Development of chemometric model for characterization of non-wood by FT-NIR data

Mohammad Nashir Uddin\textsuperscript{a},*, Taslima Ferdous\textsuperscript{b}, Zahidul Islam\textsuperscript{a}, M. Sarwar Jahan\textsuperscript{a}, M.A. Quaiyyum\textsuperscript{b}

\textsuperscript{a} BCSIR Laboratories, Dhaka, Bangladesh Council of Scientific and Industrial Research (BCSIR), Dhaka 1205, Bangladesh
\textsuperscript{b} Department of Applied Chemistry and Chemical Engineering, University of Dhaka, Dhaka 1000, Bangladesh

**A R T I C L E  I N F O**

Keywords:
Agricultural residue
Multivariate modeling
Lignocellulose component

**A B S T R A C T**

In this study, a model for prediction of lignocellulose components of agricultural residues has been developed with Fourier Transformed Near Infrared (FT-NIR) spectroscopy data. Two calibration techniques (Principal Component Regression (PCR) and Partial Least Square Regression (PLSR)) were assessed for prediction of lignin, holocellulose, \(\alpha\)-cellulose, pentosan and ash, and found the PLSR better for lignin, holocellulose and \(\alpha\)-cellulose. The PCR also produced better results for quantification of pentosan and ash. Spectral range (7000–5000 cm\(^{-1}\)) showed more informative than other parts of the spectral data. The PLSR showed maximum value of \(R^2=0.91\%\) for prediction of holocellulose. For the prediction of pentosan, the PCR was better (\(R^2=0.68\%\)). The PCR also showed better results (\(R^2=86\%\)) for quantification of ash. To determine amount of lignin, the PLSR was the best (\(R^2=0.83\%\)) when the spectral data were de-trained and smoothed with Savitzky-Golay (S-G) filtering simultaneously. For prediction of \(\alpha\)-cellulose, the PLSR was the best model (\(R^2=0.94\%\)) when the data were pretreated with mean normalization. Considering the best alternatives in Near Infrared (NIR) data preprocessing and calibration techniques, methods for quantification of lignocellulose components of agricultural residues have been developed which is rapid, cost effective, and less chemical intensive and easily usable in pulp and paper industries and pulp testing laboratories.

1. Introduction

Non-wood plants have been used as raw material for pulp and paper industries in many countries of the world (Saijonkari-Pahkala, 2001). Being an agricultural country, Bangladesh is generating substantial number of agricultural residues such as rice straw, wheat straw, bagasse, which could be a potential raw material in pulp and paper industries in Bangladesh (Jahan et al., 2016).

The chemical compositions of lignocelluloses determine its suitability for pulping. The chemical characteristics of agricultural wastes generated in Bangladesh like rice straw, wheat straw, corn stalks, Dhainchha, were determined in order to assess the suitability of these non-wood raw materials for pulp and cellulose industries (Nuruddin et al., 2011; Ferdous et al., 2020). Cellulose is the prime factor of a pulping raw material, which determines the pulp yield (Ferdous et al., 2020), on the other hand, lignin and ash are undesirable parts of the raw materials. Pentosan is also important part in selecting pulping the materials, which facilitates fiber bonding. Moreover, many studies also have been done on pulping of agricultural residues (Akhtaruzzamen and Shafi, 1993; Roy et al., 1998; Jahan et al., 2005; Jahan, 2009). Based on the available literature, chemical components are determined by traditional wet

* Corresponding author.

\textit{E-mail address: m2nashir@yahoo.com} (M.N. Uddin).

https://doi.org/10.1016/j.jobab.2020.07.005

Available online 29 July 2020

2369-9698/© 2020 Published by Nanjing Forestry University. This is an open access article under the CC BY-NC-ND license.

(\url{http://creativecommons.org/licenses/by-nc-nd/4.0/})
laboratory method (Hussain et al., 2002), which are tedious, time consuming and involves of hazardous chemicals. Therefore, the possibility of using easy, rapid, cost-effective techniques which could predict chemical characteristics of agricultural residues was tempting.

Recently interest is growing on non-destructive methods for chemical characterization of pulp, wood and non-wood using spectroscopy and chemometric (Rajesh and Ray, 2006; Hattori et al., 2012; Fahey et al., 2019). Application of infrared spectroscopy as an alternative approach allows for fast and reliable analysis by non-destructive means (Fiserova et al., 2012). This technique requires much smaller samples than those necessary for analysis by conventional wet methods. An advantage to the method is also the fact that in most cases no sample preparation prior to analysis is needed.

In recent years, spectroscopic instruments such as Near Infrared (NIR), Raman, Ultra-violete (UV) spectroscopy are used to predict these parameters in different wood samples with the assistance of multivariate techniques (Silva et al., 1999; Poke and Raymond, 2006; Alves et al., 2007; Colares et al., 2015; He et al., 2015; Li et al., 2015; Kothiyal et al., 2015; Santos et al., 2016; Uddin et al., 2017, 2019). For lignin content determination in Sitka Spruce, a chemometric method was developed by Silva et al. (1999) by Fourier Transform Infrared (FT-IR) Spectrometry and Principal Component Regression (PCR). Poke and Raymond (2006) developed a chemometric method with NIR spectral data and linear calibration for determination of lignin and cellulose. Another method was developed by Li et al. (2015) with Artificial Neural Network (ANN) in NIR spectroscopic data for determination of hemicelluloses, cellulose and lignin in Moso Bambo. A few researches have been conducted to quantify these components in non-wood samples using chemometric models (Huang et al., 2010; Uddin et al., 2017, 2019).

Now it is necessary to develop universal models for quantification of lignin, holocellulose, a-cellulose, pentosan and ash which would be applicable for all type of agricultural residues, potential raw materials for pulp industries. But none of the research has been carried out to develop such spectroscopic method for this purpose by chemometric techniques.

Therefore, the present study is attempting to develop cost effective, non-destructive and rapid method for quantification of lignocellulose components in agricultural residues based raw materials of pulp industries with NIR spectroscopic data of agricultural residues, potential non-wood raw materials for pulp and paper industries and chemometric modeling techniques.

In order to develop models for quantification of lignocelluloses components (namely, lignin, holocellulose, a-cellulose, pentosan, extractives and ash), efficiency of two calibration techniques, PCR and Partial Least Square Regression (PLSR), have been compared. At the same time, most informative region of spectral data has been chosen by calibrating model with different segments of the full spectral data as a part of dimension reduction of data matrix. Finally, different de-noising techniques have been assessed to get the best performing model with pretreated data.

2. Materials and methods

2.1. Samples

Twenty-two distinct non-wood samples, mostly agriculture waste were collected from different parts of the country. The samples are: 1) bamboo (Bambusoideae); 2) bagasse (Saccharum); 3) rice (Oryza sativa) straw; 4) wheat (Triticumaestivum) straw; 5) corn (Zea mays) stalks; 6) mustard (Brassica juncea) stalks; 7) banana (Musa cavendish) pseudo stem; 8) banana leaf; 9) banana; 10) eggplant (Solanummelongena) stalks peduncle; 11) chia (Salvia hispanica) stalks; 12) kash (Saccharumspontaneum) stalks; 13) kaun (Setearia-litilika) straw; 14) okra (Abelmoschus esculentus) stalks; 15) jute (Corchorus) fiber; 16) jute sticks; 17) cotton (Gossypium) stalks; 18) red lentil (Lensculinaris) stalks; 19) dhaincha (Sesbaniaaculeata) stalks; 20) pineapple (Ananascomosus) stalks; 21) mulberry (Morus) plant; and 22) cassava (Manihot esculenta) plant.

2.2. Quantification of lignocellulosic components

The agricultural residues were collected from Mymensingh region of Bangladesh. These residues were chopped to the size of 2–3 cm. Then they were ground in a Wiley mill of about 40–60 mesh size which were used for chemical analysis.

The composition of these agricultural residues were carried out by following Technical Association of Pulp and Paper Industries (Tappi) Test Methods: the extractive (T204 om88), klonl lignin (T211 om83), ash content (T211 om93) and pentosan (T223). The holocellulose was prepared by treating extractive free wood meal with NaClO2 solution (Rajesh and Ray, 2006). The pH of the solution was maintained at 4 by adding CH3COOH–CH3COONa buffer and cellulose was determined by treating holocellulose with 17.5% NaOH (T203 om93).

2.3. Fourier transformed near infrared (FT-NIR) spectroscopic data acquisition

The FT-NIR spectroscopy was performed using a PerkinElmer FT-NIR spectrometer (Model: Frontier, Perkin Elmer, USA) with Indium Gallium Arsenide (InGaAs) detector. The spectral range used was 10,000–4000 cm–1. For each sample, 32 scans were collected at a spectral resolution of 16 cm–1 with an interval of 4 cm–1, then the 32 scans were averaged and stored as reflectance percentage (%R). Here PerkinElmer Spectrum (Version 10.4.4) software was used for spectral data processing. Reflectance percentage of light by certain sample against each wave number was the main consideration for spectral data.

In each spectrum there was huge number of reflectance value for each wave point. Each wave point or data point was considered as spectroscopic variable. These variables were huge in number and were mutually correlated. Therefore, they required dimension reduction and preprocessing.
2.4. Preprocessing of spectroscopic data

After acquisition of spectroscopic data, they had been be preprocessed with different de-noising techniques commonly used in chemometric techniques.

2.4.1. Dimension reduction

Three segments of the full spectral data were taken to find the most informative part of the dataset. They were 8000–7000, 7000–5500 and 5200–4060 cm⁻¹. Calibration was done with full range (8000–4000 cm⁻¹) and other three parts of the dataset.

2.4.2. Smoothing or de-noising

The data were smoothed and denoised with mean normalization, de-trending, de-trending and smoothing with Savitzky-Golay (S-G) filtering as well as de-trending and second derivative together.

2.5. Chemometric model development and validation

Two most popular chemometric calibration techniques namely Principal Component Regression (PCR) and Partial Least Square Regression (PLSR) were assessed for quantification of properties of pulp like lignin, holocellulose, α-cellulose, pentosan and ash with NIR spectral data, and selected the best performing one for method development.

Developed models for quantification of lignin, holocellulose, α-cellulose, pentosan and ash were validated by creating 6-fold cross validation dataset in each case. All the samples were presented in different colored line diagram in the Fig. 1. Data of reflectance percentage are plotted against wavelength.

3. Results and discussion

3.1. Principal component analysis (PCA)

The PCA, a multivariate statistical tool that allows us to summarize the information content in large dataset by means of smaller set of summary indices called Principal Components. It has been widely used for pattern recognition, signal processing and factor analysis. In practical, the PCA has been used in this study to find multicollinearity, outliers in the dataset and finally, to development models with this principal component.

From the PCA it could be found that the first principal component 69% and second one 23%, and both together expressed 92% of the total variation in the data (Fig. 2).

Score plot of the PCA, also known as map of samples, showed that samples were scattered without following any pattern which was the indicator that there was no multicollinearity in the data (Fig. 2). Therefore, the spectroscopic data were suitable for model development.

Influence plot is popularly used to diagnose whether there is any outlier in the dataset which might have adverse impact on the developed models. From Fig. 3, it could be found that there was no outlier in the FT-NIR spectroscopic data of agricultural residues as there was no sample plotted beyond the read lines.

All parts of spectral data of a sample are not equally informative. Therefore, in order to find the most important section of the spectral data, full spectra were divided into three parts and calibration models were developed with data of full spectral data, and the segmented ranges. Efficiency of the two calibration techniques PCR and PLSR were assessed with the full range data (8000–4000 cm⁻¹) and partial datasets 8000–7000, 7000–5500 and 5200–4060 cm⁻¹. Calibrations were also done separately with calibration dataset and 6-fold cross validation (CV) datasets. Coefficients of multiple determination (R²) is one of the parameters to test the model efficiency. Here, comparisons were shown in terms of the values of R².
Fig. 2. Score plot of FT-NIR data of agricultural residues.

Fig. 3. Influence plot of FT-NIR data of agricultural residues.

Table 1
Range selection of FT-NIR data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Dataset</th>
<th>Full range (8000–4000 cm⁻¹)</th>
<th>8000–7000 cm⁻¹</th>
<th>7000–5500 cm⁻¹</th>
<th>5200–4060 cm⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PCR</td>
<td>PLSR</td>
<td>PCR</td>
<td>PLSR</td>
</tr>
<tr>
<td>Lignin</td>
<td>Calibration</td>
<td>0.242</td>
<td>0.262</td>
<td>0.228</td>
<td>0.237</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.141</td>
<td>0.092</td>
<td>0.150</td>
<td>0.096</td>
</tr>
<tr>
<td>Holocellulose</td>
<td>Calibration</td>
<td>0.746</td>
<td>0.882</td>
<td>0.006</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.215</td>
<td>0.473</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>α-cellulose</td>
<td>Calibration</td>
<td>0.781</td>
<td>0.745</td>
<td>0.193</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.271</td>
<td>0.208</td>
<td>0.000</td>
<td>0.057</td>
</tr>
<tr>
<td>Pentosan</td>
<td>Calibration</td>
<td>0.014</td>
<td>0.016</td>
<td>0.001</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Extractive</td>
<td>Calibration</td>
<td>0.006</td>
<td>0.032</td>
<td>0.005</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Ash</td>
<td>Calibration</td>
<td>0.580</td>
<td>0.549</td>
<td>0.392</td>
<td>0.399</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>0.329</td>
<td>0.190</td>
<td>0.141</td>
<td>0.203</td>
</tr>
</tbody>
</table>

Notes: PCR, Principal Component Regression; PLSR, Partial Least Square Regression; CV, cross validation.

Among full range and all three segmented ranges of spectral data most information range is 7000–5500 cm⁻¹ (Table 1) for prediction models of lignin, holocellulose, α-cellulose, pentosan, extractive and ash in different agricultural wastes, potential raw materials for pulp industries. In another study, NIR range 7502–4246 cm⁻¹ was found the most informative region for predicting holocellulose in plantation timber (Kothiyal et al., 2015).
Between two most popular calibration techniques PCR and PLSR, the later one showed better prediction efficiency in all cases with calibration and cross-validation datasets except pentosan. In case of the pentosan, the values of $R^2$ were very close, but a slight better result could be found for the PCR ($R^2 = 68\%$) over the PLSR ($R^2 = 63\%$). Similar result could be noticed for Ash prediction model.

### 3.2. Pre-treatment of spectral data

Spectral data contain some noise or unwanted background information which should be reduced or eliminated in order to improve the robustness of calibration techniques used in this study. Therefore, the FT-NIR spectral data had gone through some pretreatment processes, namely, mean normalization, de-trending, de-trending and smoothing with Savitzky-Golay (S-G) technique combined and finally, de-trending and 2nd derivative together. Like earlier, efficiency of the PCR and PLSR are assessed both for calibration and cross-validation datasets.

As shown in Table 2, calibration model for prediction of $\alpha$-cellulose in agricultural residues is the best when the FT-NIR spectral data are preprocessed with mean normalization with calibration dataset ($R^2 = 94\%$) and calibration dataset ($R^2 = 79\%$). For lignin prediction model, detrending and smoothing with Savitzky-Golay (S-G) techniques showed the best results ($R^2 = 83\%$) when they were used simultaneously. However, a severe effect could be observed when the data were pretreated with detrending and 2nd derivative combined. Here $R^2$ was tremendously high ($R^2 \approx 99\%$) for calibration dataset but for cross-validation data, and the figure was horrifically low, which made the model very unstable for using to predict these parameters from FT-NIR spectral data. Calibration performance was found very poor in terms of $R^2$ for prediction of $\alpha$-cellulose (65%–82%) and lignin (52%–95%) in solid wood without any pre-treatment (Yeh et al., 2005).

### 3.3. Calibration models

For prediction of lignin, the PLSR calibration model showed better performance than the PCR with FT-NIR spectral data detrended and smoothed with Savitzky-Golay (S-G) filtering simultaneously ($R^2 = 83\%$) (Fig. 4). It required seven factors to reach the destination of getting the best performing model.

Holocellulose could be predicted with raw spectral data of range 7000–5500 cm$^{-1}$ with the PLSR ($R^2 = 91\%$) (Fig. 5). Here, model needed seven factors for getting the best performing model.

To predict $\alpha$-cellulose non-wood samples especially agricultural residues, the PLSR performed better than the PCR with FT-NIR spectral data of range 7000–5500 cm$^{-1}$ when they were preprocessed with mean normalization ($R^2 = 94\%$) (Fig. 6). The model required seven factors to perform best. The result obtained under this study for predicting $\alpha$-cellulose was very close to another study conducted by Uddin et al. (2017) with Dhaincha sample ($R^2 = 95\%$) and Jute sample ($R^2 = 99\%$) where FT-NIR spectral data were pre-processed with standard normal variate (SNV) and calibrated with PLSR and artificial neural network (ANN) respectively.

The PCR performed better than the PLSR with raw spectral data of range 7000–5500 cm$^{-1}$ (Fig. 7). The model consisted of first six principal components (PCs) but the model was not recommendable one to use in practice as its predictive performance was not up to the mark ($R^2 = 68\%$).

Lastly, the performance of the PCR and PLSR were same for prediction of ash in the sample, even though slightly better results could be obtained from the later one with raw spectral data of range 7000–5500 cm$^{-1}$ ($R^2 = 86\%$) (Fig. 8). Here the PLSR model was developed with first six factors. The model was more stable than models with other conditions.

The results obtained in the study were much better than those of Huang et al. (2010) who used the NIR for rice straw samples. In that study, $R^2 > 0.80$ for cellulose, 0.60 < $R^2$ < 0.70 for lignin, and $R^2 < 0.70$ for hemicelluloses by the PLSR. Another study involving *Eucalyptus globules* predicted by multivariate linear regression (MLR) showed $R^2$ ranged from 0.67 to 0.87 (Poke and Raymond, 2006).
Fig. 4. Model for prediction of lignin (Blue line indicates calibration model; red line indicates validation model; and the black line indicates target line. This connotations are same in other figures as well (Figs. 5–8). Here ‘Y’ is the data matrix of dependent variable (lignin).

Fig. 5. Model for prediction of holocellulose.

Fig. 6. Model for prediction of α-cellulose.
The model for prediction of pentosan is shown in Fig. 7.

Fig. 7. Model for prediction of pentosan.

The model for prediction of ash is shown in Fig. 8.

Fig. 8. Model for prediction of ash.

A study conducted by Li et al. (2015) on moso bamboo found at best $R^2$ of 0.91 for hemicelluloses, 0.98 for cellulose, and 0.94 for lignin, where the NIR was used for spectral data processing and the ANN for calibration.

4. Conclusions

Among the segmented parts of the spectral data, the range 7000–5000 cm$^{-1}$ showed as most informative region. The PLSR (calibration model) showed maximum $R^2$ value (0.91%) for prediction of holocellulose. For prediction of pentosan, the PCR was better ($R^2 = 0.68$%). The PCR also showed better results ($R^2 = 0.86$%) for quantification of ash. However, some preprocessing treatment was required for quantification of lignin and $\alpha$-cellulose. In case of lignin, the PLSR is the best ($R^2 = 0.83$%) when the spectral data were de-trained and smoothed with Savitzky-Golay (S-G) filtering simultaneously. For prediction of $\alpha$-cellulose, the PLSR was the best model ($R^2 = 0.94$%) when the data were pretreated with mean normalization. Calibration models for quantification of lignin, holocellulose, $\alpha$-cellulose, pentosan and ash were developed with chemometric techniques and FT-NIR spectroscopic data. Finally, these models could be used by pulp and paper industries with minimum cost and time as no further chemicals is needed for quantification of these parameters by producing no chemical waste and keeping the environment clean thereby.

References


