Use of multivariate NMR analysis in the content prediction of hemicellulose, cellulose and lignin in greenhouse crop residues

Luis M. Aguilera-Sáez[a,1], Francisco M. Arrabal-Campos[a,1], Ángel J. Callejón-Ferre[b], María D. Suárez Medina[c], Ignacio Fernández[a,*]

[a] Department of Chemistry and Physics, Research Centre CIAIMBITAL, University of Almería, Ctra. Sacramento, s/n, 04120, Almería, Spain
[b] Department of Engineering– CIESOL, ceiA3, University of Almería, Ctra. Sacramento, s/n, 04120, Almería, Spain
[c] Department of Biology and Geology, University of Almería, Ctra. Sacramento, s/n, 04120, Almería, Spain

Keywords:
NMR
Biomass
Greenhouse crop residues
Predictive models
Cellulose
Hemicellulose
Lignin

A B S T R A C T

We have introduced the use of multivariate NMR analysis in the development of accurate and robust prediction models, potentially arising from a correlation between soluble metabolite profiles and cell wall composition, for the determination of hemicellulose, cellulose and lignin contents in 8 species of greenhouse crop residues. The present paper demonstrates that discriminant buckets coming from a PLS-DA model in combination with linear models provide a useful and rapid tool for the determination of cell wall composition of these plant wastes. Regularized linear regression methods have also been applied to avoid overfitting, producing improved models specifically for lignin and cellulose determinations. The predictive models are also presented in a desktop application available at http://www2.ual.es/NMRMBC/solutions. To verify the rationality and reliability of the models, control experiments following generally accepted protocols have been performed and compared to our predicted values.

1. Introduction

The potential for conversion of cellulosic, non-food-source biomass into biofuels is yet to be fully developed as a replacement for fossil fuels (Wyman, 2007). A lthough the inorganic components of biomass, especially chlorine, can cause pollution problems or deterioration of furnaces during burning, lignocellulosic biomass is recognized as one of the most important renewable resources available for conversion to fuels and other chemicals (US Department of Energy, 2011). In the last years, wood agricultural or forest residues have become an alternative bio-resource for obtaining bioethanol (Hallac et al., 2009). Thermal biomass processing via gasification or pyrolysis produces syngas and oil intermediates that are flexible feedstocks for fuel production (Chum and Overend, 2001; Ni et al., 2006).

Other uses of biomass generated from greenhouses crop residues are fertilizers, organic amendments, textile fibers, gardening and building materials, food industry (human and animal), timber industry (conglomerates, boards, etc.), pharmaceuticals and even for cosmetics (Vargas-Moreno et al., 2012). It is important to note that the production of biomass is about eight times the total annual world consumption of energy from all sources, and only a 7 percent of this annual production of biomass is reused, which indicates that we are only partially exploiting nature's abundant renewable resources.

The chemical composition of cell walls of plants varies among species, but in general, it consists of 25 percent lignin and 75 percent carbohydrates. The latter is mainly attributed to the polysaccharides cellulose and hemicellulose. Cellulose is the major carbohydrate comprising the cell wall common to all plants, which is a β-1,4-linked glucose polysaccharide. Hemicelluloses are a class of polysaccharides that have variable compositions and structures depending on the plant source. They form hydrogen bonds with cellulose, covalent bonds with lignin, and ester linkages with acetyl units and hydroxycinnamic acids. Their general formulates are $(C_5H_8O_4)_n$ and $(C_6H_{10}O_5)_n$, which are called pentosans and hexosans, respectively (Ren and Sun, 2010). The final main structural component, lignin, is a complex three-dimensional polyphenolic polymer that partially encases the plant cell-wall polysaccharides and cellulose microfibrils in lignified (i.e., secondary) plant cell walls. In addition to these three main polymers of lignocellulose, there are other non-structural components within the plant cell wall. These components, such as extractives, protein, ash, and pectin, vary greatly with species, tissue, plant maturity, harvest times, and storage, and are greatly influenced by environmental factors and stress (Davison

et al., 2013). Depending on the plant species, there is considerable variation in the relative amounts of cellulose, hemicellulose and lignin within the cell walls (Davison et al., 2013), and is for this reason that potent predictive models of biomass of unknown composition and source are required towards large-scale implementation of a biomass-to-biofuels industry (Lynd et al., 2008).

Almería (Spain) is the region all over the world with the highest density of greenhouses (CAPMA 2013), with more than 28,500 ha, largely given over to the production of tomatoes, peppers, melons, watermelons, aubergines, courgettes, cucumbers and beans. After harvest, about a million of tons per year of fresh weight crop residues are derived (Callejón-Ferre et al., 2011), which in terms of mean energy potential equals approximately a million of MW per hour and year. Greenhouse crop residues contain a mixture of highly volatile and highly nonvolatile compounds along with both low and high polarity compounds. Chromatographic techniques will only provide partial information about the composition of biomass. Several different analytical techniques, such as elemental analysis, infrared spectroscopy, gel permeation chromatography, and wet chemistry methods have been employed to characterize biomass (Lupoi et al., 2015). Established wet chemical techniques (Lupoi et al., 2014) for studying greenhouse crop residues composition do not meet the requirement of rapid and real-time detection in large-scale industrial biomass utilization, since they are time consuming, laborious and require harsh reagents (Lupoi et al., 2014). Even with a combination of all these analyses, a comprehensive view of the chemical properties of the biomass mixture is still not achieved. Thus, further research for rapid determination of components of biomass is required. Visible and near infrared (NIR) spectroscopy has been recognized as one of the most promising techniques for prediction of physical and chemical properties of mass materials, due to its pow-erful, rapid, nondestructive, simple sample preparation and good re-producibility (Xu et al., 2013). Very recently, Li et al. (2015) have developed a predictive model for the determination of hemicellulose, cellulose and lignin in Moso Bamboo based on characteristic NIR wa-velengths, obtaining R^2 values of 0.921, 0.909 and 0.892, respectively. In fact, they established the use of 20–22 independent variables in their models, which significantly reduced the number of variables employed by linear methods described until then (Huang et al., 2008).

Despite NIR has become pervasive in the literature during the last decade, nuclear magnetic resonance (NMR) spectroscopy can offer direct structural elucidation and in many cases quantification of the majority of these small molecules detected in just one analytical step without prior treatment of the sample. Limitations are seen in method sensitivity since only the most abundant metabolites are detected, and in signal overlapping when the number of metabolites is relatively high (Coen et al., 2008; Keun et al., 2002; Lenz and Wilson, 2007). In ad-dition, NMR is established to be about four times more expensive than NIR spectroscopy (see Table S4 for cost analysis). On the positive side, NMR does not require frequent instrument recalibration between ana-lyses and provides relative standardization across a variety of samples. Consequently, ^{1}H NMR spectroscopy has been used as an analytical tool for quantitative analysis of functional groups in biomass pyrolysis oils from a variety of biomasses (Mullen et al., 2009), aging reactions in bio-oil (Joseph et al., 2010), or to determine water content and relative viscosity (Dalitz et al., 2012), among many other applications related to biofuels (de Peinder et al., 2009; Filgueiras et al., 2015; Masili et al., 2012). Thus, NMR instruments and instrument time available for this type of work are both limited, but still the analytical tool that produces the highest number of publications on metabolomics and metabonomics (Theodoridis et al., 2012).

In evaluation of agricultural materials, NMR can give much more structural information than other analytical techniques; especially, for obtaining information about the molecular structure of components from one- and two-dimensional (1D and 2D) NMR spectra of several nuclei such as ^{1}H, ^{13}C, ^{19}F, ^{31}P, etc. In addition, the combination with chemometric techniques, has made possible the differentiation of, for instance, the geographical origin of grapes, the year of vintage, and even the grape variety (Du et al., 2007; López-Rituerto et al., 2012; Papotti et al., 2013; Son et al., 2008, 2009; Viggiani and Morelli, 2008).

In this work, we have used NMR spectroscopy together with multivariate analysis, for the development of linear and regularized models for predicting the content of hemicellulose, cellulose and lignin of the 8 main species of greenhouse crop residues generated in Almería (Spain).

Table 1
Mean values of percentages (% w/w) of hemicellulose, cellulose and lignin obtained by standardized analytical methods for the 8 species of crop residues analyzed.

Species	Hemicellulose	Cellulose	Lignin
Cucurbita pepo L.	13.5	15.4	9.9
Cucumis sativus L.	19.0	19.2	6.5
Solanum melongena L.	22.7	24.8	11.7
Solanum lycopersicum L.	16.8	23.3	6.6
Phaseoulus vulgaris L.	18.3	18.0	8.1
Capsicum annuum L.	20.0	21.8	10.9
Citrillus vulgaris Schrad	18.6	14.5	9.8
Cucumis melo L.	15.3	18.6	6.6

2. Results and discussion

All the results of the structural analysis obtained by standardized analytical methods for the 8 crop residues studied herein were obtained from a previous study (Callejón-Ferre et al., 2014) and are summarized in Table 1. Briefly, lignin reached their maximum and minimum values in *S. melongena* L. (11.7%) and in *C. sativus* L. (6.5%), respectively. The cellulose content was maximum in *S. melongena* L. (24.8%) and minimum in *C. vulgaris* Schrad (14.5%). *S. melongena* L. was the crop residue with the highest amount of hemicellulose (22.7%). On the other hand, the biomass with the lowest percentage of hemicellulose was *C. pepo* L. with 13.5%.

As mentioned earlier, NMR can provide direct identification and quantification of large set of compounds in just one measurement without prior treatment of the biomass sample. The ^{1}H NMR spectrum and assignments of tomato crop residues are presented in Fig. 1. Signal assignments were based on reference spectra, the analysis of 1D and 2D-NMR experiments such as ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC, ^{1}H-^{13}C HMBC, and literature (Le Gall et al., 2003; Mounet et al., 2007).

The set of 2D-NMR experiments had a key role in the confirmation of some molecular structures. For instance, HMBC spectrum allowed to assign the doublet at δ_H 1.33 ppm to threonine instead of lactic acid, and to confirm the anomeric proton signals to the corresponding sugars. The optimized NMR acquisition and processing parameters (see Experimental section) were satisfactory for these kind of biomass samples, with highly reproducible spectra, suitable baseline correction and water suppression. Examples of NMR spectra of each of the 8 biomass species analyzed are reported in Figs. S1–S7.

Although for the construction of the predictive models there is no need of any kind of assignment or knowledge regarding the metabolites contained in the samples, to demonstrate the potential of NMR in terms of structure identification, Table 2 shows some of the information deduced in terms of chemical shift and coupling constants for the 28 most abundant metabolites found in the crop residues of tomato.

To provide the best possible conditions for multivariate data analysis, the highest reproducibility should be achieved. Therefore, all samples were buffered and adjusted to pH 7.0 in order to minimize the variation in the chemical shifts produced by protonation or deprotonation phenomena. In the present study, principal component analysis (PCA) was performed on the ^{1}H NMR data for visualizing variation in large, high-throughput datasets. Two types of plots were generated from the analysis: (1) the PCA scores plot that groups similar samples based on the input data and (2) PCA loadings plot that indicates which

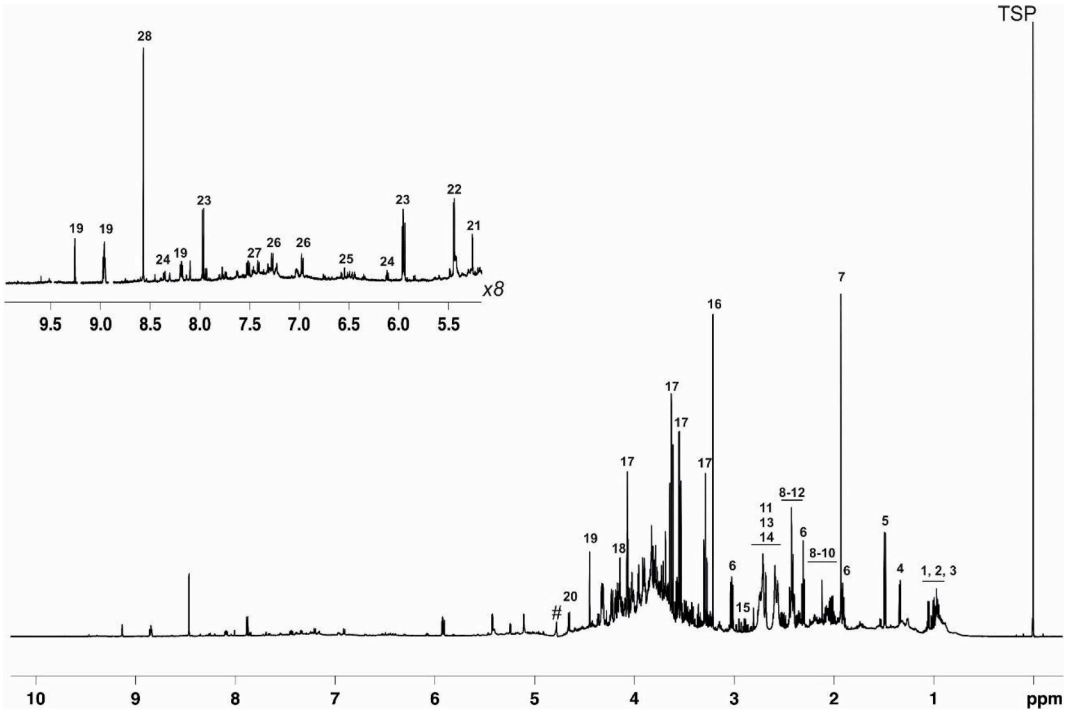

Fig. 1. ^{1}H NMR spectra (600 MHz) of a D_2O extract (pH 7.0) of tomato (*Solanum lycopersicum* L.) crop residues. Some of the identified metabolites are marked as: 1) valine; 2) isoleucine; 3) leucine; 4) threonine; 5) alanine; 6) γ-amino-butyrate (GABA); 7) acetate; 8) proline; 9) glutamate; 10) glutamine; 11) malate; 12) succinate; 13) citrate; 14) aspartate; 15) asparagine; 16) choline; 17) *myo*-inositol; 18) fructose; 19) trigonelline; 20) *β*-glucose; 21) *α*-glucose; 22) sucrose; 23) uridine; 24) adenosine; 25) fumarate; 26) tyrosine; 27) phenylalanine; 28) formate (Table 2). The suppressed water signal is marked with #.

Table 2
Summarizes the spectral information deduced for the metabolites identified on tomato (*Solanum lycopersicum* L.) crop residues.

	Metabolite	Chemical shifts (ppm) and coupling constants (Hz)
1	Valine	1.00 (d, J = 7.0 Hz), 1.05 (d, J = 7.0 Hz)
2	Isoleucine	1.02 (d, J = 7.0 Hz), 0.95 (t, J = 7.2 Hz)
3	Leucine	0.97 (t, J = 6.3 Hz)
4	Threonine	1.33 (d, J = 6.7 Hz)
5	Alanine	1.48 (d, J = 7.2 Hz)
6	GABA	1.91 (m), 2.31 (t, J = 7.4 Hz), 3.02 (t, J = 7.3 Hz)
7	Acetate	1.93 (s)
8	Proline	1.95–2.09 (m), 2.35 (m), 3.35 (m), 3.42 (m)
9	Glutamate	2.05 (m), 2.12 (m), 2.36 (m)
10	Glutamine	2.13 (m), 2.43 (m)
11	Malate	2.45 (dd, J = 15.8; 8.3 Hz), 2.70 (dd, J = 15.8; 3.9 Hz), 4.31 (dd, J = 8.3; 3.9 Hz)
12	Succinate	2.42 (s)
13	Citrate	2.55 (d, J = 16.1 Hz), 2.71 (d, J = 16.1 Hz)
14	Aspartate	2.65 (dd, J = 17.5; 9.1 Hz), 2.81 (dd, J = 17.5; 3.6 Hz)
15	Asparagine	2.87 (dd, J = 16.9; 7.7 Hz), 2.96 (dd, J = 16.9; 4.2 Hz)
16	Choline	3.21 (s)
17	*Myo*-inositol	3.28 (t, J = 9.4 Hz), 3.54 (dd, J = 9.9; 2.8 Hz), 3.62 (t, J = 9.9 Hz), 4.06 (t, J = 2.8 Hz)
18	Fructose	4.09 (m)
19	Trigonelline	4.44 (s), 8.08 (dd, J = 7.5; 6.5 Hz), 8.85 (m), 9.13 (s)
20	*β*-glucose	4.65 (d, J = 7.9 Hz)
21	*α*-glucose	5.24 (d, J = 3.7 Hz)
22	Sucrose	5.42 (d, J = 3.8 Hz)
23	Uridine	5.91 (d, J = 8.2 Hz), 5.92 (d, J = 4.7 Hz), 7.88 (d, J = 8.2 Hz)
24	Adenosine	6.08 (d, J = 5.5 Hz), 8.25 (s), 8.35 (s)
25	Fumarate	6.52 (s)
26	Tyrosine	6.90 (d, J = 8.3 Hz), 7.19 (d, J = 8.3 Hz)
27	Phenylalanine	7.33 (m), 7.38 (m), 7.43 (m)
28	Formate	8.46 (s)

spectral areas contribute more to the variation between groups. The PCA scores and loadings plots for the NMR spectra collected from 80 samples that span 8 biomass feedstock species are shown in Fig. 2 and

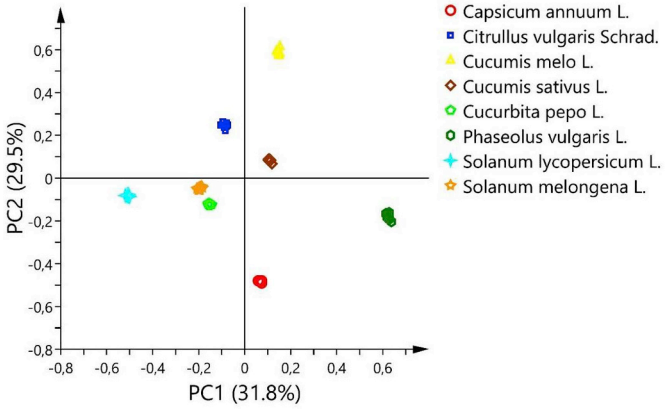

Fig. 2. PCA scores plot derived from 80 ^{1}H NMR spectra for the 8 different crop residues plant species evaluated.

Fig. S8, respectively. In overall, 80 measurements with 10 replicates for each species were included in the statistical analysis. The PCA scores plot (Fig. 2) between the two first principal components (PC1/PC2) accounted for 61.3% of the total variance of the data set. Pareto scaling was used as data preprocessing to give enough importance to the less intense peaks without overinflating them. As expected, very good discrimination was observed between all the species under study. To assess which metabolites were mostly responsible for this discrimination, the loadings plot of PCA can be inspected (Fig. S8). However, PCA scores plot does not provide any information about the proximity between species. For this reason, hierarchical cluster analysis (HCA) based on Euclidian distance coupled with the Ward's minimum variance method (Ward, 1963) was applied (Fig. 3). For this analysis, 7 principal components reduced from the original ^{1}H NMR data were used, showing statistical distances in the resulting dendrogram. These distances were calculated by the Ward linkage method, and the tree was sorted by size.

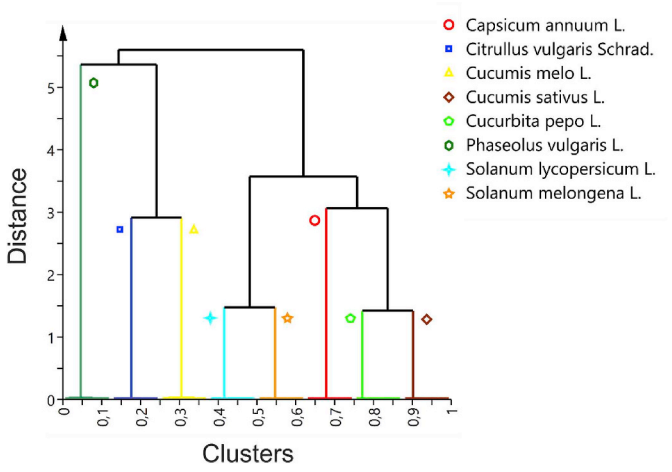

Fig. 3. Hierarchical distance cluster analysis of the first seven principal components generated by 80 ^{1}H NMR spectra. The distances between groups were calculated using Ward linkage method and the tree was sorted by size. These seven components encompass 98.9% of the total variance within the dataset.

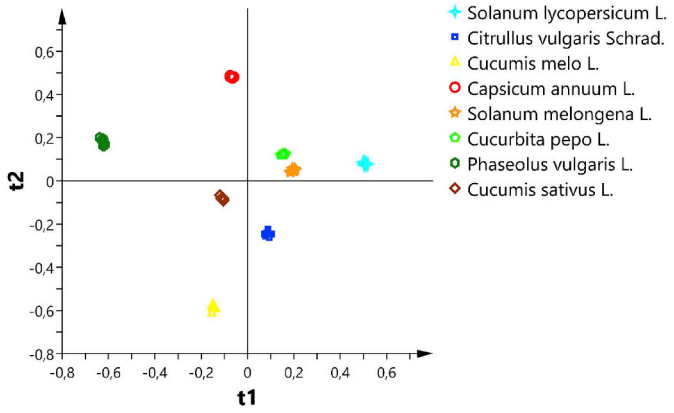

Fig. 4. PLS-DA scores plot derived from 80 ^{1}H NMR spectra for the 8 different crop residues plant species evaluated.

By analyzing the dendrogram obtained, it is possible to observe the grouped samples according to their similarities, without taking into account their class membership. The dendrogram illustrates that the samples were grouped in three clusters, i.e. melon (*C. melo* L.) and water melon (*C. vulgaris* Schrad), tomato (*S. lycopersicum* L.) and eggplant (*S. melongena* L.), and finally courgette (*C. pepo* L.), cucumber (*C. sativus* L.) and pepper (*C. annuum* L.). Regarding greenbean (*P. vulgaris* L.), the diagram shows that these samples are closer in terms of metabolic profile to melon and watermelon biomasses than to any other species.

In order to improve the discrimination among the different species and specially to select the most discriminant variables, a partial least squares discriminant analysis (PLS-DA) model was performed. This type of model is a supervised analysis that possesses high-efficiency resolving ability, because it extracts the general characteristic classification information of the full spectrum, considers class member information provided by the auxiliary matrix in code during factor configuration, adds grouping variables artificially and intensifies intergroup differences. It has widely been applied for qualitative identification of food, drug and agricultural products (Gromski et al., 2015; Pontes et al., 2017). The root mean squared error of cross validation (RMSECV) was calculated for the first ten latent variables for each plant species (Fig. S9), in order to define the optimal number of principal components necessary for the PLS-DA model (Rieppo et al., 2012). A PLS-DA model with 7 latent variables was generated based on the criteria of minimizing RMSECV with the least number of latent variables. The quality of the PLS-DA model is indicated by the cross-validation parameters, R2 and Q2, representing the explained variance and the predictive capability of the model, respectively. R2X and R2Y represent the fraction of variance of the X and Y matrix, respectively, and Q2Y represents the predictive accuracy of the model, with cumulative (cum) values of R2X, R2Y and Q2 equating to 0.990, 0.997 and 0.997 indicating an effective model. PLS-DA model was validated applying a permutation test for each class (Fig. S10). As expected, the PLS-DA scores plot (Fig. 4) of the first two latent variables showed a slight increase of clustering between species when compared to the PCA scores plot (Fig. 2). In addition, while PCA model with 7 latent variables explained 98.9% of the total variance, PLS-DA model explained 99.7% with the same number of latent variables. The variable importance in projection (VIP) scores estimate the contribution of the individual variables on the PLS-DA model. It is considered that variables with VIP values less than 1 do not influence significantly to the supervised model. The discriminant buckets from PLS-DA model with VIP values higher than 1 are reported in Fig. S11, in which it is possible to observe that the most important variables for PLS-DA model were found in the region below 5.5 ppm of the ^{1}H NMR spectra. The spectral areas showing the five largest VIP coefficients, thus contributing more significantly to the discrimination, contain the characteristic resonances of malic acid, citric acid and GABA (Table 2). In this way, the calculation of VIP scores from the PLS-DA model allowed for the selection of 59 discriminant buckets.

These results indicate that, by using ^{1}H NMR spectra coupled to PLS-DA, all the crop residues can be rapidly differentiated between each other, and therefore we envisage that these 59 discriminant buckets will provide the best predictive models.

To predict composition from the NMR spectra, 59 mathematical multivariate models based on linear combinations were formulated, which included the 59 discriminant buckets with VIP coefficients higher than 1 found previously. The set of equations were progressively reduced to one equation (one for each of the desired components: hemicellulose, cellulose and lignin) by attending to their standardized coefficients (also called beta coefficients), which are used to compare the relative weights of the used variables. When the confidence interval around standardized coefficients of a specific variable has value zero, the weight of this variable in the model is not significant and therefore is rejected. In the case of hemicellulose, from the whole set of starting variables, only 18 discriminants buckets out of the 59 found in the PLS-DA contributed with beta coefficients between −30.546 and 16.147, which afforded a final equation with an adjusted R^2 coefficient of 0.636. The same protocol of rejecting variables with beta coefficients close to zero was applied for the prediction of cellulose and lignin. For these compounds, only 14 and 8 variables were needed to obtain equations with adjusted R^2 coefficients of 0.937 and 0.906, respectively. Figs. S12–14 show the significance of the variables selected for the prediction of the content of hemicellulose, cellulose and lignin by the standardized beta coefficient.

The performances of the models were assessed by calculation of the mean squared error (MSE), the root mean squared error (RMSE) and the mean absolute percentage error (MAPE). Together with these, we provided the Akaike information criterion (AIC), which offers a relative estimation of the information lost when the model is applied, and the Schwarz criterion (SBC), which takes into account the statistical goodness of fit and the number of parameters that have to be estimated. Table 3 shows the list of correlation coefficients and errors for each of the predicted components together with the number of discriminant buckets employed for each equation. Interestingly, cellulose and lignin represent the components with lower absolute values of AIC and SBC parameters, indicating that both predictive equations minimize the information lost during the prediction, which is correlated with the highest adjusted coefficient of determination (adjusted R^2). The predictive equations are summarized in Table 4. The predictive models

Table 3
Linear regression evaluation parameters for hemicellulose, cellulose and lignin predictive models.

	Nº	R^2	Adj. R^2	MSE	RMSE	MAPE	AIC	SBC
Hemicellulose	18	0.636	0.528	3.592	1.895	7.744	142.296	185.554
Cellulose	14	0.937	0.923	0.802	0.896	3.720	14.384	48.115
Lignin	8	0.906	0.896	0.364	0.603	5.734	−60.941	−41.503

Table 4
Predictive linear regression equations for hemicellulose, cellulose and lignin.

$$Y_{hemicellulose} = -229{,}05 + 1178{,}20 * F - 19635{,}49 * I + 5720{,}91 * U + 5935{,}34 * G' - 924{,}851 * V' - 13391{,}95 * W' + 49651{,}35 * X' - 66848{,}34 * Y' + 35461{,}98 * Z' + 29914{,}26 * A'' - 426{,}65 * B'' + 1889{,}89 * C'' - 3987{,}93 * D'' - 11087{,}90 * E''$$

$$Y_{cellulose} = -54{,}90 + 1835{,}99 * S + 188{,}31 * N' - 206{,}58 * O' - 334{,}34 * P' + 3067{,}67 * R' + 481{,}38 * V' + 2160{,}55 * W' - 13990{,}30 * X' + 26541{,}92 * Y' - 25534{,}76 * Z' + 4787{,}12 * B'' - 986{,}41 * C'' - 1314{,}31 * D'' + 3416{,}79 * E''$$

$$Y_{lignin} = 18{,}48 - 1025{,}57 * F' - 161{,}26 * J' - 152{,}42 * K' + 2497{,}36 * Q' - 2817{,}72 * Y' + 6152{,}03 * Z' - 3640{,}10 * A'' - 2330{,}44 * E''$$

were fitted to a training dataset that contains 60 samples and evaluated using a test dataset formed by 20 samples, which were always independent of the samples from the training set. It should be pointed out that test samples never included samples from the training set. The letters in the equations represent the discriminant buckets selected by PLS-DA analysis, and their assignments are listed in Table 5. Detailed identification of these and much more metabolites, including chemical shifts and multiplicities, can be found in Table S1. Interestingly, only citrulline and leucine contain signals that are present in the buckets employed in the three predictive equations, and only succinic acid was exclusively found in the predictive equation of cellulose. As expected, sucrose signals are contained in the buckets used for cellulose and hemicellulose predictive models since the substrate for cellulose synthesis UDP-glucose is formed by catabolism of sucrose via sucrose synthase (Fujii et al., 2010) and for example, in the case of mannans, sucrose is involved in the nucleotide sugar conversion to GDP-mannose and UDP-galactose (Pauly et al., 2013). Moreover, the tricarboxylic acid (TCA) cycle may explain the correlation found between the discriminant amino acids and organic acids detailed in Table 5 and the cell wall composition. It is well-known that a number of reactions involved in cellulose biosynthesis require ATP consumption, for instance, large amounts of A TP are consumed in the formation of matrix polysaccharides and their transport towards the cell wall (Tarchevsky and Marchenko, 1991). In this regard, TCA cycle occupies a central position in metabolism and meets most of cell energy requirement by the complete oxidation of acetyl-CoA, a key product in the catabolism of carbohydrates, fatty acids and amino acids (Desideri et al., 2015). Therefore, the energy demand for cellulose biosynthesis produces changes in amino acids and organic acids content since amino acids are considered substrate for the TCA cycle whereas malic acid, citric acid and succinic acid are intermediate compounds that are found in the TCA cycle. The same can be extrapolated to hemicellulose and lignin biosynthesis. The rest of metabolites encountered are present in two or three predictive equations as it is indicated in Table 5. Whole cell wall information in terms of lignin subunit composition and lignin interunit linkage distribution, can be found in Mansfield et al. (2012) report.

Table 5
Metabolites associated with the discriminant buckets and the predicted component hemicellulose (H), cellulose (C) and lignin (L).

Component	Bucket	Metabolite
H/C	V′,W′,X′,Y′,Z′,A″/V′,W′,X′,Y′,Z′	Citrulline
L	F′,Y′,Z′,A″	
H,C/L	E''/E′	Leucine
H/C/L	F/O′,P'/J′,K′	Malic acid
H/C	I,U/S	Sucrose
H,C	B″	Alanine
H,C	C″	Threonine
H,C	C″	2-hydroxyisobutyric acid
C/L	N'/J′,K′	Citric acid
H/C	R′,S′,T′,U'/R′	Glutamic acid
H/L	G'/Q′	γ-aminobutyric acid
C	O'P'	Succinic acid

Fig. 5 provides the prediction values (X axis) of all the models with respect to the values found experimentally (Y axis), as well as the error limits. In Fig. 5 the dashed lines represent the curves fitted by the linear regression models whereas solid lines denote the confidence interval. The active values for formulating the model appear as blue circles and the independent values for the evaluation of the model are identified by red triangles. All the graphs presented good correlations between the observed and predicted data.

K-fold c ross-validation tests were a pplied i n order to e valuate the predictive v alidity of o ur l inear regression e quations. F or this purpose, the whole dataset was randomly p artitioned i nto two sets o f 60 a nd 2 0 samples. The f ormer set i s e mployed to train the model and the l atter to v alidate i t v ia the root mean squared e rror o f prediction (RMSEP) o ut of these 2 0 data. We p erformed this test o ne hundred times a nd the c al-culated RMSEP o f e ach o f the l inear expressions o f e ach o f the c om-ponents under study a re shown i n Fig. 6 (black dots). The RMSEP v alues a re highly dispersed that g o from 2.9 to 6.0 i n hemicellulose, f rom 3.0 to 7.0 i n c ellulose, a nd from 2.0 to 3.5 i n the p rediction o f l ignin (Table S 2). I n order to avoid this b road dispersion o f the RMSEP v alues, a nd the p otential over-fitting, a n a lternative method b ased o n a r egularized l inear regression was a pplied. I t i ntroduces a c ost f unction that tries to push the c oefficients f or m any v ariables to z ero b y m eans o f a regularized term, which c ontains the l ambda p arameter, λ. (Evgeniou e t al., 2 000). To f ind these n ew c oefficients, we h ave used the g radient des-cent method, which i s a n i terative method that uses the derivative o f the c ost f unction a nd then tries to c onverge such that the c ost f unction i s minimized a lways selecting the direction with the most p ronounced i ncrement. As a n e xample, Fig. 7 shows the comparison o f b oth l inear a nd regularized methods i n the hemicellulose p ercentage a long the 8 species under study. F igs. S 15 and S 16 show the rest o f the p redicted c omponents. The λ p arameter was set i n a ll the c ases to 10^{-6}. I t i s c learly o bserved h ow the regularized relationships (red l ine i n Fig. 7) a void o ver-fitting since the red l ine p roceeds a lmost straight a long the ten samples p er species o f the training set. I n order to e valuate the regularized models, we p erform the k -fold c ross-validation o nce a gain. F ig. 6 shows i n triangles the sharper dispersion o f RMSEP v alues (Table S3) f or e ach o f the p redicted c omponents, p ointing o ut that the regularized e xpressions h ave a s ignificantly h igher p redictive a ccuracy.

The new regularized predictive equations including the new coefficients are summarized in Table 6. The new adjusted R^2 values for the prediction of hemicellulose, cellulose and lignin are 0.692, 0.940 and 0.908, respectively. Interestingly the prediction of hemicellulose significantly improved their fitting parameters, whereas the other two remain almost unchanged but as has been shown previously, with no

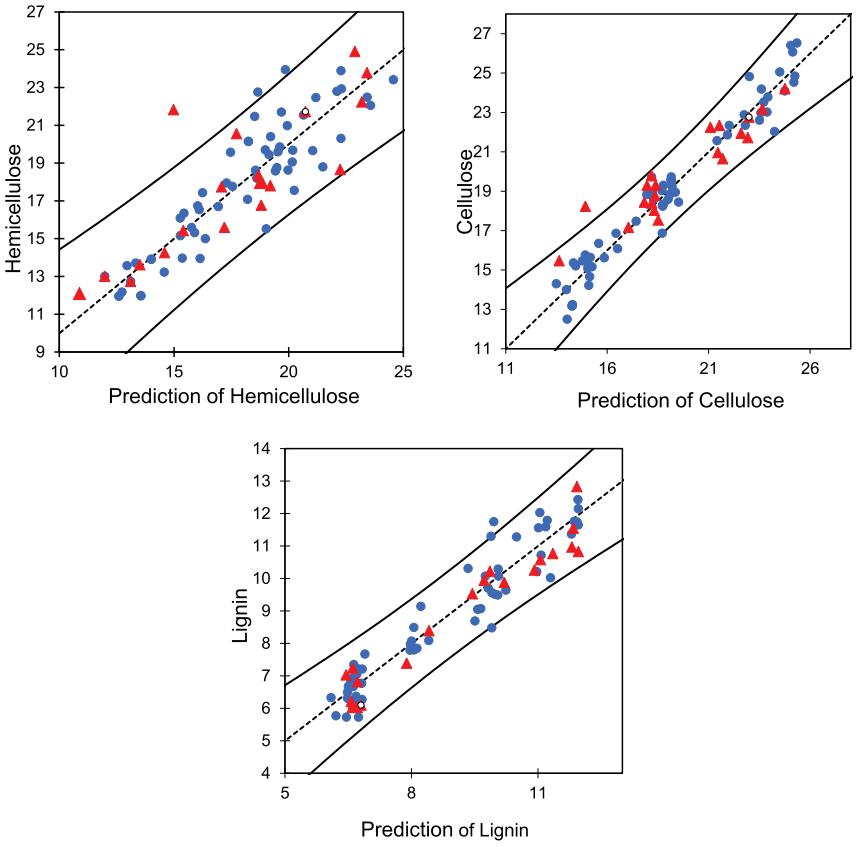

Fig. 5. Scatter graphs for the three regularized models showing their reliability as predictors of crop residues composition. A ctive values are represented in blue circles whereas testing values are shown in red triangles. Both solid lines are the upper and lower 95% prediction limits. Linear equations and their evaluation parameters are provided in Tables 3 and 6 for non-regularized and regularized linear models, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

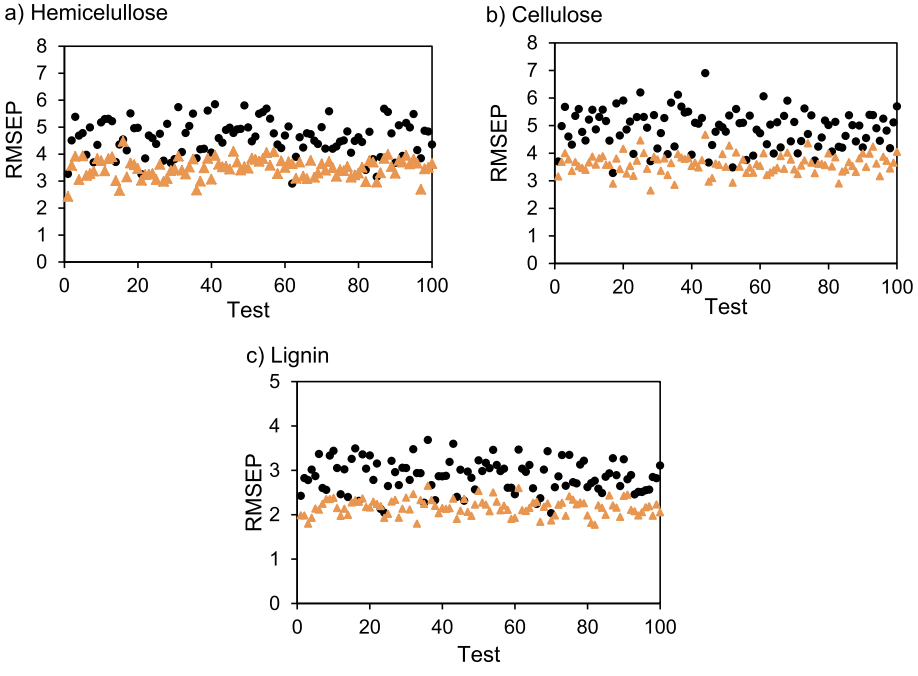

Fig. 6. Scatter root mean squared error of prediction (RMSEP) using k-fold cross-validation test for the prediction of hemicellulose, cellulose and lignin contents. Dots (black) and triangles (orange) correspond to linear and regularized models, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

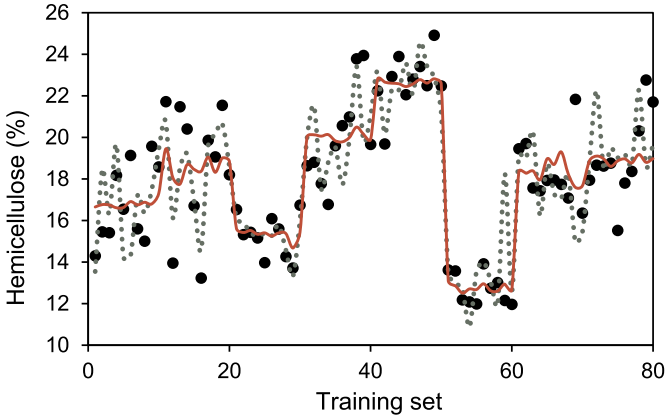

Fig. 7. Hemicellulose (%) *vs* training set. The value of λ was set to 10^{-6}. The red and green (dashed) lines correspond to the regularized and linear regressions, respectively. The black dots represent the experimental data. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and regularized prediction values together with the committed errors, are all given in Fig. 8. As expected, the prediction of hemicellulose present higher error values (up to 22%), while for the other two components the errors are in all the cases below 1%.

3. Conclusions

We have described the use of multivariate NMR analysis in the development of accurate and robust prediction models, originated from a correlation between NMR fingerprints of soluble extracts and cell wall composition, for the determination of hemicellulose, cellulose and lignin on 8 species of greenhouses-derived biomass. These results demonstrate that ^{1}H NMR discriminant buckets coming from PLS-DA in combination with linear models provide a useful and rapid tool for the determination of cell-wall biomass composition of greenhouse crop residues. Some of these specific spectral regions belong to malic acid, sucrose, GABA, glutamic acid, citrulline, alanine, threonine, leucine, citric acid, 2-hydroxyisobutyric acid and succinic acid. This straightforward procedure uses just a ^{1}H NMR spectrum and avoids the use of tedious and time-consuming chemical methods. We have also shown that regularized linear regression produces a sharper dispersion of errors in the content predictions and minimize overfitting, which provides improved models that are available in a desktop application. This study provides relevant information for unraveling greenhouse crop residues composition and therefore tackles important issues regarding its adequate management.

We believe these models proved to be the ones most applicable for bioenergy industries, with highest fitting parameters described so far and therefore the most applicable to real biomass management.

Table 6
Regularized linear regression evaluation parameters for hemicellulose, cellulose and lignin predictive models.

	Nº	R^2	Adj. R^2	MSE	RMSE	MAPE	AIC	SBC
Hemicellulose	18	0.692	0.601	2.628	1.621	6.790	177.300	160.559
Cellulose	14	0.940	0.927	0.691	0.831	3.399	2.388	36.119
Lignin	8	0.908	0.896	0.364	0.603	5.734	−60.941	−41.503

evidences of over-fitting. Again, among the three predicted components, cellulose and lignin represent those with lower absolute values of AIC and SBC parameters, indicating that both predictive models minimize information lost better than the rest. This statement agrees with highest adjusted R^2 values of 0.940 and 0.908, found for both of them.

The predictive equations are summarized in Table 7 and they are presented in a desktop application available at http://www2.ual.es/NMRMBC/solutions.

To verify the rationality and reliability of these models, control experiments were conducted with biomasses mixtures whose composition have been validated according to the accepted US Department of Energy-National Renewable Energy Laboratory (NREL) methods. The mixtures were based on the 8 evaluated biomasses, where eggplant and melon represented a 40% of the mixture and the rest of plant species in a 10% each. The ratios between eggplant and melon in the 5 mixtures analyzed varied between 1:1 to 1:2.3, respectively. The experimental

Table 7
Predictive regularized linear regression equations for hemicellulose, cellulose and lignin.

$$Y_{hemicellulose} = -4.34 + 126.03 * F - 1654.51 * I + 862.42 * U + 149.66 * G' + 14.43 * R' + 658.26 * S' + 315.54 * T' + 732.39 * U' - 69.92 * V' - 363.42 * W' + 1292.45 * X' - 200.26 * Y' - 503.04 * Z' - 650.10 * A'' - 912.59 * B'' - 624.16 * C'' + 720.64 * D'' - 1186.46 * E''$$

$$Y_{cellulose} = 36.36 - 344.30 * S - 31.81 * N' - 37.10 * O' - 0.59 * P' + 218.82 * R' - 112.11 * V' + 483.29 * W' - 186.68 * X' - 368.74 * Y' - 316.46 * Z' - 30.84 * B'' - 264.37 * C'' - 545.96 * D'' + 124.14 * E''$$

$$Y_{lignin} = 17.70 - 1202.47 * F' - 160.93 * J' - 142.50 * K' + 2323.08 * Q' - 2495.51 * Y' + 6143.54 * Z' - 3480.20 * A'' - 2230.56 * E''$$

4. Experimental

4.1. Chemicals

All chemical reagents were of analytical grade. D_2O (99.9% D) and TSP (98.0% D) were purchased from Eurisotop (Saint-Aubin, France) whereas the enzyme inhibitor sodium azide (NaN_3) and the monopotassium phosphate (KH_2PO_4) were purchased from Sigma Aldrich (Steinheim, Germany).

4.2. Greenhouse crop residue sampling

The plant species studied were courgette (*Cucurbita pepo* L.), cucumber (*Cucumis sativus* L.), eggplant (*Solanum melongena* L.), tomato (*Solanum lycopersicum* L.), greenbean (*Phaseoulus vulgaris* L.), pepper (*Capsicum annuum* L.), watermelon (*Citrillus vulgaris* Schrad), and melon (*Cucumis melo* L.). The experimental design used to produce the samples analyzed was the same for each species, where two adult plants (at the end of their life cycle) were randomly collected on different areas of the southeast of Almería (Spain), and from this pool of plants multiple samples were collected and taken to the laboratory for analysis within 24 h.

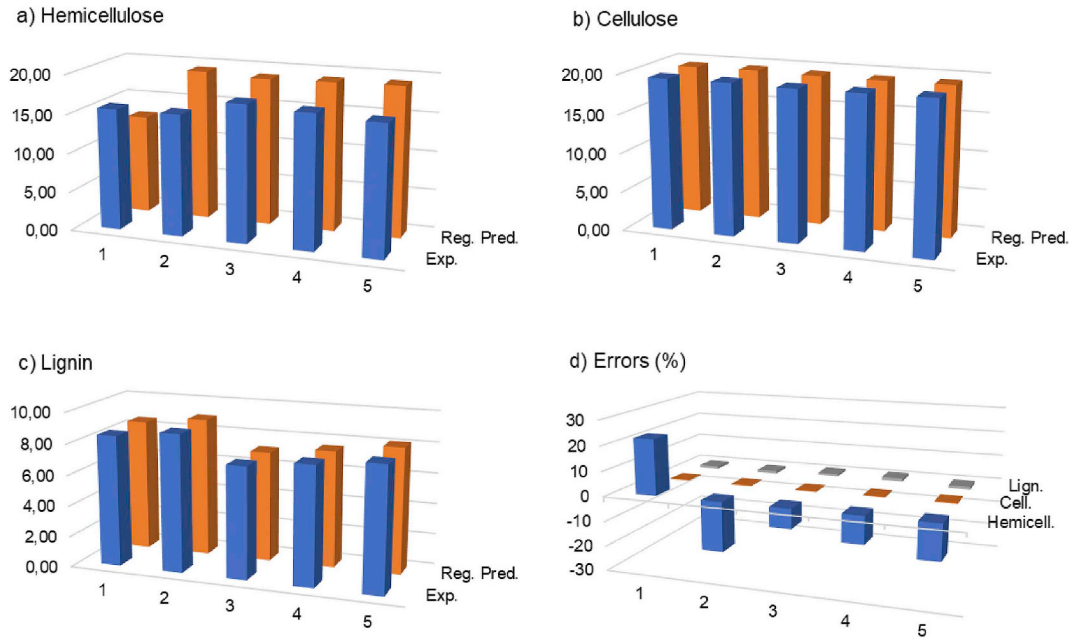

Fig. 8. Experimental values for hemicellulose, cellulose and lignin compared to those predicted with equations given in Table 7. Errors committed are given in d).

4.3. Cell wall component analysis

The roots were separated, and all foreign elements were removed, analyzing only the aerial part. The analyses were performed according to UNE-CEN/TS 14780:2008 EX (AENOR, 2008) and ASTM D1107-84(ASTM, 1984) protocols. The parameters studied were hemicellulose (H), cellulose (C) and lignin (L). All these data have been previously reported (Callejón-Ferre et al., 2014), and we will use them here with no further modifications. To verify the models, control experiments following protocols derived from the National Renewable Energy La-boratory (NREL) were performed (technical report NREL/TP-510-42619 (Sluiter et al., 2005) and technical report NREL/TP-510-42618)(Sluiter et al., 2008), and compared to our predicted values.

4.4. NMR sample preparation

Fifty milligrams of freeze-dried sample of each biomass were extracted in 1.5 mL Eppendorf tubes with 0.85 mL of KH_2PO_4 buffer (pH 7) in D_2O containing the sodium salt of 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid (TSP, 0.01%$_{ww}$) and sodium azide (NaN_3, 90 μM). The extracts were vigorously vortex-stirred for 20 min and centrifuged at 13,500 rpm for 10 min. Five hundred microliters of the supernatants were transferred in oven-dried 5 mm NMR tubes for spectral analysis.

4.5. NMR analysis and statistics

All ^{1}H-NMR spectra were recorded on a Bruker Avance III 600 spectrometer operating at a proton frequency of 600 MHz using a 5 mm QCI quadruple resonance pulse field gradient cryoprobe and equipped with a SampleCase that allowed the automatic analysis of 24 samples in a row. All samples were measured at 293 ± 0.1 K, without rotation and using 8 dummy scans prior to 80 scans. Acquisition parameters have been set as follows: size of fid = 64 K, spectral width = 20.5 ppm, acquisition time = 2.73 s, relaxation delay = 5 s, FID resolution = 0.36 Hz. Data acquisition was achieved using an experiment with a NOESY presaturation pulse sequence (Bruker 1D noesygppr1d) with water suppression via irradiation of the water frequency during the recycle and mixing time delays. The spectra were automatically phased, baseline-corrected, and calibrated to the TSP signal at 0.0 ppm. The t_1 time was set to 4 μs and the mixing time (d8) to 10 ms. Acquisition and processing of spectra were carried out with TOPSPIN software (version 3.1; Bruker Biospin GmbH, Germany). The spectrometer transmitter was locked to D_2O frequency using a mixture H_2O-D_2O (9:1). The NMR experiments employed in the statistics were carried out with a fixed receiver gain (RG) of 57, which was estimated adequate through several tests. ^{1}H–^{1}H total correlation spectroscopy (TOCSY), ^{1}H-^{13}C heteronuclear single quantum coherence (HSQC), ^{1}H-^{13}C heteronuclear multiple bonds coherence (HMBC) spectra were recorded using standard Bruker sequences. The TOCSY spectra were generated applying a relaxation delay of 2.0 s, spectral width in both dimensions of 7194.25 Hz and a RG of 64.0. It was processed using sine-bell window function (SSB = 2.0). The HSQC spectra were acquired using a relaxation delay of 1.0 s, spectral width of 7211.54 Hz in F2 and 24900.71 Hz in F1. In this case, a quadratic sine window function (SSB = 2.0) was applied. The HMBC spectra were recorded with the same parameters used in the HSQC spectra except for 37729.71 Hz of spectral width in F1. The coupling constant for HSQC experiment was fixed to 145 Hz whereas HMBC experiment was obtained using fixed coupling constants of 145 and 8 Hz (long range).

For the statistics, the ^{1}H NMR spectral data (from δ_H 0.5–10.5 ppm) were reduced into 0.04 ppm spectral buckets (the value of each bucket represents the total area within the respective spectral region) using Amix software (version 3.9.4; Bruker Biospin GmbH, Germany). The spectral region corresponding to water (from δ_H 4.74–4.82 ppm) was removed. The spectra were then normalized to total spectral area (Craig et al., 2006) and imported into SIMCA -P software (version 14.0; Umetrics, Sweden) for multivariate statistical analysis. Normalization step was carried out on the data matrix in order to correct vertical scale errors originated from the different water content in the samples.

Subsequent multivariate data analysis was performed on the assumption of normally distributed data and the mean center was applied for all multivariate analysis. Prior to application of the PCA and PLS-DA models on the NMR data matrix, it was pretreated to put the spectra in the most suitable form for the successive data analysis. Pareto scaling was chosen over other pretreatment methods (as autoscaling or variable stability (VAST) scaling) as it allows to upweight the contribution of lower intensity peaks without overinflating excessively the noise (Ritota et al., 2010).

For the predictive linear models, a total of 59 variables, selected thanks to PLS-DA model, were originally used and then reduced down to 18 for hemicellulose, 14 for cellulose and 8 for lignin, by attending to their standardized beta coefficients. The evaluation and validation of the models involved the coefficient of determination (R^2), adjusted R^2, MSE, RMSE, RMSEP, MAPE, AIC and SBC (Dempster, 1969).

All predictive equations employed 60 (75%) data as the training set and 20 (25%) random data for the validation set. The RMSEP was determined for validation purposes. While the RMSE measures the error between the estimator and the true value, the RMSEP calculates the error between what the predictive linear regression predicts for a defined value and true value. For this reason, RMSEP allows to assess the quality of the predictive linear regression. Once the random data for validation set were selected, the predictive linear regression was applied to the random data set. The RMSEP obtained using the validation data set was compared with the RMSE obtained by training data set. As expected, in all cases both error values were similar. It proves the great prediction power of the predictive linear regression models herein developed.

The linear regression models were elaborated using the statistical software XLSTAT 2009, which allowed the calculation of the significance of the variables of the mathematical prediction models by the use of the beta coefficients and Student's t-test. Beta coefficients provide the relative contribution of each bucket to the final estimated value whereas the Student's t-test provides a corrector factor with the aim to approximate the collected samples to the total population. Thus, the relative significance values for hemicellulose, cellulose and lignin were lower to 0.05. The regression models were validated with independent samples that were not used to create the models. The data observed in the new experiments and predicted by the models were compared with a paired-sample test based on Student's t-test. The validation test takes the differences between the observed and predicted values in independent samples and assesses whether the mean is statistically different from zero. The results obtained from the paired Student's *t*-test indicate that there are not statistical differences between the true values and the predictive values calculated through the linear regression models since the probability of the null hypothesis was higher than 0.05 in all cases. Therefore, this result agrees with the RMSEP and RMSE values previously introduced.

Conflicts of interest

The authors declare no conflicts of interest associated with the current study.

Acknowledgments

Financial support was given by Bruker Española SA and by Junta de Andalucía(Spain) and Ministerio de Ciencia, Innovación y Universidades (Spain) under the project numbers P12-FQM-2668 and CTQ2017-84334-R, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phytochem.2018.11.013.

References

AENOR, 2008. Biocombustibles sólidos. Métodos para la preparación de muestras (Madrid, Spain). .

ASTM, 1984. Standard Test Method for Alcohol–benzene Solubility of Wood (West Conshohocken, USA). .

Callejón-Ferre, A.J., Carreño-Sánchez, J., Suárez-Medina, F.J., Pérez-Alonso, J., Velázquez-Martí, B., 2014. Prediction models for higher heating value based on the structural analysis of the biomass of plant remains from the greenhouses of Almería (Spain). Fuel 116, 377–387.

Callejón-Ferre, A.J., Velázquez-Martí, B., López-Martínez, J.A., Manzano-Agugliaro, F., 2011. Greenhouse crop residues: energy potential and models for the prediction of their higher heating value. Renew. Sustain. Energy Rev. 15, 948–955.

CAPMA, 2013. Cartografía de invernaderos en el litoral de Andalucía oriental. Campaña 2012. Consejería de Agricultura Pesca y Medio Ambiente, Junta de Andalucía, Andalucía (Spain), pp. 21.

Chum, H.L., Overend, R.P., 2001. Biomass and renewable fuels. Fuel Sci. Technol. 71, 187–195.

Coen, M., Holmes, E., Lindon, J.C., Nicholson, J.K., 2008. NMR-Based metabolic profiling and metabonomic approaches to problems in molecular toxicology. Chem. Res. Toxicol. 21, 9–27.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K., Lindon, J.C., 2006. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Anal. Chem. 78, 2262–2267.

Dalitz, F., Steiwand, A., Raffelt, K., Nirschl, H., Guthausen, G., 2012. ^{1}H NMR techniques for characterization of water content and viscosity of fast pyrolysis oils. Energy Fuels 26, 5274–5280.

Davison, B.H., Parks, J., Davis, M.F., Donohoe, B.S., 2013. Plant cell walls: basics of structure, chemistry, accessibility and the influence on conversion. In: Wyman, C.E. (Ed.), Aqueous Pretreatment of Plant Biomass for Biological and Chemical Conversion to Fuels and Chemicals. John Wiley & Sons, Ltd.

de Peinder, P., Visser, T., Petrauskas, D.D., Salvatori, F., Soulimani, F., Weckhuysen, B.M., 2009. Partial least squares modeling of combined infrared, ^{1}H NMR and ^{13}C NMR spectra to predict long residue properties of crude oils. Vib. Spectrosc. 51, 205–212.

Dempster, A.P., 1969. Elements of Continuous Multivariate Analysis, first ed. Addison Wesley Longman Publishing Co.

Desideri, E., Vegliante, R., Ciriolo, M.R., 2015. Mitochondrial dysfunctions in cancer: genetic defects and oncogenic signaling impinging on TCA cycle activity. Cancer Lett. 356, 217–223.

Du, Y.Y., Bai, G.Y., Zhang, X., Liu, M.L., 2007. Classification of wines based on combination of ^{1}H NMR spectroscopy and Principal Component Analysis. Chin. J. Chem. 25, 930–936.

Evgeniou, T., Pontil, M., Poggio, T., 2000. Regularization networks and support vector machines. Adv. Comput. Math. 13, 1.

Filgueiras, P.R., Terra, L.A., Castro, E.V.R., Oliveira, L.M.S.L., Dias, J.C.M., Poppi, R.J., 2015. Prediction of the distillation temperatures of crude oils using ^{1}H NMR and support vector regression with estimated confidence intervals. Talanta 142, 197–205.

Fujii, S., Hayashi, T., Mizuno, K., 2010. Sucrose synthase is an integral component of the cellulose synthesis machinery. Plant Cell Physiol. 51, 294–301.

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R., 2015. A tutorial review: metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Anal. Chim. Acta 879, 10–23.

Hallac, B.B., Sannigrahi, P., Pu, Y., Ray, M., Murphy, R.J., Ragauskas, A.J., 2009. Biomass characterization of *Buddleja davidii*: a potential feedstock for biofuel production. J. Agric. Food Chem. 57, 1275–1281.

Huang, A., Li, G., Fu, F., Fei, B., 2008. Use of Visible and Near Infrared Spectroscopy to predict klason lignin content of bamboo, Chinese fir, paulownia, and poplar. J. Wood Chem. Technol. 28, 194–206.

Joseph, J., Baker, C., Mukkamala, S., Beis, S.H., Wheeler, M.C., DeSisto, W.J., Jensen, B.L., Frederick, B.G., 2010. Chemical shifts and lifetimes for Nuclear Magnetic Resonance (NMR) analysis of biofuels. Energy Fuels 24, 5153–5162.

Keun, H.C., Beckonert, O., Griffin, J.L., Richter, C., Moskau, D., Lindon, J.C., Nicholson, J.K., 2002. Cryogenic probe ^{13}C NMR spectroscopy of urine for metabonomic studies. Anal. Chem. 74, 4588–4593.

Le Gall, G., Colquhoun, I.J., Davis, A.L., Collins, G.J., Verhoeyen, M.E., 2003. Metabolite profiling of tomato (*Lycopersicon esculentum*) using ^{1}H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. J. Agric. Food Chem. 51, 2447–2456.

Lenz, E.M., Wilson, I.D., 2007. Analytical strategies in metabonomics. J. Proteome Res. 6, 443–458.

Li, X., Sun, C., Zhou, B., He, Y., 2015. Determination of hemicellulose, cellulose and lignin in moso bamboo by Near Infrared Spectroscopy. Sci. Rep. 5, 17210.

López-Rituerto, E., Savorani, F., Avenoza, A., Busto, J.H., Peregrina, J.M., Engelsen, S.B., 2012. Investigations of La Rioja Terroir for wine production using ^{1}H NMR meta-bolomics. J. Agric. Food Chem. 60, 3452–3461.

Lupoi, J.S., Singh, S., Parthasarathi, R., Simmons, B.A., Henry, R.J., 2015. Recent innovations in analytical methods for the qualitative and quantitative assessment of lignin. Renew. Sustain. Energy Rev. 49, 871–906.

Lupoi, J.S., Singh, S., Simmons, B.A., Henry, R.J., 2014. Assessment of lignocellulosic biomass using analytical spectroscopy: an evolution to high-throughput techniques. BioEnergy Res. 7, 1–23.

Lynd, L.R., Laser, M.S., Bransby, D., Dale, B.E., Davison, B., Hamilton, R., Himmel, M., Keller, M., McMillan, J.D., Sheehan, J., Wyman, C.E., 2008. How biotech can transform biofuels. Nat. Biotechnol. 26, 169.

Mansfield, S.D., Kim, H., Lu, F., Ralph, J., 2012. Whole plant cell wall characterization using solution-state 2D NMR. Nat. Protoc. 7, 1579.

Masili, A., Puligheddu, S., Sassu, L., Scano, P., Lai, A., 2012. Prediction of physical–-chemical properties of crude oils by ^{1}H NMR analysis of neat samples and chemometrics. Magn. Reson. Chem. 50, 729–738.

Mounet, F., Lemaire-Chamley, M., Maucourt, M., Cabasson, C., Giraudel, J.-L., Deborde, C., Lessire, R., Gallusci, P., Bertrand, A., Gaudillère, M., Rothan, C., Rolin, D., Moing, A., 2007. Quantitative metabolic profiles of tomato flesh and seeds during fruit development: complementary analysis with ANN and PCA. Metabolomics 3, 273–288.

Mullen, C.A., Strahan, G.D., Boateng, A.A., 2009. Characterization of various fast-pyrolysis bio-oils by NMR spectroscopy. Energy Fuels 23, 2707–2718.

Ni, M., Leung, D.Y.C., Leung, M.K.H., Sumathy, K., 2006. An overview of hydrogen

production from biomass. Fuel Sci. Technol. Int. 87, 461–472.
Papotti, G., Bertelli, D., Graziosi, R., Silvestri, M., Bertacchini, L., Durante, C., Plessi, M., 2013. Application of one- and two-dimensional NMR spectroscopy for the characterization of protected designation of origin lambrusco wines of modena. J. Agric. Food Chem. 61, 1741–1746.
Pauly, M., Gille, S., Liu, L., Mansoori, N., de Sousa, A., Schultink, A., Xiong, G., 2013. Hemicellulose biosynthesis. Planta 238, 627–642.
Pontes, J.G.M., Brasil, A.J.M., Cruz, G.C.F., de Souza, R.N., Tasic, L., 2017. NMR-based metabolomics strategies: plants, animals and humans. Anal. Method. 9, 1078–1096.
Ren, J.-L., Sun, R.-C., 2010. Chapter 4 - hemicelluloses. Cereal Straw as a Resource for Sustainable Biomaterials and Biofuels. Elsevier, Amsterdam, pp. 73–130.
Rieppo, L., Rieppo, J., Jurvelin, J.S., Saarakkala, S., 2012. Fourier Transform Infrared Spectroscopic Imaging and Multivariate Regression for prediction of proteoglycan content of articular cartilage. PloS One 7, e32344.
Ritota, M., Marini, F., Sequi, P., Valentini, M., 2010. Metabolomic characterization of Italian sweet pepper (*Capsicum annum* L.) by means of HRMAS-NMR spectroscopy and multivariate analysis. J. Agric. Food Chem. 58, 9675–9684.
Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., Crocker, D., 2008. Technical Report NREL/TP-510-42618.
Sluiter, A., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., 2005. Technical Report NREL/TP-510-42619.
Son, H.-S., Hwang, G.-S., Ahn, H.-J., Park, W.-M., Lee, C.-H., Hong, Y.-S., 2009. Characterization of wines from grape varieties through multivariate statistical analysis of ^{1}H NMR spectroscopic data. Food Res. Int. 42, 1483–1491.
Son, H.-S., Kim, K.M., van den Berg, F., Hwang, G.-S., Park, W.-M., Lee, C.-H., Hong, Y.-S., 2008. ^{1}H Nuclear Magnetic Resonance-based metabolomic characterization of wines by grape varieties and production areas. J. Agric. Food Chem. 56, 8007–8016.
Tarchevsky, I.A., Marchenko, G.N., 1991. Chapter 6 - Cellulose: Biosynthesis and Structure. Springer, Berlin, pp. 94–97.
Theodoridis, G.A., Gika, H.G., Want, E.J., Wilson, I.D., 2012. Liquid chromatography–mass spectrometry based global metabolite profiling: a review. Anal. Chim. Acta 711, 7–16.
U.S. Department of Energy, 2011. In: Perlack, R.D., Stokes, B.J. (Eds.), Billion-ton Update: Biomass Supply for a Bioenergy and Bioproducts Industry. Oak Ridge National Laboratory, Oak Ridge, TN.
Vargas-Moreno, J.M., Callejón-Ferre, A.J., Pérez-Alonso, J., Velázquez-Martí, B., 2012. A review of the mathematical models for predicting the heating value of biomass materials. Renew. Sustain. Energy Rev. 16, 3065–3083.
Viggiani, L., Morelli, M.A.C., 2008. Characterization of wines by Nuclear Magnetic Resonance: a work study on wines from the basilicata region in Italy. J. Agric. Food Chem. 56, 8273–8279.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–245.
Wyman, C.E., 2007. What is (and is not) vital to advancing cellulosic ethanol. Trends Biotechnol. 25, 153–157.
Xu, F., Yu, J., Tesso, T., Dowell, F., Wang, D., 2013. Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: a mini-review. Appl. Energy 104, 801–809.