Received: 4 November 2014 ■ *Revised: 27 January 2015* ■ *Accepted: 29 January 2015* ■ *Publication: 17 February 2015*

JNIRS JOURNAL OF NEAR INFRARED SPECTROSCOPY

Prediction of chemical characteristics of fibrous plant biomasses from their near infrared spectrum: comparing local versus partial least square models and cross-validation versus independent validations

Bruno Godin,* Richard Agneessens, Jérôme Delcarte and Pierre Dardenne

Walloon Agricultural Research Center—CRA-W. Valorisation of Agricultural Products Department—Biomass, Bioproducts and Energy Unit, Chaussée de Namur, 146. B-5030 Gembloux, Belgium. E-mail: b.godin@cra.wallonie.be

The reliability of local and partial least square (PLS) near infrared (NIR) models to predict the chemical characteristics of fibrous plant biomasses was compared. Validations with different degrees of independence were used. The developed NIR models were reliable for the prediction of different main chemical characteristics of various fibrous plant species using multispecies datasets. The local models were more reliable in terms of prediction error compared with the PLS models because the local method appears to cope with the non-linearity and non-homogeneity associated with a large multispecies dataset. The degree of independence of samples in the validation set relative to samples used in the calibration set had a major impact on the prediction performance, especially for the local method. It affected the local method more because of the lower number of samples used in its specific regressions. There was a decrease in the reliability of local and PLS models according to the increase in the degree of independence of the validation set (i.e. the similarity of the predicted samples in regard to the calibration samples). The additions of a few independent samples of the predicted plant-species group to their calibration set that did not contain samples of the predicted plant-species group improved the prediction performance of multispecies models, especially for the local method. The type of NIR models developed in the present study can be used for screening, ranking and quantitative analyses of the main chemical components contents in fibrous biomasses, and for the assessment of their suitability to be converted into biofuels.

Keywords: chemical characteristics, biomass, near infrared, multivariate data analysis, chemometrics, validation

Introduction

Fibrous plant biomasses are an important potential source of renewable fuels and chemicals because of their great availability and sustainability.[1–3] They represent therefore an important biomass resource for a bio-based economy. The chemical characteristics (e.g. cellulose, hemicelluloses, lignin, starch, soluble sugars, protein and mineral compounds) of fibrous plant biomasses vary widely depending on genetics, growth environments, storage conditions, harvesting methods and periods.[4] This variability in chemical characteristics is difficult to control, but it determines the suitability of fibrous plant biomasses to be converted by biological (e.g. ethanolic and methanogenic fermentation) and thermochemical (e.g. combustion, pyrolysis and gasification) processes, and the efficiency of these conversion processes.[2,3,5]

ISSN: 0967-0335
doi: 10.1255/jnirs.1138

To monitor the variability in plant biomass chemical characteristics amounts and composition (e.g. monosaccharidic composition of hemicelluloses) and to adjust the process parameters of their conversion, robust analytical methods are needed.[4] Currently, standard wet chemical methods are used to determine the chemical characteristics of plant biomasses. These methods are reliable, but they are also often tedious, resource- and time-consuming, and expensive, and/or use hazardous chemicals. A broad chemical characterisation of plant biomasses using standard wet chemical methods costs US$800–2000 per sample and takes numerous days to complete.[4] Thus, reliable alternative methods without the disadvantages of wet chemical methods are needed for screening large numbers of samples and for industrial process optimisation.

Near infrared (NIR) spectroscopy is one such alternative method. This simple, fast, cheap, clean, non-destructive and reliable method is widely used for the quantitative and qualitative analysis of pharmaceutical, food, feed and plant products.[6] This method can predict different variables for about US$10 (1–5% of the wet chemistry procedure cost) per sample and can be used for online process control.[4] The main disadvantage of the NIR method is that it is a secondary analytical method. Consequently, it must be calibrated to a primary (reference) analytical method. For this calibration, advanced multivariate models are now used.[6] To calculate a reliable value of a chemical characteristic based on information obtained from an NIR spectrum, the calibration model must be built with:[7] (1) data determined by an accurate primary (reference) method; (2) a dataset containing a large variability of the predicted population (spectral variability) and of the chemical characteristic (large concentration range). At least 100–300 samples are usually used to build a robust prediction model of a chemical characteristic for a given population of an agricultural product. Therefore, the development of such NIR models for a broad chemical characterisation of a given population of plant biomasses is expensive and resource- and time-consuming, costing about US$300,000.[4]

NIR predictions are generally based on linear multivariate models such as partial least square (PLS) regression. To obtain a realistic prediction performance in terms of prediction error, these models require a large number of samples that are representative of the whole population variability to cover its spectral space. A large number of samples will usually reduce the accuracy in terms of the prediction error of these models.[7] One technique that can improve the accuracy of a prediction model when there are a large number of samples in the dataset (e.g. multiproduct) is to split the samples into small specific datasets (e.g. per type of product or species). This procedure will reduce the non-linearity present in a large dataset.[7] To solve the issue of splitting a dataset into small specific datasets, the local (specific regression and non-linear) method can be used, for example the local method of Shenk *et al.*[8] This local method builds a specific PLS regression with a low number of samples for each sample by selecting its most similar spectral neighbours from the library based on the highest correlation between spectra. To minimize the under- and overfitting of local prediction models, the optimum number of selected samples and the minimum and maximum PLS components for its specific regressions have to be determined.[8] The local method enables a large dataset (e.g. multiproduct) to be used to obtain accurate predictions, in terms of prediction error, and to have realistic prediction performance because a specific regression is built for each sample. The prediction model is less influenced by non-linearity and non-homogeneity present in a large dataset.[7,9] Local prediction models are considered to be 10–30% more reliable depending on the spectral diversity of the dataset, as compared with the PLS prediction models.[7,9,10]

The aim of this study was to compare the reliability of local (specific regression and non-linear) and PLS (linear regression) NIR models to predict the chemical characteristics of fibrous plant biomasses. The study also used multispecies datasets, to predict numerous chemical characteristics and to test the influence of the type of validation on the parameters used to assess the prediction performance of the models.

The chemical characteristics that have been considered in the present study were: (1) cellulose and hemicelluloses, the most abundant structural carbohydrates in nature and the resource for the production of cellulosic ethanol; the optimal conversion of the hemicelluloses into fuels and chemicals depends on its monosaccharidic composition;[2,3,5] (2) lignin, the most abundant component in nature with an aromatic ring structure;[3] (3) water-soluble glucose, fructose and sucrose, non-structural carbohydrates that can be fermented without a complex pretreatment of the biomass;[3] (4) protein, the source of organic nitrogen of plant biomasses;[3] (5) mineral compounds, which represent the inorganic part of plant biomasses and its complement, which is the organic matter; (6) the higher heating value (HHV), which can be used to assess the thermal energy of biomasses;[2,5] (7) the enzymatically digestible organic matter (eDOM) and the biochemical methane potential (BMP), which are relevant methods to assess the suitability of the plant biomasses to be converted by anaerobic digestion.[5,11]

Material and methods

Biomass material

The analysed samples consisted of a "grass–alfalfa" mixture (*Dactylis glomerata* L.– *Medicago sativa* L.; cultivar Terrano—*Medicago sativa* L. cultivar Europe; three harvest cycles late spring–late summer–late autumn; 38–30–30 samples), cocksfoot (*Dactylis glomerata* L.; cultivar: Terrano; three harvest cycles late spring–late summer–late autumn; 6–6–2 samples), fibre corn (*Zea mays* L.; cultivars: Aayrton, Athlético, Aventura, Beethoven, Cannavaro, Coryphée, Dominator, Franky, Ladifférence, LG Azelo, Olympus, Ricardinio and Ronaldinio; early autumn and late winter harvest, respectively, 237 and 17 samples), fibre sorghum [*Sorghum bicolor* (L.) Moench; cultivars: CA25, ENR10, H133, Maja and Zerberus; early autumn and late winter harvest, respectively, 179 and 58 samples],

hemp (*Cannabis sativa* L.; cultivars: Epsilon 68, Fedora 17 and Futura 75; early autumn harvest; 109 samples), immature rye (*Secale cereale* L.; cultivars: Protector and Vitalio; early spring harvest; 45 samples), immature spelt [*Triticum aestivum* L. ssp. *spelta* (L.) Thell.; cultivars: Badengold and Cosmos; late spring harvest; nine samples], Jerusalem artichoke leaves and stalks (*Helianthus tuberosus* L.; cultivar: Volkenroder spindel; early autumn harvest; 62 samples), miscanthus giganteus (*Miscanthus × giganteus* J.M. Greef & Deuter ex Hodk. & Renvoize; cultivars: Bical; early autumn and late winter harvest, respectively, 34 and 178 samples), spelt straw (*Triticum aestivum* L. ssp. *spelta* (L.) Thell.; cultivars: Badengold and Cosmos; late summer harvest; 95 samples), switchgrass (*Panicum virgatum* L.; cultivars: Alamo, Blackwell, Cave-in-Rock, Dacotah, Kanlow, Nebraska 28, Shelter and Traiblazer; early autumn and late winter harvest, respectively, 25 and 186 samples) and tall fescue (*Festuca arundinacea* Schreb.; cultivars: Hykor, Jordane, Kora, Perun and Soni; three harvest cycles late spring–late summer–late autumn; 215–217–212 samples). These samples came from randomised block trials performed in 2008, 2009, 2010 and/or 2011 at Libramont [498 m above sea level (asl); average annual temperature: 8.6°C; average annual precipitation: 1260 mm; 49°55′N, 05°24′E; Belgium], Gembloux (161 m asl; average annual temperature: 9.8°C ; average annual precipitation: 856 mm; 50°33′N, 04°43′E), Tinlot (255 m asl; average annual temperature: 9.7°C; average annual precipitation: 871 mm; 50°28′N, 05°23′E; Belgium), Mötsch (330 m asl; average annual temperature: 8.4°C; average annual precipitation: 675 mm; 49°57′N, 06°33′E; Germany) or Gerbéviller (260 m asl; average annual temperature: 9.9°C; average annual precipitation: 1022 mm; 48°29′N, 06°31′E; France). The crops came from trials that were performed with different harvest periods, cultivars and/or nitrogen fertilisation levels (0–240 kg of nitrogen per hm^2). From plots between 9 m^2 and 24 m^2, the whole above-ground biomass was harvested at 10 cm from the ground and chopped (particle size 1–2 cm).

For the data analyses, the small groups of cocksfoot, "grass–alfalfa" mixture, immature spelt and immature rye samples were incorporated into the large group of tall fescue. These samples of plant species were merged together into one group of plant species because they are similar plant species. This enabled the prediction of these small groups of plant species. This group of merged samples of similar plant species is called the grasses group.

Two representative fresh subsamples of 750 g each were dried at 60°C for 72 h in a forced air oven immediately after the harvest. After the drying process, the two subsamples of the green-dried form of biomass were milled first using a BOA hammer mill (Waterleau, Herent, Belgium) through a 4 mm screen and then milled using a Cyclotec cyclone mill (FOSS, Hillerød, Denmark) through a 1 mm screen. For the storage of the samples, airtight bags were used, kept at room temperature and protected from the light in a dark box.

In addition, a third sample was packed in a plastic bag under vacuum to simulate the silage process. The vacuum-sealed plastic bag containing the silage-wet form of biomass was stored at room temperature for at least 3 weeks before the laboratory analysis. If gas was produced during silaging, the plastic bag was opened and put again under vacuum.

Chemical reagents and analyses

All chemicals were of analytical grade or equivalent. Various chemical characteristics were determined for the analysed biomasses.

Neutral detergent fibre (NDF) residue corrected for its mineral compound), acid detergent fibre (ADF) residue corrected for its mineral compounds and acid detergent lignin (ADL) residue corrected for its mineral compounds were determined by the Van Soest (VS) method.[12–15] These residues were used to estimate the cellulose VS (ADF–ADL), hemicelluloses VS (NDF–ADF) and lignin VS (ADL) content of plant biomasses.[12–15] The sulfuric acid hydrolysis (SAH) method of Godin *et al.*[5,14,16] was used to determine the cellulose SAH, hemicelluloses SAH, xylan, arabinan, mannan, galactan, hemicellulosic glucan and insoluble Klason lignin (corrected for its mineral compounds) content. The cellulose SAH (cellulosic glucan; i.e. D-glucose of cellulose in its polymeric form) content was calculated as the difference between the total glucan (cellulosic and hemicellulosic glucan) and hemicellulosic glucan content. The hemicelluloses SAH content was calculated as the sum of the xylan, arabinan, galactan, mannan and hemicellulosic glucan content. The monosaccharidic compounds (D-xylose, L-arabinose, D-glucose, D-mannose and D-galactose) of hemicelluloses were expressed in their polymeric form (xylan, arabinan, hemicellulosic glucan, mannan and galactan).[5,14,16]

The total soluble sugars were determined by the Luff–Schoorl method.[5,17] The water-soluble sucrose, glucose and fructose content were measured by a liquid chromatography method.[5] The protein content was measured by the Kjeldahl method using 6.25 as a conversion factor of nitrogen to protein.[5,18] The mineral compounds content was determined by the use of a muffle furnace set at 550°C for 3 h. The organic matter (OM) content was the complement of the mineral compounds (calculated as 100 minus mineral compounds content). The HHV was measured using a Parr controlled oxygen bomb calorimeter.[5,19] The HHV can be used to assess the thermal energy of biomasses.[5] The enzymatically digestible organic matter (eDOM) was determined by the De Boever method.[20] This relatively simple and fast method can be used to assess the suitability of the plant biomasses to be converted by anaerobic digestion.[5,21,22] The eDOM can be considered as the minimum level of anaerobic digestibility of the plant biomass. Indeed, the micro-organisms of the anaerobic digestion are expected to produce more enzymes *in situ* and for a longer period of time compared with the enzyme cocktail used in the analysis.[5,21,22] The BMP was determined according to the VDI 4630 standard as described by Mayer *et al.*[23] The volumes of biomethane were normalised at 0°C and 1013 hPa according to the temperature and pressure conditions of each measurement. BMP is the most relevant method used to determine the biogas production potential of biomasses.[11]

All the measurements were carried out on the green-dried form of the biomass, except for the BMP analysis, which was carried out on the silage-wet form of the biomass because this form of biomass is closer to the industrial biomethanation process. Duplicate aliquots were measured on each sample, except for the BMP analysis, for which one aliquot was measured on each sample. The dry matter (DM) content of the samples dried at 60°C for 72 h was determined at 103°C for 4 h.

Near infrared analysis

The near infrared (NIR) reflectance spectra were obtained using an NIRSystems 5000 (FOSS) on the biomass samples in green-dried form in a Black Metal Ring Cup (FOSS). Each spectrum was collected in the range of 1100–2498 nm and was the average of 32 scans. The spectra were normalised by a standard normal variate (SNV) transformation followed by a first-order derivation (1, 4, 4, 1; first derivative, 4 nm gap, 4 points of first smoothing, 1 point of second smoothing).

Statistical analysis

Descriptive statistics were obtained using JMP 11 (SAS Institute, Cary, NC). The local (specific regression and nonlinear procedure)[8] and PLS (Modified-PLS algorithm; linear regression model) techniques were used to develop prediction models. The spectral preprocessing, local and PLS procedures were performed with WinISI 4.6.8 (FOSS)

The specific factors for each local and PLS model were optimised according to the software (WinISI 4.6.8). The aim of this optimisation was to build models with the best prediction performance in terms of prediction error to minimize the under- and overfitting of the model. These factors were the optimum number of selected samples and the minimum and maximum PLS components for the local prediction models (details in Table S1 of the supplementary data). The factor to optimise for the PLS prediction models was the number of PLS components (details in Table S1 of the supplementary data). The local models used a higher number of PLS components than the PLS models (details in Table S1 of the supplementary data). For the local method, the final prediction PLS component number for a given sample was calculated using the weighted average of all the prediction values obtained from the minimum to the maximum PLS. The weights calculated were inversely proportional to the value of the β-coefficients and the residuals of the PLS regressions.[8]

To prevent an overestimation of the prediction performance of the models, each was evaluated using independent validation (V) datasets in addition to a leave-one-out cross-validation (*CV-LOO*). In a leave-one-out cross-validation, each sample was predicted by an equation developed from whole the other samples. Cross-validation *CV-LOO* was used to assess the prediction error for the samples of the library. For independent validation, a group of independent samples (validation set) was predicted by a model developed using other samples (calibration set). Samples in a validation set did not come from the same cropping site, year, harvest period and/or species as the samples of the calibration set.[24]

The independent validation dataset V1 contained samples of each plant-species group and was built by manually splitting the whole dataset into a calibration set and a validation set. The validation set contained approximately 20% of total samples (approximately 20% per plant-species group) and comprised representative and independent samples with regard to the calibration set. These two sets were considered independent because, for each plant-species group, the samples of these two sets were not from the same cropping site, year or harvest period. Validation V1 was designed to assess the prediction error of future new samples of plant species contained in the library.

For each of the eight other independent validation datasets (fibre corn, fibre sorghum, grasses, hemp, Jerusalem artichoke, miscanthus giganteus, spelt straw and switchgrass), V2 only contained the samples of one plant-species group. All the samples of one plant-species group were predicted using all the samples of the seven other plant-species groups. For the independent validation V2, the predicted values of each plant-species group were corrected by the median prediction bias of its plant-species group. The bias correction made it possible to eliminate the systematic error that can occur when a plant species is predicted without being in the calibration set. The prediction results of each of the eight independent validation datasets V2 were pooled to assess the prediction performance of the models. Validation V2 was designed to assess the prediction error of future samples of plant species that were similar to but not contained in the library used to develop the calibration model.

The samples of validation V2 were the most independent with regard to the samples of their respective calibration set. To reduce this degree of independence of validation V2, a few independent samples (5, 10, 15, 20 and 25) of the predicted plant-species group were added to their calibration sets. For each plant-species group, these 25 samples were randomly selected and excluded from being in the validation V2 before their addition to their respective calibration set. For each level of addition, the prediction results of each of the eight new independent validations were pooled together to assess the prediction performance of the models. This analysis was conducted for only a few chemical characteristics: NDF, ADF, ADL and mineral compounds. These characteristics were chosen because many of the samples had data for these constituents enabling estimates of cellulose, hemicelluloses, lignin and mineral compounds of fibrous biomass.

To evaluate the prediction performance of the models, the following parameters were determined for the cross-validation *CV-LOO*, validation V1 and validation V2: the coefficient of determination of prediction based on medians (r^2Med) [Equation (1)]; the median standard residual error of prediction (*MedRE*) [Equation (2)] (instead of the root mean square error of prediction, *RMSEP*);[25] the ratio of the median standard deviation (*SDMed*) of the variable to *MedRE* ($RPDMed = SDMed * MedRE^{-1}$); the ratio of *MedRE* to the standard error of laboratory (*SEL*); and the median spectral distance of Mahalanobis (*GHMed*). These parameters were calculated based on medians to be

robust and to avoid deleting subjectively outlier samples (which have high residual values) of a given model without deleting them from other models. Therefore, the statistical parameters (r^2Med, $MedRE$, $RPDMed$, $MedRE*SEL^{-1}$ and $GHMed$) used to qualify the models were determined based on medians:

$$r^2Med = \frac{SDMed^2 - MedRE^2}{SDMed^2} \quad (1)$$

where $SDMed$ = median standard deviation of the variable; $MedRE$ = median standard residual error of prediction:

$$MedRE = MAD*1.4826 \quad (2)$$

where MAD = median of the absolute deviation of the residuals.[25]

In order to evaluate the reliability of the prediction performance, the r^2Med and the RDPMed of each prediction model were used. The following guidelines for agricultural products were suggested by Malley *et al.*[26] for these parameters: excellent prediction model, $r^2Med \geq 0.95$ and $RPDMed \geq 4.0$; successful prediction model, $r^2Med \geq 0.90$ and $RPDMed \geq 3.0$; moderately successful prediction model, $r^2Med \geq 0.80$ and $RPDMed \geq 2.3$; moderately useful prediction model (semi-quantitative), $r^2Med \geq 0.70$ and $RPDMed \geq 1.8$.

Results and discussion

Predicted chemical characteristics

The chemical characteristics of the analysed samples are listed in Table 1. The samples were fibrous plant biomasses: fibre corn, fibre sorghum, grasses, hemp, Jerusalem artichoke leaves and stalks, miscanthus giganteus, spelt straw and switchgrass. Chemical characteristics have been analysed by two different methods, because they are both often used in commercial laboratories, in the case of cellulose, hemicelluloses, lignin and water-soluble sugars, and the suitability to be converted by anaerobic digestion (eDOM and BMP). The analysis of hemicelluloses by the SAH method had the advantages of yielding carbohydrate composition and being more accurate, compared with the VS method. For the analysis of the water-soluble sugars, the liquid chromatography method had the same advantages as the Luff–Schoorl method.[5,14,16]

The content and variability of the determined chemical characteristics of the analysed biomasses have been reported previously[5,21,22,27] and are related to the plant species, harvest periods (autumn and winter) and cropping conditions (year, area, cultivar and nitrogen fertilisation levels).

Table 1. Whole dataset summary of the reference values of the chemical characteristics.

Chemical characteristic	*n*	Min.	Max.	Median	Median *SD*	Interplant-species group *SD*	Intraplant-species group *SD*	Relative *SEL* (%)
NDF[a] (g 100 g^{-1} DM)	1169	29.59	91.40	66.77	18.04	59.69	7.01	0.6
ADF[a] (g 100 g^{-1} DM)	1167	17.15	70.91	42.31	15.08	50.04	6.31	0.7
ADL[a] (g 100 g^{-1} DM)	1167	1.13	13.59	6.54	3.91	12.99	1.41	2.3
Insoluble Klason lignin[a] (g 100 g^{-1} DM)	112	2.62	17.01	8.92	4.39	4.38	1.86	3.4
Cellulose SAH (g 100 g^{-1} DM)	413	13.32	50.63	28.47	8.79	21.26	3.80	2.8
Hemicelluloses SAH (g 100 g^{-1} DM)	413	7.16	34.30	21.68	4.42	13.20	1.99	2.8
Xylan (g 100 g^{-1} DM)	413	4.22	27.80	15.12	4.80	12.06	1.92	3.3
Arabinan (g 100 g^{-1} DM)	413	0.31	4.28	2.51	0.59	2.21	0.35	4.8
Mannan (g 100 g^{-1} DM)	413	0.01	2.74	0.57	0.22	1.14	0.20	11
Galactan (g 100 g^{-1} DM)	413	0.25	3.43	0.87	0.46	1.41	0.31	11
Hemicellulosic glucan (g 100 g^{-1} DM)	413	0.63	05.26	1.95	0.64	1.03	0.62	10
Total soluble sugars (g 100 g^{-1} DM)	643	0.01	38.23	5.50	7.44	17.97	5.90	4.5
Sum of soluble sucrose + glucose + fructose (g 100 g^{-1} DM)	269	0.19	17.18	5.23	2.75	34.65	27.54	3.8
Soluble sucrose (g 100 g^{-1} DM)	269	0.01	8.12	1.85	1.71	16.75	14.90	4.3
Soluble glucose (g 100 g^{-1} DM)	269	0.06	5.49	0.95	0.62	13.86	7.89	7.4
Soluble fructose (g 100 g^{-1} DM)	269	0.03	5.51	1.80	1.18	10.72	10.74	6.1
Protein (g 100 g^{-1} DM)	864	0.62	23.03	5.64	3.87	11.41	3.07	3.5
Mineral compounds (g 100 g^{-1} DM)	1377	0.84	20.09	6.53	3.38	11.01	2.17	1.5
HHV (MJ kg^{-1} DM)	810	16.74	19.73	18.34	0.61	1.45	0.41	0.3
eDOM (g 100 g^{-1} OM)	484	7.29	83.59	38.64	23.26	1.16	0.26	1.9
BMP (dm^3 kg^{-1} OM)	587	147	635	387	80	205	56	5.2

[a]Residue corrected for its mineral compounds.

For each chemical characteristic, the number of analysed samples was generally high (n = 269–1377), except for the insoluble Klason lignin (n = 112). The concentration range of each chemical characteristic was usually very wide, except for the HHV. The HHV is known to have a relatively stable value.[5,21,22] The interplant-species group *SD* of the analysed chemical characteristic was generally considerably higher than the intraplant-species group *SD*, except for the sum of soluble sucrose + glucose + fructose, soluble sucrose, soluble glucose and soluble fructose because the analysed plant-species groups had similar concentration ranges for these parameters. The data for each chemical characteristic were therefore well spread across its concentration range. This can be explained by the diversity of the whole dataset, which comprised samples of various plant species coming from different harvest periods and cropping conditions (year, area, cultivar, nitrogen fertilisation levels). To develop reliable NIR prediction models, it is important to have a large number of samples uniformly distributed over a wide concentration range.

The relative values of the *SEL* (relative *SEL* = *SEL**median^{-1}) were usually low, except for the contents of mannan, galactan, hemicellulosic glucan, soluble glucose, soluble fructose and the BMP where the relative *SEL* values were above 5% (Table 1). Having a low relative *SEL* of the primary method is also a critical precursor for achieving a reliable NIR prediction model.

Thus, based on the aforementioned criteria (number of analysed samples, concentration range, distribution of the data around the concentration range and *SEL*), the whole dataset used to build NIR prediction models was quite suitable.

Comparison of local and partial least square models

The prediction performances (r^2*Med*, *RPDMed*, *MedRE*, *MedRE***SEL*$^{-1}$ and *GHMed*) of the NIR predicted chemical

Table 2. r^2*Med* and *RPDMed* for the cross-validation *CV-LOO*, validation V1 (calibration containing samples of the predicted plant-species group) and validation V2 (calibration not containing samples of the predicted plant-species group) of the prediction models of the assessed chemical characteristics.

Chemical characteristic	r^2*Med*						*RPDMed*					
	Cross-validation *CV-LOO*		Validation V1		Validation V2		Cross-validation *CV-LOO*		Validation V1		Validation V2	
	Local	PLS	Local	PLS	Local	PLS	Local	PLS	Local	PLS	Local	PLS
NDF[a]	0.997	0.994	0.997	0.996	0.990	0.990	19	13	19	15	9.4	9.2
ADF[a]	0.997	0.995	0.996	0.994	0.990	0.990	19	14	15	13	9.5	8.7
ADL[a]	0.995	0.987	0.98	0.98	0.97	0.97	14	8.9	7.9	7.9	6.0	5.6
Insoluble Klason lignin[a]	0.97	0.97	0.97	0.96	0.95	0.92	6.2	5.9	5.7	5.1	4.5	3.6
Cellulose SAH	0.991	0.990	0.97	0.97	0.98	0.97	11	8.5	5.6	5.5	6.5	5.8
Hemicelluloses SAH	0.98	0.97	0.95	0.93	0.950	0.90	6.5	5.4	4.6	3.8	4.5	3.2
Xylan	0.990	0.98	0.98	0.96	0.97	0.96	8.9	7.3	7.1	4.8	5.3	5.1
Arabinan	0.93	0.88	0.93	0.91	0.74	0.79	3.7	2.8	3.7	2.0	2.0	2.2
Mannan	0.85	0.60	0.92	0.83	0.00	0.00	2.6	1.6	3.6	2.4	0.8	0.7
Galactan	0.93	0.84	0.79	0.77	0.55	0.68	3.8	2.5	2.2	2.1	1.5	1.8
Hemicellulosic glucan	0.81	0.70	0.59	0.30	0.34	0.17	2.3	1.8	1.6	1.2	1.2	1.1
Total soluble sugars	0.994	0.990	0.991	0.98	0.97	0.94	13	8.6	10	7.9	5.6	4.1
Sum of soluble sucrose + glucose + fructose	0.94	0.91	0.88	0.87	0.63	0.62	4.0	3.2	2.9	2.8	1.6	1.6
Soluble sucrose	0.92	0.80	0.64	0.59	0.58	0.43	3.5	2.2	1.7	1.6	1.5	1.3
Soluble glucose	0.79	0.59	0.83	0.53	0.00	0.00	2.2	1.6	2.4	1.5	0.6	0.7
Soluble fructose	0.93	0.88	0.82	0.92	0.53	0.25	3.7	2.9	2.3	3.5	1.5	1.2
Protein	0.993	0.990	0.98	0.97	0.97	0.97	12	8.4	7.2	6.0	6.1	6.0
Mineral compounds	0.990	0.96	0.97	0.95	0.93	0.92	8.3	5.1	5.8	4.3	3.9	3.6
HHV	0.96	0.94	0.91	0.87	0.80	0.86	4.7	4.2	3.4	2.8	2.3	2.6
eDOM	0.993	0.990	0.98	0.97	0.98	0.96	12	8.7	6.7	5.8	6.6	4.8
BMP	0.85	0.80	0.81	0.83	0.49	0.25	2.6	2.2	2.3	2.4	1.4	1.2

[a]Residue corrected for its mineral compounds.

characteristics of local and PLS models for each type of validation (cross-validation *CV-LOO*, validation V1 and validation V2) are listed in Table 2 for r^2*Med* and *RPDMed*, Table 3 for *MedRE*, Table 4 for *MedRE*SEL*$^{-1}$ and Table 5 for *GHMed*. To obtain a good prediction model for a given characteristic and dataset, the aim was to have the lowest *MedRE* and *GHMed*. To compare the performances of prediction models of different characteristics and datasets, the r^2 and RPD were considered because they are independent of the used units. The r^2 and RDP are directly linked to each other. The RDP was interesting in addition to the r^2 because the RPD is more discriminatory than the r^2 when r^2 is close to 1.[24]

For each type of validation, the local prediction models were generally more reliable than the PLS prediction models, because with the local method, a specific regression with a low number of samples was built for each sample by selecting its most similar spectral neighbours from the library based on the highest correlation between spectra. In the present study, the optimum number of selected samples based on the median was 100 samples (Table S1 of the supplementary information). The similar spectral neighbours were presumably samples of the same nature and plant species (i.e. lower independence between samples). This enabled a prediction to negate non-linearity and non-homogeneity associated with a large multispecies dataset.[7,9]

The chemical characteristics that had a lower prediction performance for each type of validation were arabinan, mannan, galactan, hemicellulosic glucan, sum of soluble sucrose + glucose + fructose, soluble sucrose, soluble glucose and soluble fructose and BMP (Tables 2–5). They had high relative *SEL* and low concentration levels. The relative *SEL* and concentration levels are linked together by the Horwitz curve.[28] This curve shows that the relative *SEL* increases with decreasing levels of concentration. Therefore, chemical characteristics with low levels of concentration were predicted less reliably because of the high relative *SEL* of the primary analytical methods used to build the prediction models. This was especially the case for mannan, which had the highest relative

Table 3. Median standard residual error of prediction (*MedRE*) for the cross-validation *CV-LOO*, validation V1 (calibration containing samples of the predicted plant-species group) and validation V2 (calibration not containing samples of the predicted plant-species group) of the prediction models of the assessed chemical characteristics.

Chemical characteristic		*MedRE*					
		Cross-validation *CV-LOO*		Validation V1		Validation V2	
	SEL	Local	PLS	Local	PLS	Local	PLS
NDF[a] (g 100 g^{-1} DM)	0.40	0.96	1.41	1.06	1.28	1.92	1.96
ADF[a] (g 100 g^{-1} DM)	0.30	0.81	1.05	1.01	1.17	1.58	1.73
ADL[a] (g 100 g^{-1} DM)	0.15	0.29	0.44	0.50	0.50	0.65	0.70
Insoluble Klason lignin[a] (g 100 g^{-1} DM)	0.30	0.71	0.75	0.76	0.85	0.97	1.23
Cellulose SAH (g 100 g^{-1} DM)	0.80	0.83	1.04	1.50	1.53	1.36	1.52
Hemicelluloses SAH (g 100 g^{-1} DM)	0.60	0.68	0.81	0.74	0.90	0.99	1.37
Xylan (g 100 g^{-1} DM)	0.50	0.54	0.66	0.58	0.86	0.90	0.95
Arabinan (g 100 g^{-1} DM)	0.12	0.16	0.21	0.22	0.24	0.30	0.27
Mannan (g 100 g^{-1} DM)	0.06	0.09	0.14	0.13	0.19	0.28	0.31
Galactan (g 100 g^{-1} DM)	0.10	0.12	0.18	0.23	0.24	0.31	0.26
Hemicellulosic glucan (g 100 g^{-1} DM)	0.20	0.28	0.35	0.37	0.48	0.52	0.58
Total soluble sugars (g 100 g^{-1} DM)	0.25	0.59	0.86	0.74	0.97	1.34	1.81
Sum of soluble sucrose + glucose + fructose (g 100 g^{-1} DM)	0.20	0.69	0.85	0.81	0.85	1.68	1.69
Soluble sucrose (g 100 g^{-1} DM)	0.08	0.49	0.76	1.12	1.20	1.11	1.29
Soluble glucose (g 100 g^{-1} DM)	0.07	0.29	0.40	0.36	0.60	1.11	0.92
Soluble fructose (g 100 g^{-1} DM)	0.11	0.32	0.41	0.50	0.34	0.81	1.02
Protein (g 100 g^{-1} DM)	0.20	0.33	0.46	0.45	0.54	0.63	0.65
Mineral compounds (g 100 g^{-1} DM)	0.10	0.41	0.67	0.58	0.77	0.87	0.95
HHV (MJ kg^{-1} DM)	0.06	0.13	0.15	0.15	0.18	0.27	0.23
eDOM (g 100 g^{-1} OM)	0.75	2.01	2.68	3.38	3.94	3.54	4.80
BMP (dm^3 kg^{-1} OM)	20	31	36	32	30	57	69

[a]Residue corrected for its mineral compounds.

Table 4. Ratio of the median standard residual error of prediction (*MedRE*) to the *SEL* for the cross-validation *CV-LOO*, validation V1 (calibration containing samples of the predicted plant-species group) and validation V2 (calibration not containing samples of the predicted plant-species group) of the prediction models of the assessed chemical characteristics: ratio of *MedRE* of validation to *MedRE* of cross-validation *CV-LOO* for these models.

Chemical characteristic	*MedRE*SEL*$^{-1}$						*MedRE*v**MedRE*cv^{-1}			
	Cross-validation *CV-LOO*		Validation V1		Validation V2		Validation V1		Validation V2	
	Local	PLS	Local	PLS	Local	PLS	Local	PLS	Local	PLS
NDF[a]	2.4	3.5	2.7	3.2	4.8	4.9	1.10	0.91	2.00	1.39
ADF[a]	2.7	3.5	3.4	3.9	5.3	5.8	1.25	1.11	1.95	1.65
ADL[a]	1.9	2.9	3.3	3.3	4.3	4.7	1.72	1.14	2.24	1.59
Insoluble Klason lignin[a]	2.4	2.5	2.5	2.8	3.2	4.1	1.07	1.13	1.37	1.64
Cellulose SAH	1.0	1.3	1.9	1.9	1.7	1.9	1.81	1.47	1.64	1.46
Hemicelluloses SAH	1.1	1.4	1.2	1.5	1.7	2.3	1.09	1.11	1.46	1.69
Xylan	1.1	1.3	1.2	1.7	1.8	1.9	1.07	1.30	1.67	1.44
Arabinan	1.3	1.7	1.8	2.0	2.5	2.3	1.38	1.14	1.88	1.29
Mannan	1.5	2.4	2.2	3.2	4.7	5.2	1.44	1.36	3.11	2.21
Galactan	1.2	1.8	2.3	2.4	3.1	2.6	1.92	1.33	2.58	1.44
Hemicellulosic glucan	1.4	1.7	1.9	2.4	2.6	2.9	1.32	1.37	1.86	1.66
Total soluble sugars	2.3	3.5	3.0	3.9	5.4	7.2	1.25	1.13	2.27	2.10
Sum of soluble sucrose + glucose + fructose	3.5	4.3	4.0	4.2	8.4	8.5	1.17	1.00	2.43	1.99
Soluble sucrose	6.1	9.5	14.0	15.0	13.8	16.2	2.29	1.58	2.27	1.70
Soluble glucose	4.1	5.7	5.2	8.6	15.9	13.2	1.24	1.50	3.83	2.30
Soluble fructose	2.9	3.7	4.5	3.0	7.4	9.3	1.56	0.83	2.53	2.49
Protein	1.6	2.3	2.3	2.7	3.2	3.3	1.36	1.17	1.91	1.41
Mineral compounds	4.1	6.7	5.8	7.7	8.7	9.5	1.41	1.15	2.12	1.42
HHV	2.2	2.4	2.5	3.0	4.5	3.8	1.15	1.20	2.08	1.53
eDOM	2.7	3.6	4.5	5.3	4.7	6.4	1.68	1.47	1.76	1.79
BMP	1.5	1.8	1.6	1.5	2.9	3.5	1.03	0.83	1.84	1.92

[a]Residue corrected for its mineral compounds.

SEL and lowest concentration. In addition, at low levels of concentration, the absorbance of the NIR peaks was reduced. Thus, the prediction models became less accurate for these low concentration levels.

The local prediction models of hemicellulosic glucan and soluble glucose for each type of validation had a considerably higher prediction performance than the PLS prediction models (Tables 2–5). Furthermore, the prediction performance for these chemical characteristics of validation V2 (calibration not containing samples of the predicted plant-species group) was especially low because these validation samples were highly independent relative to the calibration samples. The prediction models of hemicellulosic glucan and soluble glucose must have been impacted by the high cellulose content (Tables 2–5). This glucose polymeric carbohydrate was present in much higher amounts in the analysed biomasses, compared with hemicellulosic glucan and soluble glucose. Therefore, using NIR spectra, it is difficult to distinguish the lower contents of hemicellulosic glucose and soluble glucose from the much higher content of cellulosic glucose. With the local method, it was most likely possible to manage this issue because for each sample, a specific regression was built, as previously explained.

Comparison of cross-validation and independent validations

The three validations (cross-validation *CV-LOO*, validation V1 and validation V2) and their prediction performance (r^2*Med*, *RPDMed*, *MedRE*, *MedRE*SEL*$^{-1}$ and *GHMed*) of the NIR predicted chemical characteristics of local and PLS models are shown in Tables 2–5. Figure 1 shows the cross-validation *CV-LOO*, validation V1 and validation V2 of the local prediction models for hemicelluloses SAH, protein and HHV. These three validations were evaluated to prevent an overestimation of

Table 5. Median spectral distance of Mahalanobis (*GHMed*) for the cross-validation *CV-LOO*, validation V1 (calibration containing samples of the predicted plant-species group) and validation V2 (calibration not containing samples of the predicted plant-species group) of the prediction models of the assessed chemical characteristics.

Chemical characteristic	*GHMed*					
	Cross-validation *CV-LOO*		Validation V1		Validation V2	
	Local	PLS	Local	PLS	Local	PLS
NDF[a]	0.73	0.74	0.91	0.84	1.56	1.20
ADF[a]	0.70	0.70	0.89	0.85	1.39	1.19
ADL[a]	0.73	0.81	0.91	0.91	1.83	1.39
Insoluble Klason lignin[a]	0.83	0.81	1.10	0.98	1.08	1.00
Cellulose SAH	0.71	0.81	0.71	0.81	0.91	1.29
Hemicelluloses SAH	0.80	0.84	0.83	0.86	1.56	1.28
Xylan	0.68	0.85	0.82	0.83	1.13	1.43
Arabinan	0.84	0.88	0.84	0.88	1.98	1.90
Mannan	0.82	0.89	0.92	0.95	2.87	2.40
Galactan	0.81	0.90	0.96	0.98	2.60	2.33
Hemicellulosic glucan	0.82	0.89	1.09	0.95	4.17	3.04
Total soluble sugars	0.79	0.81	1.00	0.95	1.19	1.31
Sum of soluble sucrose + glucose + fructose	0.85	0.83	1.01	0.93	2.50	2.14
Soluble sucrose	0.84	0.85	0.98	0.98	2.77	4.01
Soluble glucose	0.84	0.84	1.17	1.08	2.54	4.12
Soluble fructose	0.91	0.82	1.55	0.96	3.22	2.34
Protein	0.73	0.77	0.94	0.84	1.65	1.19
Mineral compounds	0.81	0.80	0.86	0.78	1.82	1.56
HHV	0.84	0.87	1.05	1.00	2.85	1.97
eDOM	0.56	0.83	0.96	1.73	1.10	1.09
BMP	0.81	0.87	1.80	1.25	9.48	11.1

[a]Residue corrected for its mineral compounds.

the performances of the prediction models. Cross-validation *CV-LOO* was designed to assess the prediction error for the samples of the library. Validation V1 was designed to assess the prediction error of future new samples of plant species contained in the library. Validation V2 was designed to assess the prediction error of future new samples of plant species not contained in the library but similar to the plant species contained in the library.

The ratio of *MedRE* (median standard residual error of prediction) to *SEL* was used to compare the accuracy of local and PLS prediction models with the primary analytical method and according to the type of validation (Table 4). This ratio was independent of the units used. Based on this ratio, the following observations were made: (1) for the cross-validation *CV-LOO*, the predictions were 2.4 times less accurate than the primary analytical methods based on their median value and local models were 23% more accurate than PLS models based on their median value; (2) for the validation V1, the predictions were 2.9 times less accurate than the primary analytical methods based on their median value and local models were 14% more accurate than PLS models based on their median value; (3) for the validation V2, the predictions were 4.5 times less accurate than the primary analytical methods based on their median vale and local models were 8.7% more accurate than PLS models based on their median value.

The ratio of *MedRE* of validation (validation V1 or V2) to *MedRE* of cross-validation *CV-LOO* was used to compare the overestimation of the prediction performance of local and PLS models of cross-validation *CV-LOO* with regards to the independent validations (validation V1 and V2) (Table 4). This ratio was independent of the units used. Based on this ratio, the following observations were made: (1) for the validation V1 in regard to the cross-validation *CV-LOO*, the predictions were 1.25 times less accurate based on their median value and PLS models showed 14% less accuracy than local models based on their median value; (2) for the validation V2 in regard to the cross-validation *CV-LOO*, the predictions were 1.86 times less accurate based on their median value and PLS models showed 18% less accuracy than local models based on their median value.

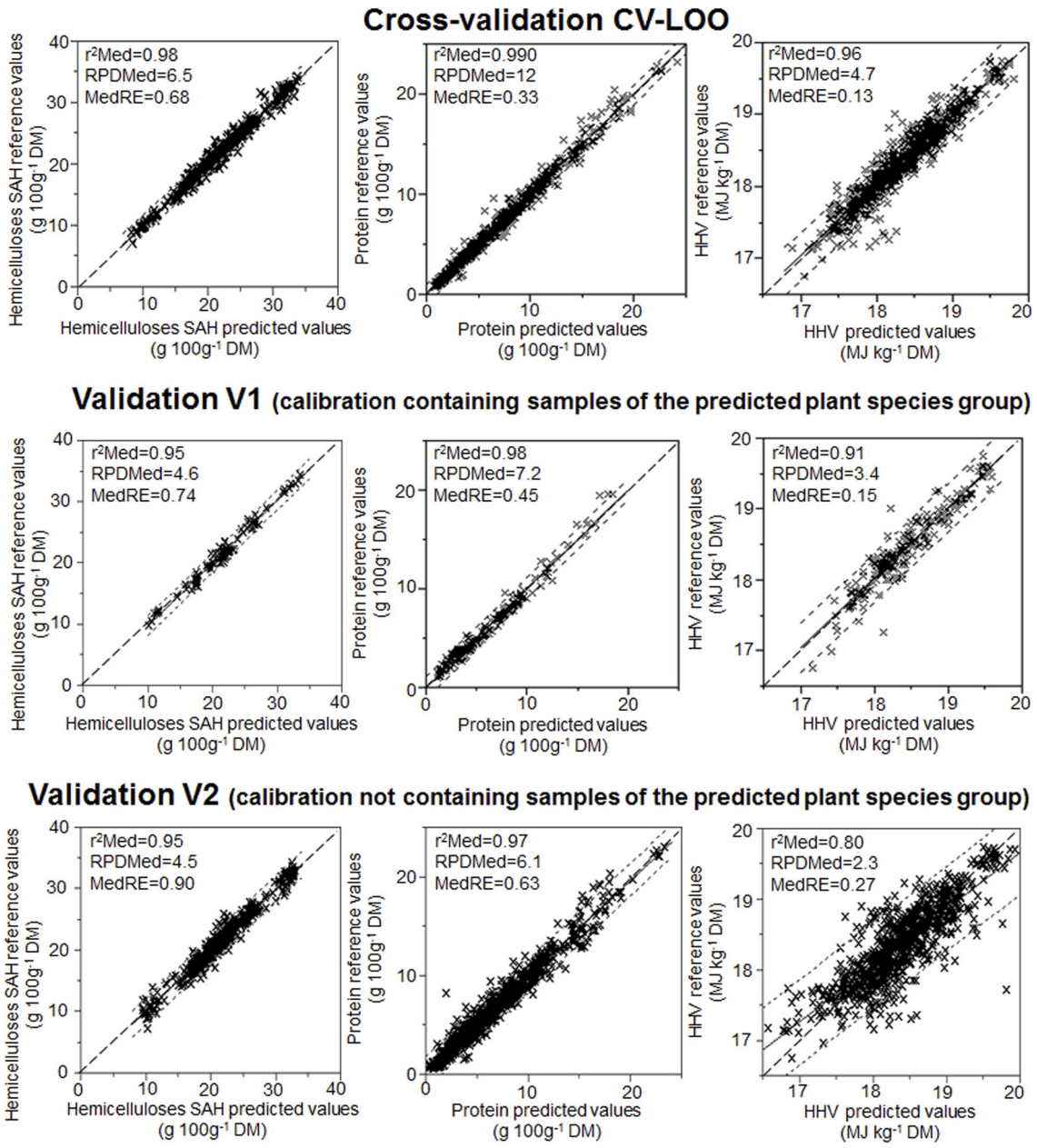

Figure 1. Reference values versus the predicted values for the cross-validation *CV-LOO*, validation V1 (calibration containing samples of the predicted plant-species group) and validation V2 (calibration not containing samples of the predicted plant-species group) for the local prediction models of hemicelluloses SAH, protein and HHV. The black dashed line in each plot is the line of equality ($y = x$). The grey plain line in each plot is the regression line. The dashed grey line in each plot is the 95% confidence line of the regression.

The analysis of the ratios of *MedRE* to *SEL* and of *MedRE* of validation (validation V1 or V2) to *MedRE* of cross-validation *CV-LOO* showed three interesting points about the accuracy and overestimation of the prediction performance (in terms of prediction error) of local and PLS models according to the type of validation.

First, the local models were more reliable than PLS models independently of the type of validation, as explained in the previous section. Furthermore, there was a decrease in the accuracy of local and PLS models according to the increase in the degree of independence of the validation set (i.e. the similarity of the predicted samples in regard to the calibration samples). This is illustrated by Figure 1, which shows that the prediction error increased according to the degree of independence of the validation. The degree of independence of the validation set was the lowest for the cross-validation *CV-LOO*, increased for validation V1 and was the highest for validation V2, as explained in the "Statistical analysis" section.

Second, the difference in accuracy between local and PLS models decreased according to the increase in the degree

of independence of the validation set. The local method was more affected by the degree of independence of the validation set than the PLS method. This can be explained by the fact that, with the local method, a specific regression was built for each sample, as explained in the previous section. An additional explanation was that the regressions of the PLS models were built with many more samples than the specific regressions of the local models. This enabled the PLS method to be less optimistic with multispecies predictions while losing some accuracy, as compared with the local method.

Third, the prediction performance of the cross-validation *CV-LOO* was too optimistic (overestimated) compared with an independent validation set such as validation V1, especially for the local method. This explained why some *MedRE* values of cross-validation *CV-LOO* of local models were very close to their respective *SEL* value (Table 3). The samples of validation V2 were too independent (different plant-species groups) in regard to the calibration samples. Therefore, the prediction performance of validation using a V2 approach should not be questioned. The prediction performance of a validation such as validation V1 should therefore be considered for future new samples of plant species contained in the library.

Thus, the degree of independence of samples in the validation set relative to samples used in the calibration set had a major impact on the results of prediction performance of multispecies models. The prediction performance of multispecies model for agricultural products should be estimated with an independent validation set such as validation V1: a small set of representative samples (independent in regard to the calibration samples; not from the same cropping site-year-harvest period) containing all types of plant-species groups that are predicted by the remaining samples set (which contains samples of all the predicted plant-species groups). This is specially the case for local models.

The prediction models of the chemical characteristics with high reliability ($r^2Med \geqslant 0.80$ and $RPDMed \geqslant 2.3$ of cross-validation *CV-LOO*, validation V1 and validation V2) can be used for quantitative purposes owing to their r^2Med and *RPDMed* (Table 2). These models were those of the main chemical characteristics of a fibrous biomass: NDF, ADF, ADL, insoluble Klason lignin, cellulose SAH, hemicelluloses SAH, xylan, total soluble sugars, protein, mineral compounds, HHV and eDOM. The high reliability of these models can be explained by the more accurate primary analytical methods because of the low relative *SEL* and high concentration level of the main chemical characteristics, as explained in the previous section. These reliable models enable the use of NIR for screening, ranking and quantitative analyses of the main chemical components contents in fibrous biomasses, and for the assessment of their suitability to be converted into biofuels. Normally, these fibrous biomasses should be the same plant species as those of the library because a prediction model is specific to the types of samples with which it is built. However, the validations V2 of these reliable models still had quantitative prediction performance ($r^2Med \geqslant 0.80$ and $RPDMed \geqslant 2.3$ of validation V2). This validation V2 involved predicting a fibrous plant species by other fibrous plant species. The prediction performance of a validation such as validation V2 should therefore be considered for future new samples of plant species not contained in the library but similar to plant species contained in the library. Nevertheless, validation V1 and validation V2 showed that the prediction models were, respectively, 2.9 and 4.5 times less accurate than the primary analytical methods based on their median value, as previously mentioned. Yet, even with such a loss of accuracy, NIR prediction models can still be used for screening and ranking purposes to detect statistical differences in the chemical composition of biomasses.[29]

The prediction models of the chemical characteristics with low reliability ($r^2Med < 0.80$ or $RPDMed < 2.3$ of cross-validation *CV-LOO*, validation V1 or validation V2) cannot be used for quantitative purposes owing to their r^2Med and *RPDMed*. However, these prediction models can be used for semi-quantitative screening and ranking of future new samples of plant species contained in the library. These models were those of the minor chemical characteristics of a fibrous biomass: arabinan, mannan, hemicellulosic glucan sum of soluble sucrose + glucose + fructose, soluble sucrose, soluble glucose, soluble fructose and BMP. The low reliability of these models was explained in the previous section. The prediction models of mannan, hemicellulosic glucan and soluble glucose were those with an even lower reliability (r^2Med and *RPDMed*), as explained in in the previous section.

The median spectral distance of Mahalanobis (*GHMed*) of the prediction models is shown for each chemical characteristic and type of validation in Table 5. The *GHMed* had a median value of 0.82 for the cross-validation *CV-LOO*, 0.95 for validation V1 and 1.83 for validation V2. The increases in *GHMed* according to the type of validation can be explained by the degree of independence of the validation set, as previously explained. NIR prediction models must have a low *GHMed* to yield reliable values. In the present study, the prediction models with a high reliability (as defined above) generally had a *GHMed* value smaller than 2, whereas those models with a low reliability (as defined above) generally had a *GHMed* value above 2.

Addition of a few samples of the predicted plant-species group into their calibration set

The samples of validation V2 had the specificity of being highly independent (different plant species) in regard to their respective calibration set. Thus, the reliability of the prediction models of the validation V2 was consistently lower than the cross-validation *CV-LOO* and validation V1, as explained in the previous section. To reduce the degree of independence of validation V2, a few independent samples (5, 10, 15, 20 and 25) of the predicted plant-species group were added into their calibration sets corresponding to validation V2.

Table 6 shows that such additions of a few samples to the calibration sets of validation V2 led to improved prediction performances (*MedRE*, $MedRE*SEL^{-1}$ and *GHMed*). The improvement can be explained by the fact that, with such additions, the calibration sets also contained plant-species samples that

Table 6. Performances of validation V2 (calibration not containing samples of the predicted plant-species group) with the addition of a few samples of the predicted plant-species group into their calibration set for the prediction models of the NDF, ADF, ADL and mineral compounds.

Chemical characteristic	Number of samples added for the predicted group	*SEL*	Local *MedRE*	PLS *MedRE*	Local *MedRE* *SEL^{-1}	PLS *MedRE* *SEL^{-1}	Local GH Med	PLS GH Med
		(g 100 g^{-1} DM)						
NDF[a]	0 (Validation V2)	0.40	1.92	1.97	4.8	4.9	1.56	1.20
	5	0.40	1.60	1.94	4.0	4.9	1.45	1.18
	10	0.40	1.49	1.92	3.7	4.8	1.39	1.16
	15	0.40	1.45	1.79	3.6	4.5	1.37	1.13
	20	0.40	1.41	1.75	3.5	4.4	1.28	1.11
	25	0.40	1.38	1.74	3.5	4.4	1.21	1.09
Validation V1		0.40	1.06	1.28	2.7	3.2	0.91	0.84
ADF[a]	0 (Validation V2)	0.30	1.58	1.73	5.3	5.8	1.39	1.19
	5	0.30	1.42	1.70	4.7	5.7	1.25	1.14
	10	0.30	1.38	1.68	4.6	5.6	1.23	1.13
	15	0.30	1.29	1.66	4.3	5.5	1.25	1.15
	20	0.30	1.22	1.63	4.1	5.4	1.16	1.13
	25	0.30	1.21	1.60	4.0	5.3	1.12	1.11
Validation V1		0.30	1.01	1.17	3.4	3.9	0.89	0.85
ADL[a]	0 (Validation V2)	0.15	0.65	0.70	4.3	4.7	1.83	1.39
	5	0.15	0.60	0.65	4.0	4.3	1.68	1.27
	10	0.15	0.52	0.63	3.5	4.2	1.54	1.26
	15	0.15	0.48	0.61	3.2	4.1	1.41	1.25
	20	0.15	0.47	0.60	3.1	4.0	1.29	1.24
	25	0.15	0.47	0.60	3.1	4.0	1.20	1.22
Validation V1		0.15	0.50	0.50	3.3	3.3	0.91	0.91
Mineral compounds	0 (Validation V2)	0.10	0.87	0.95	8.7	9.5	1.82	1.56
	5	0.10	0.79	0.89	7.9	8.9	1.71	1.48
	10	0.10	0.72	0.88	7.2	8.8	1.56	1.42
	15	0.10	0.66	0.88	6.6	8.8	1.47	1.41
	20	0.10	0.63	0.86	6.3	8.6	1.40	1.33
	25	0.10	0.60	0.85	6.0	8.5	1.35	1.28
Validation V1		0.10	0.58	0.77	5.8	7.7	0.86	0.78

[a]Residue corrected for its mineral compounds.

were the same as the plant species in the prediction set. The first of these samples additions consistently improved these performances (decrease in *MedRE*, *MedRE***SEL*$^{-1}$ and *GHMed*), whereas these performances began to stabilize with the additions of the last of these samples. Beyond the addition of more than 25 samples, the improvement in these performances would probably be minimal.

Based on the ratio of *MedRE* to *SEL* (Table 6), the mean accuracy of prediction (in terms of prediction error) of validation V2 after the addition of 25 samples was improved by 28% for local models and by 11% for PLS models. The mean *GHMed* of validation V2 after the addition of 25 samples decreased by 26% for local models and 12% for PLS models. The increase in the prediction performance was superior for local models, as compared with PLS models. This can be explained by the fact that, with the local method, a specific regression was built for each sample, as explained previously. Thus, the addition of a few samples of the same plant species in the calibration sets (not containing samples of the predicted plant-species group) had a much more important impact on the local models, as compared with the PLS models.

Therefore, it is interesting to use the local method to predict a given plant species, even if there are only a few samples of them that are present in a large multispecies dataset with

similar plant-species samples. This approach is very practical for cost-effective and fast NIR screening of plant biomasses for which the specific NIR prediction models are not yet available.

Conclusions

The developed NIR models were reliable for the prediction of different main chemical characteristics of various fibrous plant species using multispecies datasets. The types of NIR models developed in the present study can be used for screening, ranking and quantitative analyses of the main chemical components contents in fibrous biomasses, and for the assessment of their suitability for conversion into biofuels.

The local models were more reliable in terms of prediction error compared with the PLS models. The local method, by developing specific regressions for each sample, appears to cope with the non-linearity and non-homogeneity associated with a large multispecies dataset.

The degree of independence of samples in the validation set (cross-validation *CV-LOO*, validation V1, validation V2) relative to samples used in the calibration set had a major impact on the prediction performance, especially for the local method. It affected the local method more because of the lower number of samples used in its specific regressions. There was a decrease in the reliability of local and PLS models according to the increase in the degree of independence of the validation set (i.e. the similarity of the predicted samples in regard to the calibration samples). An independent validation such as V1 should be used to determine the prediction performance of NIR models for agricultural products. The validation V1 set contained independent and representative samples that did not come from the same cropping site, year or harvest period in regard to the samples of the calibration set. Owing to their degree of independence, the prediction performance of a validation such as validation V1 should be considered for future new samples of plant species contained in the library, whereas a validation such as validation V2 should be considered for future new samples of plant species not contained in the library but similar to the plant species contained in the library.

The additions of a few independent samples of the predicted plant-species group to their calibration set of validation V2 (calibration not containing samples of the predicted plant-species group) resulted in improved prediction performances of multispecies models, especially for the local method. However, these performances began to stabilize with the last sample additions (20 and 25 samples). Thus, the use of the local method is also interesting for predictions of a given plant species when there are only a few samples of them that are present in a large multispecies dataset of similar plant-species samples. This approach will enable fast cost-effective NIR screening, ranking and quantitative analyses of the chemical characteristics of new plant biomasses that are similar to those of the library.

Supplementary data

The optimum number of selected samples, minimum and maximum PLS components for the local models, and number of PLS components for the PLS models are shown in Table S1 of the supplementary information.

The number of samples of each plant-species group for each predicted chemical characteristics is shown in Table S2 of the supplementary information.

Acknowledgements

This research was funded by the Walloon Agricultural Research Center (CRA-W) with the support of the Belgian Science Policy and by the ENERBIOM project (ENERBIOM project no. 14GR23024 of the European territorial cooperate programme in the context of INTERREG IV A "Grande Région" 2007-2013 no. CCI2007CB163PO064, co-financed by the FEDER-EU funds). The authors are grateful to the technicians of the BIOETHA2 and the ENERBIOM projects for their technical support. The authors acknowledge Stéphane Lamaudière and Patrick Gerin (Université catholique de Louvain) for the soluble sucrose, glucose and fructose analyses, and Frédéric Mayer and Philippe Delfosse (Centre de Recherche Public—Gabriel Lippmann) for the BMP analyses.

References

1. A. Demirbas, "Biomass resource facilities and biomass conversion processing for fuels and chemicals", *Energ. Convers. Manage.* **42,** 1357 (2001). doi: http://dx.doi.org/10.1016/S0196-8904(00)00137-0
2. P. McKendry, "Energy production from biomass (part 1): overview of biomass", *Bioresour. Technol.* **83,** 37 (2002). doi: http://dx.doi.org/10.1016/S0960-8524(01)00118-3
3. B. Kamm and M. Kamm, "Principles of biorefineries", *Appl. Microbiol. Biotechnol.* **64,** 137 (2004). doi: http://dx.doi.org/10.1007/s00253-003-1537-7
4. B. Hames, S. Thomas, A. Sluiter, C. Roth and D. Templeton, "Rapid biomass analysis", *Appl. Biochem. Biotech.* **105,** 5 (2003). doi: http://dx.doi.org/10.1385/ABAB:105:1-3:5
5. B. Godin, S. Lamaudière, R. Agneessens, T. Schmit, J.-P. Goffart, D. Stilmants, P. Gerin and J. Delcarte, "Chemical composition and biofuel potentials of a wide diversity of plant biomasses", *Energy Fuels* **27,** 2588 (2013). doi: http://dx.doi.org/10.1021/ef3019244
6. D. Bertrand and E. Dufour, *La spectroscopie infrarouge et ses applications analytiques, 2nd Edn.* Lavoisier, Paris, France (2006).
7. P. Berzaghi, J. Shenk and M. Westerhaus, "Local prediction with near infrared multi-product databases", *J. Near Infrared Spectrosc.* **8,** 1 (2000). doi: http://dx.doi.org/10.1255/jnirs.258

8. J. Shenk, M. Westerhaus and P. Berzaghi, "Investigation of a local calibration procedure for near infrared instruments", *J. Near Infrared Spectrosc.* **5,** 223 (1997). doi: http://dx.doi.org/10.1255/jnirs.115

9. D. Pérez-Marin, A. Garrido-Varo and J. Guerrero, "Non-linear regression methods in NIRS quantitative analysis", *Talanta* **72,** 28 (2007). doi: http://dx.doi.org/10.1016/j.talanta.2006.10.036

10. H. Tran, P. Salgado, E. Tillard, P. Dardenne, X. Nguyen and P. Lecomte, "'Global' and 'Local' predictions of dairy diet nutritional quality using near infrared reflectance spectroscopy", *J. Dairy Sci.* **93(10)**, 4961 (2010). doi: http://dx.doi.org/10.3168/jds.2008-1893

11. C. Raju, A. Ward, L. Nielsen and H. Moller, "Comparison of near infra-red spectroscopy, neutral detergent fibre assay and in-vitro organic matter digestibility assay for rapid determination of the biochemical methane potential of meadow grasses", *Bioresour. Technol.* **102,** 7835 (2011). doi: http://dx.doi.org/10.1016/j.biortech.2011.05.049

12. P. Van Soest and R. Wine, "Use of detergents in the analysis of fibrous feeds. iv. determination of plant ceil. vaii constituents", *J. AOAC* **50(1)**, 50 (1967).

13. P. Van Soest, "Collaborative study of acid-detergent fiber and lignin", *J. AOAC* **56(4)**, 781 (1973).

14. B. Godin, R. Agneessens, P. Gerin and J. Delcarte, "Composition of structural carbohydrates in biomass: Precision of a liquid chromatography method using a neutral detergent extraction and a charged aerosol detector", *Talanta* **85,** 2014 (2011). doi: http://dx.doi.org/10.1016/j.talanta.2011.07.044

15. B. Godin, R. Agneessens, S. Gofflot, S. Lamaudière, G. Sinnaeve, P. Gerin and J. Delcarte, "Revue sur les méthodes de caractérisation des polysaccharides structuraux des biomasses lignocellulosiques", *Biotechnol. Agron. Soc. Environ.* **15,** 165 (2011).

16. B. Godin, R. Agneessens, P. Gerin and J. Delcarte, "Structural carbohydrates in a plant biomass: correlations between the detergent fiber and dietary fiber methods", *J. Agric. Food Chem.* **62,** 5609 (2014). doi: http://dx.doi.org/10.1021/jf500924q

17. European Union, "Commission Regulation No 152/2009", *Off. J. Eur. Union* **L54**, 1 (2009).

18. AOAC, *Official Methods of Analysis 15th Edn.* Association of Official Analytical Chemist, Washington, DC (1990).

19. European Committee for Standardization, *CEN/TS 14918: Solid Biofuels—Method for the Determination of Calorific Value*. European Committee for Standardization, Brussels, Belgium (2005).

20. J. De Boever, B. Cottyn, F. Buysse and J. Vanacker, "The use of an enzymatic technique to predict digestibility, metabolisable and net energy of compound feedstuffs for ruminants", *Anim. Feed Sci. Technol.* **14,** 203 (1986). doi: http://dx.doi.org/10.1016/0377-8401(86)90093-3

21. B. Godin, S. Lamaudière, R. Agneessens, T. Schmit, J.-P. Goffart, D. Stilmants, P. Gerin and J. Delcarte, "Chemical characteristics and biofuel potential of several vegetal biomasses grown under a wide range of environmental conditions", *Ind. Crops Prod.* **46,** 1 (2013). doi: http://dx.doi.org/10.1016/j.indcrop.2013.04.007

22. B. Godin, S. Lamaudière, R. Agneessens, T. Schmit, J.-P. Goffart, D. Stilmants, P. Gerin and J. Delcarte, "Chemical characteristics and biofuels potentials of various plant biomasses: influence of the harvesting date", *J. Sci. Food Agric.* **93,** 3216 (2013). doi: http://dx.doi.org/10.1002/jsfa.6159

23. F. Mayer, P. Gerin, A. Noo, G. Foucart, J. Flammang, S. Lemaigre, G. Sinnaeve, P. Dardenne and P. Delfosse, "Assessment of factors influencing the biomethane yield of maize silages", *Bioresour. Technol.* **153,** 260 (2014). doi: http://dx.doi.org/10.1016/j.biortech.2013.11.081

24. P. Dardenne, "Some considerations about NIR spectroscopy: Closing speech at NIR-2009", *NIR News* **21(1),** 8 (2010).

25. F. Hampel, "The influence curve and its role in robust estimation", *J. Am. Stat. Assoc.* **69,** 383 (1974). doi: http://dx.doi.org/10.1080/01621459.1974.10482962

26. D. Malley, C. McClure, P. Martin, K. Buckley and W. McCaughey, "Compositional analysis of cattle manure during composting using a field-portable near-infrared spectrometer", *Commun. Soil Sci. Plant Anal.* **36,** 455 (2005). doi: http://dx.doi.org/10.1081/CSS-200043187

27. B. Godin, F. Ghysel, R. Agneessens, T. Schmit, S. Gofflot, S. Lamaudière, G. Sinnaeve, J.-P. Goffart, P. Gerin, D. Stilmant and J. Delcarte, " Détermination de la cellulose, des hémicelluloses, de la lignine et des cendres dans diverses cultures lignocellulosiques dédiées à la production de bioéthanol de deuxième génération", *Biotechnol. Agron. Soc. Environ.* **14,** 549 (2010).

28. K. Boyer, W. Horwitz and R. Albert, "Interlaboratory variability in trace element analysis", *Anal. Chem.* **57,** 454 (1985). doi: http://dx.doi.org/10.1021/ac50001a031

29. D. Templeton, A. Sluiter, T. Hayward, B. Hames and S. Thomas "Assessing corn stover composition and sources of variability via NIRS", *Cellulose* **16,** 621 (2009). doi: http://dx.doi.org/10.1007/s10570-009-9325-x