 2.5 signifies the model's capacity to "approximate" quantitative prediction; a value between 2.5 and 3.0 suggests "good" quantitative prediction; and avalue > 3.0 indicates "excellent" quantitative prediction [31, 42], whereas models with RER values under 3 are considered unsuccessful, while RER

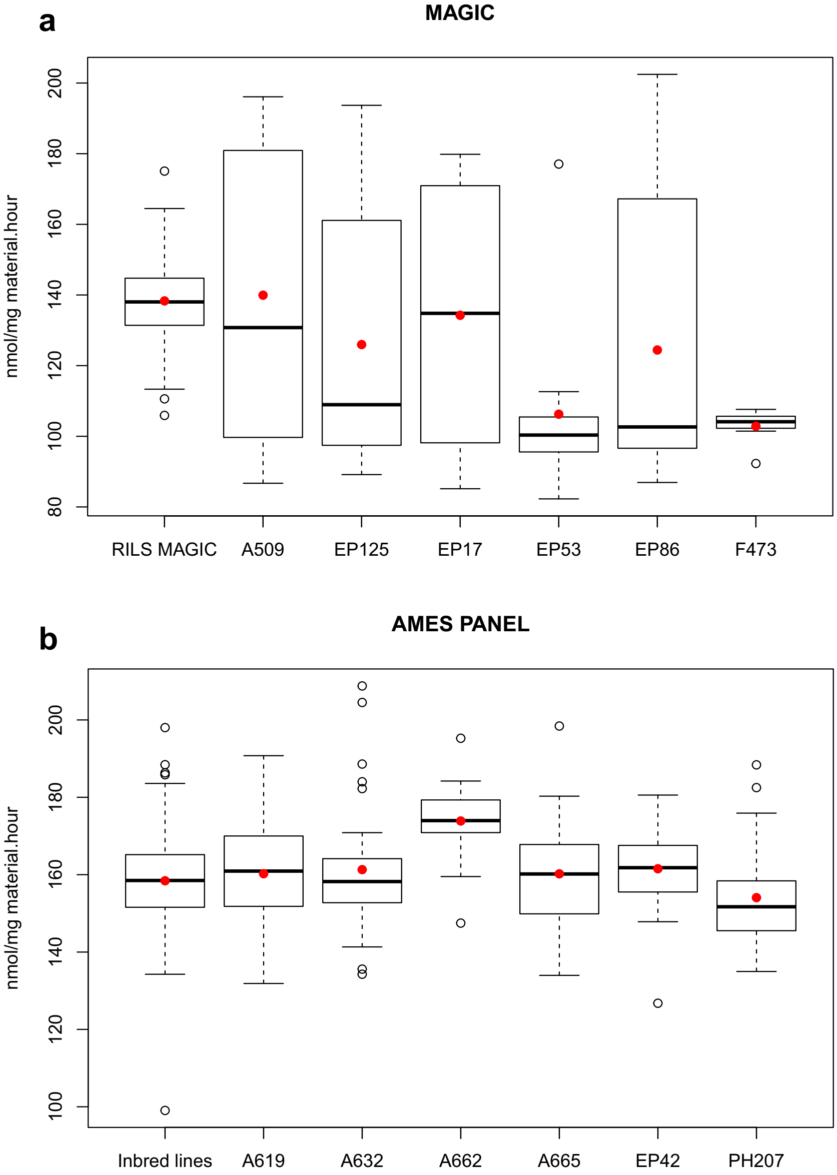

Fig. 2 Boxplots of the saccharification data obtained in the two sample subsets included in this study. a: subset of 408 lines from a MAGIC population together with six founders (EP17, EP53, EP86, F473, A509, and EP125), and b: a subset of 300 lines, belonging to the Ames association panel (North Central Regional Plant Introduction Station, USA), together with 6 controls (A619, A632, A662, A665, PH207, EP42). Red dots indicate the mean values

values between 3 and 10 indicate limited applicability (e.g., screening) and RER values higher than 10 are considered to characterize high-quality models [30, 43]. For our models, RPDcv and RERcv achieved values of 2.55 and 11.75 for MPLS model, and 1.78 and 8.54 for MLR model, respectively, indicating a reliable prediction power for MPLS model. By contrast, MLR model would be fair, but just allow classifying samples into high and low groups of saccharification efficiency.

After the calibration process, both models were validated with an external (independent) set of samples (Table 1). The values of the coefficient of determination r^2ev and the RPDev were 0.80 and 2.21 for MPLS model and 0.68 and 1.75 for MLR model, respectively. Considering the criteria previously defined, the predictive quality of the calibration models based on RPDev values were considered poor for MLR model, and suitable for quantitative predictions for MPLS model. However, we have to note that the RERev values for both models, shortly exceed the minimum value suggested by Williams and Sobering [30] for a reliable quantitative model (RER > 10), with 10.03 and 12.64 for MLR and MPLS, respectively. Although we should mention that the expected range used in RER calculation depends on the number of samples, whereas the standard deviations used in RPD was not, this dependence is the reason for preferring RPD over RER [44].

In the same way, attending to bias and slope of the external validation, both models showed good results, although MPLS displayed better results (0.27 for bias and 1.00 for slope). An ideal slope value should be 1, but any value close to 1 would also represent the accuracy of the model; whereas bias should have a value close enough to 0, a negative value

Table 1 Calibration, cross-validation and external validation statistics of NIRS models for predicting the saccharification efficiency (nmol mg^{-1} material^{-1} hora^{-1}) of corn stover biomass using the 1500 samples set. Modified Partial Least Squares Regression (MPLS) and Multiple Linear Regression (MLR) models

Calibration									**Cross Validation**			
Model	**n**	**SD**	**Outliers**	**Mean**	**Min**	**Max**	**r^{2}c**	**SEc**	**1-VR**	**SEcv**	**RPDcv**	**RERcv**
MPLS	1150	27.6	83	153.3	77.6	204.5	0.88	9.46	0.84	10.80	2.55	11.75
MLR	1150	26.5	96	150.5	77.6	204.5	0.69	14.17	0.68	14.85	1.78	8.54
External Validation												
Model	**n**	**SD**	**Outliers**	**Mean**	**Min**	**Max**	**r^{2}ev**	**SEev**	**Bias**	**Slope**	**RPDev**	**RERev**
MPLS	350	22.2	0	153.6	77.6	204.5	0.80	10.04	0.27	1.00	2.21	12.64
MLR	350	22.2	0	153.6	77.6	204.5	0.68	12.65	-1.32	0.92	1.75	10.03

CALIBRATION: *n*: number of observations; *Outliers*: number of anomalous chemical samples removed from the study; *SD*: standard deviation; *Min*: minimum; *Max*: maximum; *SEc*: standard error of calibration; *r^{2}c*: coefficient of determination of the calibration model; ***CROSS VALIDATION:*** *1-VR*: coefficient of determination in cross-validation; *SEcv*: standard error of cross-validation; *RDPcv*: ratio between the standard deviation and the error of prediction in cross-validation; *RERcv*: ratio between the range of values and the error of prediction in cross-validation; ***EXTERNAL VALIDATION:*** *r^{2}ev*: coefficient of determination in external validation; *SEev*: standard error of prediction; *Bias*: mean difference between the observed values and those predicted by the equation; slope: regression slope; *RDPev*: ratio between the standard deviation and the error of prediction in external validation; *RERev*: ratio between the range of values and the error of prediction in external validation

relates to underestimation by the model, whereas a positive bias value depicts overestimation [45].

Comparing the results, the MPLS calibration method demonstrated to have more predictive ability than the MLR to measure the prediction of saccharification efficiency (Fig. 3). MPLS is known to be a more effective model than MLR for the development of NIRS calibration models, particularly with large datasets, by reducing the dataset into a small number of orthogonal factors and to enabling avoid collinearity and over-fitting [46]. Additionally, MPLS is known to be more reliable than MLR for the calibration of complex parameters [47]. Additionally, the MPLS technique was better than MLR model in the validation on independent set. Therefore, we recommend constructed calibration model by MPLS technique in preference to MLR technique for this saccharification trait.

Finally, contrasting the results obtained with other potential species for bioethanol production, Huang and coauthors [22], using MPLS model, reported similar or slightly lower predictive ability for estimating biomass saccharification (expressed as total releases sugar) (r^{2}c = 0.75, RPDev = 2.0) in a rice straws; van der Weijde and coauthors [48] developed NIRS models to predict of saccharification efficiency of the crop *Miscanthus* and obtained good correlations (1-VR: 0,82–0,92); while Li and colleagues [10] developed a calibration model that included different sugarcane genotypes, and they found RPD values of over 2.0 in calibration, internal cross-validation, and external validation. These results are as good as the obtained in the current work. However, and related to the complexity of the parameter, we should note that the performance of our calibration models was more limited than those reported in sugarcane [23], who obtained NIRS models for fermentable hexoses and total sugar that exhibited excellent prediction capability (RPD values higher than 4.0) for predicting biomass digestibility. The usefulness of those last traits to estimate bioethanol potential could be consider in future studies evaluating corn biomass.

Alternatively, as databases get larger, this increases the complexity in terms of variability, and although this is normally seen as an advantage in global calibrations, in practice it creates a problem because prediction accuracy decreases [49, 50]. Although we do not have variability in terms of different species or local laboratory determination facilities, we tried to define the advantages or disadvantages of the use of smaller datasets in the calibration process, primarily based on genetic variability of the inbred lines included. Focussing in statistics in external validation and MPLS model (Table 2), the values of the coefficient of determination r^{2}ev and the RPDev were 0.69 and 1.73 for the MAGIC, and 0.24 and 1.15 for AMES, respectively. Considering the criteria previously defined, the predictive quality of the calibration models was considered poor or very poor. In addition,

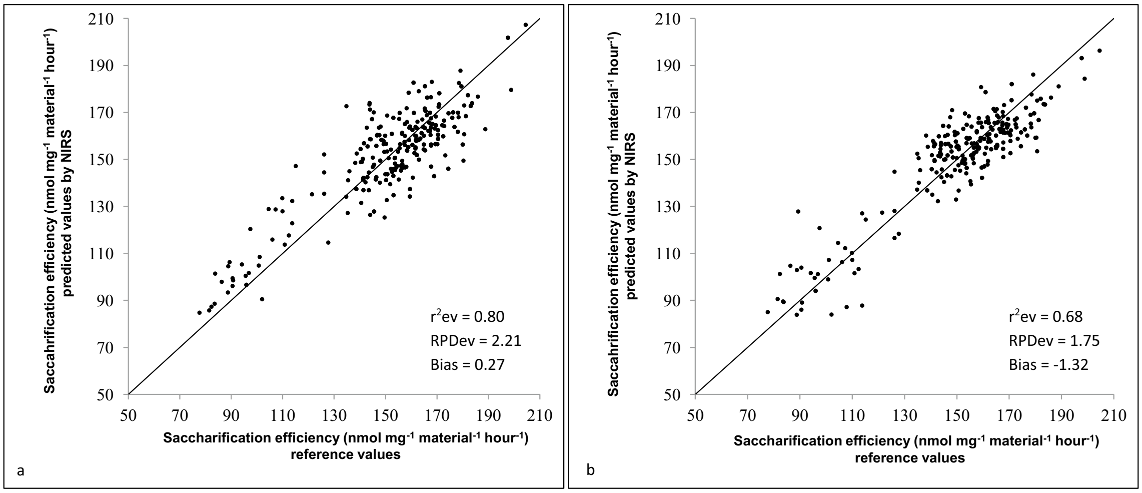

Fig. 3 Validation scatter plot of reference values *vs.* predicted values by NIRS of saccharification efficiency (nmol mg^{-1} material^{-1} h^{-1}) for samples of corn stover biomass. a: Modified Partial Least Squares Regression (MPLS) and b: Multiple Linear Regression (MLR)

parameters such as the bias indicate greater overestimation in relation to the average of residuals of laboratory and reference values (5.58 for bias).

Regarding to the genetic background of the datasets, the most outstanding different we can note is that the AMES set correspond to a non-structured panel (set of genetically diverse lines), including assorted materials, but with greater proportion of American programs (stiff stalk and non-stiff stalk germplasm groups) [25]; whereas the MAGIC population refers to a limited number of known parents of diverse origins (Spain, Italy, France and Northern North America) and just including non-stiff stalk materials [24]; nevertheless, although the MAGIC population showed better results for some calibration statistics, they are far away from the observed in the global approach.

Conclusions

We can check a better efficiency of the NIRS calibration process using larger number of observations and genetic backgrounds. In addition, the comparison of regression methods for estimating saccharification efficiency showed that the Modified Partial Least Squares was a better method than Multiple Linear Regression, based on terms of higher correlation coefficient between predicted and reference values and higher index of prediction (RPD). As a result, we can state that near-infrared spectroscopy can be effectively used in the screening of large germplasm corn collections in relation to the use of their biomass in bioethanol production.

Table 2 Calibration, cross-validation and external validation statistics of NIRS models for predicting the saccharification efficiency (nmol mg^{-1} material^{-1} h^{-1}) of corn stover biomass using the samples subsets MAGIC (Multi-parent Advanced Generation InterCrosses) and AMES (USDA North Central Regional Plant Introduction Station). Modified Partial Least Squares Regression (MPLS) and Multiple Linear Regression (MLR) models

Calibration										**Cross Validation**			
Panel	**Model**	**n**	**SD**	**Out**	**Mean**	**Min**	**Max**	**r^2c**	**SEc**	**1-VR**	**SEcv**	**RPDcv**	**RERcv**
MAGIC	MPLS	536	40.00	36	137.4	77.60	218.7	0.80	19.00	0.71	21.02	1.90	6.71
	MLR	536	40.00	34	137.4	77.60	218.7	0.77	19.64	0.76	20.00	2.00	7.06
AMES	MPLS	527	15.95	43	158.4	93.10	224.7	0.29	9.71	0.21	10.54	1.51	12.49
	MLR	527	15.95	40	158.4	93.10	224.7	0.37	9.96	0.36	10.01	1.59	13.15
External Validation													
Panel	**Model**	**n**	**SD**	**Out**	**Mean**	**Min**	**Max**	**r^2ev**	**SEev**	**Bias**	**Slope**	**RPDev**	**RERev**
MAGIC	MPLS	195	40.08	0	136.2	77.60	218.3	0.69	23.14	5.58	0.92	1.73	6.08
	MLR	195	40.08	0	136.2	77.60	218.3	0.74	20.81	1.31	0.92	1.93	6.76
AMES	MPLS	180	16.28	0	157.4	93.10	224.7	0.24	14.15	0.47	1.06	1.15	9.30
	MLR	180	16.28	0	157.4	93.10	224.7	0.38	13.17	0.55	0.91	1.24	9.99

CALIBRATION: *n:* number of observations; Out: number of anomalous chemical samples removed from the study; *SD:* standard deviation; *Min:* minimum; *Max:* maximum; *SEc:* standard error of calibration; *r^2c*: coefficient of determination of the calibration model; ***CROSS VALIDATION*****:** *1-VR:* coefficient of determination in cross-validation; *SEcv:* standard error of cross-validation; *RDPcv:* ratio between the standard deviation and the error of prediction in cross-validation; *RERcv:* ratio between the range of values and the error of prediction in cross-validation; ***EXTERNAL VALIDATION:*** *r^2ev:* coefficient of determination in external validation; *SEev:* standard error of prediction; *Bias:* mean difference between the observed values and those predicted by the equation; *slope:* regression slope; *RDPev:* ratio between the standard deviation and the error of prediction in external validation; *RERev:* ratio between the range of values and the error of prediction in external validation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12155-024-10761-4.

Acknowledgements N.G. acknowledges an FPI contract ("Contrato predoctoral para la formación de doctores") funded by project RTI2018-096776-B-C21 (financed by MCIU/AEI/FEDER, UE). A.L-M. acknowledges her postdoctoral scholarship to the Xunta de Galicia. Authors acknowledge the technical support of Servicio de Espectroscopia NIRS Aplicada al Ambito Agroforestal (MBG-SENAF).

Author Contributions Conceptualization, R.S.; Methodology, S.P.-C, N.G., A.L.-M.; Software, S.P.-C; Formal Analysis, S.P.-C. and N.G..; Investigation, R.S. and L.D.G; Data Curation, N.G. and A.L.-M.; Writing—Original Draft Preparation, S.P.-C., N.G., R.S.; Writing—Review and Editing, R.S., A.L.-M., L.D.; Supervision, R.S.; Project Administration, R.S; Funding Acquisition, R.S. All authors have read and agreed to the final version of the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was funded by subsequent coordinated projects financed by MCIU/AEI/FEDER, UE (RTI2018-096776-B-C21, RTI2018-096776-B-C22, PID2021-122196OB-C21, PID2021-122196OB-C22 and INTRAMURAL ref. 202240I051).

Data Availability The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of Interests The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Nie JM, Zhang RJ, Liu XY, Yang F, Wang JJ, Xiao J, Zhao J (2019) Technologies for lignocellulose pretreatment to produce fuel ethanol. IOP Conf Ser: Earth Environ Sci 237:042034. https://doi.org/10.1088/1755-1315/237/4/042034
2. López-Malvar A, Butrón A, Malvar RA, McQueen-Mason SJ, Faas L, Gómez LD, Revilla P, Figueroa-Garrido DJ, Santiago R (2021) Association mapping for maize stover yield and saccharification efficiency using a multiparent advanced generation intercross (MAGIC) population. Sci Rep 11:3425. https://doi.org/10.1038/s41598-021-83107-1
3. Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN Jr (2008) Plant to power: bioenergy to fuel the future. Trends Sci 13:421–429. https://doi.org/10.1016/j.tplants.2008.06.001
4. Dhugga KS (2007) Maize biomass yield and composition for biofuels. Crop Sci 47:2211–2227. https://doi.org/10.2135/cropsci2007.05.0299
5. Mosier N, Wyman C, Dale B, Elander R, Lee YY, Holtzapple M, Ladisch M (2005) Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresour Technol 96:673–686. https://doi.org/10.1016/j.biortech.2004.06.025
6. Taherzadeh MJ, Karimi K (2008) Pretreatment of lignocellulosic wastes to improve ethanol and biogas production: a review. Int J Mol Sci 9:1621–1651. https://doi.org/10.3390/ijms9091621
7. Zhao H, Li Q, He J, Yu J, Yang J, Liu C, Peng J (2014) Genotypic variation of cell wall composition and its conversion efficiency in *Miscanthus sinensis*, a potential biomass feedstock crop in China. GCB Bioenergy 6:768–776. https://doi.org/10.1111/gcbb.12115
8. Carroll A, Somerville C (2009) Cellulosic biofuels. Annu Rev Plant Biol 60:165–182. https://doi.org/10.1146/annurev.arplant.043008.092125
9. Gómez LD, Whitehead C, Barakate A, Halpin C, McQueen-Mason SJ (2010) Automated saccharification assay for determination of digestibility in plant materials. Biotechnol Biofuels 3:23. https://doi.org/10.1186/1754-6834-3-23
10. Li X, Ma F, Liang C, Wang M, Zhang Y, Shen Y, Adnan M, Lu P, Khan MT, Huang J, Zhang M (2021) Precise high-throughput online near-infrared spectroscopy assay to determine key cell wall features associated with sugarcane bagasse digestibility Biotechnol. Biofuels 14:123. https://doi.org/10.1186/s13068-021-01979-x
11. Li M, He S, Wang J, Liu Z, Xie GH (2018) A NIRS-based assay of chemical composition and biomass digestibility for rapid selection of Jerusalem artichoke clones. Biotechnol Biofuels 11:334. https://doi.org/10.1186/s13068-018-1335-1
12. Shenk JS, Westerhaus MO (1994) The application of near infrared reflectance Spectroscopy (NIRS) to forage analysis. In: Fahey Jr GC (ed) Forage quality, evaluation, and utilization. Soil science society of america/american society of agronomy/crop science society of america. Madison, pp 406–449
13. Xiao X, Lijuan X, Yibin Y (2019) Factors influencing near infrared spectroscopy analysis of agro-products: a review. Front Agr Sci Eng 6:105–115. https://doi.org/10.15302/J-FASE-2019255
14. Brereton RG (2003) Chemometrics: data analysis for the laboratory and chemical plant. Wiley, Chichester
15. Martens H, Naes T (1989) Multivariate calibration. Wiley, New York
16. Naes T, Isakson T, Fearn T et al (2002) A user-friendly guide to multivariate calibration and classification. NIR Publications, Chichester
17. Eisenstecken D, Panarese A, Robatscher P, Huck CW, Zanella A, Oberhuber M (2015) A near infrared spectroscopy (NIRS) and chemometric approach to improve apple fruit quality management: a case study on the cultivars "Cripps Pink" and "Braeburn." Molecules 20:13603–13619. https://doi.org/10.3390/molecules200813603
18. Huang J, Xia T, Li A, Yu B, Li Q, Tu Y, Zhang W, Yi Z, Peng L (2012) A rapid and consistent near-infrared spectroscopic assay for biomass enzymatic digestibility upon various physical and chemical pretreatments in *Miscanthus*. Bioresour Technol 121:274–281. https://doi.org/10.1016/j.biortech.2012.06.015
19. Lomborg CJ, Thomsen MH, Jensen ES, Esbensen KH (2010) Power plant intake quantification of wheat straw composition for 2nd generation bioethanol optimization—a near infrared spectroscopy (NIRS) feasibility study. Bioresour Technol 101:1199–1205. https://doi.org/10.1016/j.biortech.2009.09.027
20. Hou S, Li L (2011) Rapid characterization of woody biomass digestibility and chemical composition using near-infrared spectroscopy free access. J Integr Plant Biol 53:166–175. https://doi.org/10.1111/j.1744-7909.2010.01003.x

21. Wu L, Li M, Huang J, Zhang H, Zou W, Hu S, Li Y, Fan C, Zhang R, Jing H, Peng L (2015) A near-infrared spectroscopic assay for stalk soluble sugars, bagasse enzymatic saccharification and wall polymers in sweet sorghum. Bioresour Technol 177:118–124. https://doi.org/10.1016/j.biortech.2014.11.073
22. Huang J, Li Y, Wang Y, Chen Y, Liu M, Wang Y, Zhang R, Zhou S, Li J, Tu Y, Hao B (2017) A precise and consistent assay for major wall polymer features that distinctively determine biomass saccharification in transgenic rice by near-infrared spectroscopy. Biotechnol Biofuels 10:1–14. https://doi.org/10.1186/s13068-017-0983-x
23. Adnan M, Shen Y, Ma F, Wang M, Jiang F, Hu Q, Mao L, Lu P, Chen X, He G, Tahir Khan F, Deng Z, Chen B, Zhang M, Huang J (2022) A quick and precise online near-infrared spectroscopy assay for high-throughput screening biomass digestibility in large scale sugarcane germplasm. Ind Crops Prod 189:115814. https://doi.org/10.1016/j.indcrop.2022.115814
24. Jiménez-Galindo JJ, Malvar RA, Butrón A, Santiago R, Samayoa LF, Caicedo M, Ordás B (2019) Mapping of resistance to corn borers in a MAGIC population of maize. BMC Plant Biol 19:431. https://doi.org/10.1186/s12870-019-2052-z
25. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA et al (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55. https://doi.org/10.1186/gb-2013-14-6-r55
26. Gesteiro N, Butrón A, Santiago R, Gómez LD, López-Malvar A, Álvarez-Iglesias L, Revilla P, Malvar RA (2023) Breeding dual-purpose maize: grain production and biofuel conversion of the stover. Agronomy 13:1352. https://doi.org/10.3390/agronomy13051352
27. Anthon GE, Barrett DM (2002) Determination of reducing sugars with 3-methyl-2-benzothiazolinonehydrazone. Anal Biochem 305:287–289. https://doi.org/10.1006/abio.2002.5644
28. Shenk JS, Westerhaus MO (1991) Population structuring of near infrared spectra and modified partial least squares regression. Crop Sci 31:1548–1555. https://doi.org/10.2135/cropsci1991.0011183X003100060034x
29. Mark H, Workman J (2003) Statistics in Spectroscopy. Elsevier-Academic Press, Amsterdam
30. Williams PC, Sobering DC (1996) How do we do it?: a brief summary of the methods we use in developing near infrared calibrations. In: Davies W (ed) Near Infrared Spectroscopy: The Future Waves. NIR Publications, Chichester, Reino Unido, pp 185–188
31. Williams PC (2014) The RPD statistic: a tutorial note. NIR News 25:22–26. https://doi.org/10.1255/nirn.1419
32. Feng X, Jianming Y, Tesfaye T, Floyd D, Donghai W (2014) Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: A mini-review. ChemInform 104:801–809. https://doi.org/10.1016/j.apenergy.2012.12.019
33. Guimarães CC, Simeone MLF, Parrella RAC, Sena MM (2014) Use of NIRS to predict composition and bioethanol yield from cell wall structural components of sweet sorghum biomass. Microchem J 117:194–201. https://doi.org/10.1016/j.microc.2014.06.029
34. Chen JY, Zhang H, Miao Y, Asakura M (2010) Non-destructive determination of carbohydrate content in potatoes using near infrared spectroscopy. Jpn J Food Eng 11:59–64. https://doi.org/10.11301/jsfe.11.59
35. Osborne BG, Fearn T, Hindle P (2003) Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman Scientific and Technical, London
36. Clark DH, Lamb RC (1991) Near infrared reflectance spectroscopy: a survey of wavelength selection to determine dry matter digestibility. J Dairy Sci 74:2200–2205. https://doi.org/10.3168/jds.S0022-0302(91)78393-8
37. Shenk JS, Workman J, Westerhaus M (2008) Application of NIR spectroscopy to agricultural products. In: Burns DA, Ciurczac EW (eds) Handbook of near infrared analysis. CRC Press, Taylor & Francis Group, Boca Raton, FL, pp 347–386
38. Whitehead C, Gomez LD, McQueen-Mason SJ (2012) The analysis of saccharification in biomass using an automated high-throughput method. Methods Enzymol 510:37–50. https://doi.org/10.1016/B978-0-12-415931-0.00003-3
39. Whitehead C, Ostos Garrido FJ, Reymond M et al (2018) A glycosyl transferase family 43 protein involved in xylan biosynthesis is associated with straw digestibility in Brachypodium distachyon. New Phytol 218:974–985. https://doi.org/10.1111/nph.15089
40. Ostos Garrido FJ, Pistón F, Gómez LD, McQueen-Mason SJ (2018) Biomass recalcitrance in barley, wheat and triticale straw: Correlation of biomass quality with classic agronomical traits. PLoSONE 13:e0205880. https://doi.org/10.1371/journal.pone.0205880
41. Shenk JS, Workman JJ, Westerhaus MO (2001) Application of NIR spectroscopy to agricultural products. In: Burns DA, Ciurczak EW (eds) Handbook of near infrared analysis. Marcel Dekker, New York, pp 419–474
42. Chadalavada K, Anbazhagan K, Ndour A, Choudhary S, Palmer W, Flynn JR et al (2022) NIR instruments and prediction methods for rapid access to grain protein content in multiple cereals. Sensors 22:3710. https://doi.org/10.3390/s22103710
43. Quentin AG, Rodemann T, Doutreleau MF, Moreau M, Davies NW (2017) Application of near-infrared spectroscopy for estimation of non-structural carbohydrates in foliar samples of Eucalyptus globulus Labilladière. Tree Physiol 37(1):131–141. https://doi.org/10.1093/treephys/tpw083
44. Fearn T (2002) Assessing Calibrations: SEP, RPD, RER and R2. NIR News 13:12. https://doi.org/10.1255/nirn.689
45. Wu Y, Peng S, Xie Q, Han Q, Zhang G, Sun H (2019) An improved weighted multiplicative scatter correction algorithm with the use of variable selection: application to near-infrared spectra. Chemom Intell Lab Syst 185:114–21. https://doi.org/10.1016/j.chemolab.2019.01.005
46. Martens H, Næs T (1987) Multivariate calibration by data compression. In: Williams P, Norris K (eds) Near-infrared technology in the agricultural and food industries, American Association of Cereal Chemists, Inc: St. Paul. Minnesota, pp 57–88
47. Bresolin T, Dórea JRR (2020) Infrared spectrometry as a high-throughput phenotyping technology to predict complex traits in livestock systems. Front Genet 11:923. https://doi.org/10.3389/fgene.2020.00923
48. van der Weijde T, Dolstra O, Visser RGF, Trindade LM (2017) Stability of cell wall composition and saccharification efficiency in Miscanthus across diverse environments. Front Plant Sci 7:2004. https://doi.org/10.3389/fpls.2016.02004
49. Minet O, Baeten V, Lecler B, Dardenne P, Fernández Pierna JA (2019) Local vs global methods applied to large near infrared databases covering high variability. ICNIRS 17:45–49. https://doi.org/10.1255/nir2017.045
50. Baeten V, Rogez H, Fernández Pierna JA, Vermeulen P, Dardenne P (2015) Vibrational spectroscopy methods for the rapid control of agro-food products. In: Nollet LML, Toldra F (eds) Handbook of food analysis, 3rd edn. CRC Press, Boca Raton, pp 591–614. https://doi.org/10.1201/b18668

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.