RAPID DETERMINATION METHOD OF METHOXYL CONTENT IN LIGNIN BY CHEMOMETRIC ANALYSIS OF FT-IR SPECTROSCOPIC DATA

MOHAMMAD NASHIR UDDIN,[*] M. NUR ALAM LIKHON,[*] M. MOSTAFIZUR RAHMAN,[*] M. TUSHAR UDDIN,[*] M. KAMAL HOSSAIN[**] and M. SARWAR JAHAN[*]

[*]*Pulp and Paper Research Division, Bangladesh Council of Scientific and Industrial Research Laboratories, Dhaka, Dr. Qudrat-i-Khuda Road, Dhaka 1205, Bangladesh*
[**]*Soil, Water and Environment Laboratory, Bangladesh Council of Scientific and Industrial Research Laboratories, Dhaka, Dr. Qudrat-i-Khuda Road, Dhaka 1205, Bangladesh*
✉ *Corresponding author: M. S. Jahan, sarwar2065@hotmail.com*

Received August 12, 2025

Lignin is the second largest biopolymer on earth, and an important feedstock of biomaterials and biochemicals. The methoxyl group (-OCH_3) content in lignin controls its properties. To elucidate lignin's structure, quantification of –OCH_3 a critical step in lignin characterization. With FT-IR spectral data of 30 lignin samples, chemometric models – Support Vector Regression (SVR) and Partial Least Squares Regression (PLSR) – have been established. These SVR and PLSR were calibrated with the full FT-IR data and its different segments, and the range 3600-3100 cm^{-1} was found the most informative. Both models have been developed based on raw and pretreated spectroscopic data with Moving Average (MA), Mean Normalization (MN), Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC). The predictive performance of the PLSR model is better for predicting methoxyl contents (R^2=89.3%), but the model is unstable in the validation stage (R^2=26.3%) with the data pretreated by MSC. Through this exercise, the best model for predicting methoxyl group in lignin is PLSR, when it is calibrated with the raw spectral data of FT-IR spectroscopic data for the range 3600-3100 cm^{-1} (R^2=82.5% for calibration and 53.5% for validation dataset). Finally, the predicted values of the methoxyl content by the developed model are very close to the values calculated by the tedious and time-consuming wet method.

Keywords: lignin, methoxyl group, FT-IR spectral data, partial least squares regression

INTRODUCTION

Lignin, the second most abundant biopolymer on Earth, is derived from wood and other lignocellulosic materials. As the world transitions toward a circular bioeconomy, lignin plays a crucial role due to its reactive functional groups, such as hydroxyl, methoxyl, and carboxyl groups. It serves as a sustainable alternative to petroleum-based polymers in bio-based polymer production. Currently, bio-based polymers account for approximately 1% of global polymer production, with an annual output of 4.2 million tons.[1] Although cellulose, paper and other products like lignosulfonates are also biobased products, these were not included in this account.

Lignin is composed of syringyl (S), guaiacyl (G), and p-hydroxyphenyl (H) units in varying proportions. The S unit contains two methoxyl groups at the 3- and 5-positions, the G unit has one methoxyl group at either the 3- or 5-position, and the H unit lacks methoxyl groups on the benzene ring. The S/G ratio significantly influences lignocellulose delignification[2] and affects wood hydrolysis for bioethanol production.[3] Additionally, Shimizu *et al*.[4] demonstrated that the β-O-4 bond in syringyl lignin is more susceptible to alkaline-induced cleavage than that in guaiacyl lignin. To elucidate lignin's structure, its empirical formula is determined based on methoxyl group (-OCH_3) content, making –OCH_3 quantification a critical step in lignin characterization.

Traditionally, methoxyl content in lignin has been determined by using the Zeisel–Viebock–Schwappach method, which involves hazardous bromine.[5] To overcome this limitation, alternative methods have been developed. Li *et al*.[6] introduced a technique based on headspace gas

chromatography (HS-GC), where methoxyl groups are cleaved with hydroiodic acid (HI) at 130 °C for 30 min, forming methyl iodide, which is then quantified via flame ionization detection. Sumerskii *et al*.[7] advanced this approach by employing headspace-isotope dilution GC-MS (HS-ID GC-MS) for precise quantification of methoxyl and ethoxyl groups in lignin. Their method demonstrated superior precision and accuracy compared to the classical Zeisel–Viebock–Schwappach approach. These methods including the traditional ones are expensive, time consuming, complicated and need to go through sample preparation hassles. The principle of the Zeisel–Vieböck–Schwappach method for determining the methoxyl ($-OCH_3$) group content in lignin is based on the quantitative cleavage of ether linkages by concentrated hydro iodic acid followed by volumetric titration. The method proceeds through three primary chemical stages: (i) cleavage of methoxyl groups, (ii) oxidation of methyl iodide and (iii) quantitative titration. Therefore, the method is time-consuming and destructive.

As an alternative, chemometric modeling has been extensively explored in forest products research, particularly for quantifying lignocellulosic components, pulp properties, and lignin characteristics in recent years.[8] Our group has developed robust chemometric models based on multivariate analysis of spectroscopic data to analyze lignocellulose and pulp. These models effectively quantify key chemical components in agricultural residues ($R^2 \approx 0.99$, FT-NIR, ANN) and predict pulp properties from jute using FT-NIR with PLSR. Specific models have demonstrated strong predictive performance, achieving R^2 values of 83% for lignin (PLSR with Savitzky-Golay filtering) and 94% for α-cellulose (PLSR with mean normalization).[9–12]

Similarly, numerous chemometric models have been developed for determining the content of hexenuronic acid (HexA) in kraft pulps using various spectroscopic techniques, including FT-NIR spectrophotometry,[13,16] UV Resonance Raman Spectroscopy,[14] and FTIR.[15] These models have been successfully applied to both unbleached and bleached chemical pulps derived from diverse wood species, such as *Eucalyptus globulus*, *Pinus radiata*, and Norway spruce. More recently, Uddin *et al*.[13] developed a rapid FT-IR-based method for quantifying HexA in non-wood pulp, achieving high predictive accuracy ($R^2 = 94.24\%$) using a PLSR model with Savitzky-Golay filtering, derivatives, and leverage correction. The model's reliability was further validated by its exceptional performance in predicting HexA content in unknown non-wood samples. Additionally, Uddin *et al*.[17] successfully quantified the guaiacyl-to-syringyl ratio in lignin, with PLSR yielding the highest predictive accuracy ($R^2 = 99.90\%$) when applied to FT-NIR data processed with Savitzky-Golay filtering, derivatives, and leverage correction.

A few studies have been conducted for developing multivariate calibration models with spectroscopic data based chemometric techniques for quantification of OCH_3 group in lignin from wood samples,[18-20] even though these studies exercised only PLSR as calibration technique, and the full range of spectra were used in these studies. Another research has been reported[18] where a regression model for predicting methoxyl groups content (OCH_3) was developed based on Oxygen (O) and Hydrogen (H) contents.

Despite these advancements, no studies have yet reported the quantification of OCH_3 groups in lignin using multivariate analysis and chemometric modeling with lignin from non-wood multi-species samples by examining the supremacy of the developed model over other calibration methods and by searching for the most informative regions of the spectra. Therefore, this study aims to develop a feasible, rapid, and environmentally friendly and non-destructive method for quantifying methoxyl groups in lignin, utilizing chemometric modeling in combination with FT-IR spectroscopic data by selecting the most informative regions for the prediction model. In addition, three calibration models have been explored and the best performing one was found. Finally, the methoxyl groups ($-OCH_3$) in lignin in non-wood samples have been quantified by the developed models, and the results were compared with those obtained by the traditional method, which has not been encountered in previously published research papers.

EXPERIMENTAL

Materials

Fifteen non-wood lignocellulosic materials were collected from different parts of the country. The samples were: jute stick (*Corchorus olitorius*), sugarcane bagasse (*Saccharum officinarum*), mulberry (*Morus alba*), banana tree (*Musa acuminata*), corn stalk (*Dracaena fragrans*), chia (*Salvia hispanica*), rice straw (*Oryza sativa*), kash (*Saccharum spontaneum*), zara (*Pennisetum purpureum*), bamboo (*Bambusa vulgaris*),

coconut husk (*Cocos nucifera*), jute fiber (*Corchorus olitorius*), hogla (*Typha elephantina*), bringle (*Solanum melongena*) and dhaincha (*Sesbania bispinosa*). From these lignocellulosic materials 30 (15 technical and 15 dioxane lignin) samples have been prepared. The methoxyl contents of technical and proto lignin from non-wood lignocellulosic materials were calculated by the Zeisel–Viebock–Schwappach approach.[5]

Dioxane and technical lignin extraction

Dioxane-lignin was extracted from extract-free meal samples using acidic dioxane solution. The extraction involved refluxing the samples in a hydrochloric acid-dioxane solution (dioxane-to-meal ratio of 8) for 1 h under a nitrogen atmosphere (50 mL/min flow rate). The resulting meal-dioxane mixture was filtered through Whatman filter paper No. 42. The filtrate was concentrated by vacuum evaporation at 40 °C. The lignin was precipitated from the filtrate by dropwise addition into deionized water with constant stirring. The precipitate was then centrifuged, washed up to neutral pH, and vacuum-dried over P_2O_5. Further purification was achieved by dissolving the dried lignin in aqueous dioxane (9:1) and re-precipitating it in ether. Finally, the purified lignin was vacuum-dried in a desiccator containing P_2O_5. The dioxane lignin in this paper is considered equivalent to protolignin.

For getting technical lignin, pulping of these lignocellulosic materials was performed using the soda process under the conditions specified in Table 1 to isolate technical lignin. Pulping was conducted in small bombs of 100 mL capacity immersed in an electrically heated oil bath. The bombs containing the raw materials and cooking chemicals were placed in the bath at room temperature. The temperature was then raised to either 150 °C or 170 °C, taking 40 minutes and 60 minutes, respectively. After the desired cooking time, the bombs were immediately cooled in flowing tap water. Following cooling, the bombs were opened, and the cooked mass was filtered in a Büchner funnel to separate the liquor from the pulp. Technical lignins were precipitated from the soda pulping liquor by adjusting the pH to 2 using dilute sulfuric acid. The isolated lignin was then purified by dissolution in aqueous dioxane (9:1), followed by precipitation in ether. The isolated lignin was vacuum dried over P_2O_5.

Table 1
Conditions for extraction of technical lignin from non-wood lignocellulosic materials

Raw materials	NaOH (% based on raw material)	Material : Liquor	Temperature (°C)
Jute stick	18	1:10	170
Sugarcane bagasse	18	1:10	150
Mulberry	18	1:10	170
Banana tree	18	1:10	150
Corn stalk	18	1:10	150
Chia	18	1:10	170
Rice straw	18	1:10	150
Kash	18	1:10	150
Zara	18	1:10	150
Bamboo	18	1:10	170
Coconut husk	18	1:10	170
Jute fiber	18	1:10	150
Hogla	18	1:10	150
Bringle	18	1:10	170
Dhaincha	18	1:10	170

Reaction time was 120 min for all raw materials

Determination of methoxyl group

The methoxyl group of lignin was determined according to Chen *et al*.[21] For the experiment, 30.0 mg of moisture-free lignin was placed in a two-necked reaction flask. To this, 3.5 mL of concentrated hydroiodic acid, 0.5 g of phenol, and 6 drops of acetic anhydride were added. The mixture was then heated to 140 °C in an oil bath. The volatile methyl iodide produced during the reaction was quantitatively stripped from the solution by purging it with carbon dioxide. The methyl iodide gas was collected in a trap containing a 25 mL solution of sodium acetate in glacial acetic acid, to which a few drops (5-6) of concentrated bromine were added. This reaction was allowed to proceed for 30 minutes. After the reaction, the absorbed solution was transferred to a 250 mL conical flask, and 25 mL of 20% sodium acetate was added. The mixture was shaken vigorously to neutralize it, and 4% formic acid was added dropwise until discoloration occurred, indicating the removal of excess bromine. Finally, the iodine produced was titrated with a standardized sodium thiosulfate solution using methyl red as an indicator.

FT-IR spectroscopic data acquisition

Fourier Transform Infrared Spectroscopy (FT-IR) spectroscopy was performed using a PerkinElmer FT-IR spectrometer (Model: Frontier, Perkin Elmer, USA) with GAAS detector. Attenuated Total Reflectance

(ART) sampling mode was used in Fourier Transform Infrared Spectroscopy (FTIR) for getting the spectral data. The spectral range used was 4000-650 cm^{-1}. For each sample, 32 scans were collected at a spectral resolution of 16 cm^{-1} with an interval of 1 cm^{-1}, then 32 scans were averaged and stored as transmittance percentage (%). Here, PerkinElmer Spectrum (Version 10.4.4) software was used for spectral data processing.

In a spectrum, each wavenumber corresponds to a specific reflectance value, which is vital for spectral analysis. However, the vast number of reflectance values across the wavenumber range leads to a high volume of spectroscopic variables that are often strongly correlated. Therefore, dimensionality reduction and pre-processing are vital steps to handle this complexity effectively.

Informative spectral range selection

The whole FT-IR spectral data are not equally contributing to developing calibration models. Therefore, we explored different segments and tried to find the better performing models along with their full range. The most efficient ranges 2900-2800 cm^{-1} and 3600-3100 cm^{-1} showed better predictive results than other segments of the spectra.

Preprocessing of spectral data

After acquisition of spectral data from FT-IR, at first, they were preprocessed with some transformations. In the study, smoothing techniques such as Mean Normalization (MN), Moving Average (MA), Standard Normal Variage (SNV) and Multiplicative Scatter Correction (MSC) were applied and their efficiencies were assessed.

Mean Normalization (MN) is a frequently used procedure for de-noising spectroscopic data consisting in normalising the continuum to unity by dividing the observed spectrum by a smooth approximation of its continuum. This approximation can be obtained by dividing the raw data by itself after filtering or smoothing. Median filtering and running average algorithms are well suited for this purpose. A spline fit can be made to the points defined interactively as well as to the filtered data.[22]

Moving Average (MA), also known as a rolling average, running average, or moving mean, is a method used to analyze data by computing a sequence of averages over different subsets of a dataset. Mathematically, it represents a form of convolution. A moving average filter smooths data by replacing each point with the average of its neighboring values within a defined window. This process generates a new series in which each value is derived from the average of corresponding raw observations in the original dataset.[23]

Standard Normal Variate (SNV) is another commonly used pretreatment in FT-IR spectroscopy to eliminate scatter. It is applied individually to each spectrum. The mean and standard deviation of all data points are calculated, and each data point is then mean-centered by subtracting the average and scaling by the standard deviation.[24]

Multiplicative Scatter Correction (MSC) is a widely used normalization technique designed to adjust spectra so they closely align with a reference spectrum, typically the mean of the dataset. This is achieved by modifying both the scale and offset of the spectra.[25] MSC compensates for additive and multiplicative effects in spectral data through a row-oriented transformation, meaning each data point is influenced by its neighboring values. By applying MSC, physical effects such as particle size variations and surface scatter – which do not contain relevant chemical or physical information – are effectively removed from the spectra.[26]

As pretreatments of FT-IR spectroscopic data, Mean Normalization (MN) is used to remove overall intensity differences between spectra, Moving Average (MA) – to smooth the spectra by averaging points within a moving window, Standard Normal Variate (SNV) – to correct for multiplicative scatter effects (differences in particle size, packing, or light scattering) and Multiplicative Scatter Correction (MSC) – to correct both additive and multiplicative effects in spectra caused by scattering or path length differences.

Chemometric calibration techniques

Principal Component Regression (PCR) is the most frequently used among the chemometric calibration techniques applied in different studies.[9–12,14,15,17] However, in this particular study PCR demonstrated worse results than Support Vector Regression (SVR) and Partial Least Square Regression (PLSR).

Support Vector Regression (SVR) is a variant of Support Vector Machine (SVM), a supervised learning algorithm designed for both classification and regression tasks. While SVMs aim to identify the optimal hyperplane that separates different classes in a high-dimensional space, SVR focuses on finding a function that accurately predicts continuous output values based on input data.[27,28]

Partial Least Square Regression (PLSR) extracts key components from the independent variables (X) that best predict the dependent variable (Y). It achieves this by identifying a set of latent vectors that simultaneously decompose both X and Y while maximizing their covariance. This approach builds upon the principles of Principal Component Analysis (PCA). After decomposition, a regression step follows, where the latent vectors obtained from X are used to model Y. PLS regression represents both X and Y as a combination of a common set of orthogonal factors and their associated loadings.[29,30]

Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) were assessed for quantifying methoxyl (-OCH_3) group in lignin using FT-IR spectroscopic data. The model demonstrating the best performance was selected for further method development.

To validate the models, 6-fold cross-validation was performed in each case. A total of 30 lignin samples were used as calibration data for model development, while an independent test set of 3 lignin samples was employed to evaluate the predictive performance of the selected model for quantification of methoxyl group.

For PCR, five principal components (latent variables) were considered, and Singular Value Decomposition (SVD) was used as the model input algorithm. In PLSR, up to seven latent factors were included, with Nonlinear Iterative Partial Least Squares (NIPALS) used as the modeling algorithm, allowing a maximum of 100 iterations. Random cross-validation was applied in both approaches. These settings yielded the highest predictive accuracy.

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. It is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. A 6-fold cross-validation approach was employed in all cases to reduce the risk of overfitting and ensure reliable evaluation of calibration accuracy. A set of 4 test data have been used to investigate the prediction efficiency of the models.

CAMO, the Unscrambler software (version 10.3), was used for calibration and validation of PCR and PLSR. Next, spectroscopic data of the lignin samples were collected on the Fourier Transformed-Near Infrared (FT-IR) spectrophotometer.

Next, the data were pretreated with Mean Normalization (MN), Moving Average (MA), Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC). Finally, all forms of data, raw and treated were used for developing chemometric models. Here predictive efficiencies of two sets of models, Support Vector Regression (SVR) and Partial Least Square Regression (PLSR) have been assessed.

Finally, with the better alternatives, new chemometric assisted spectroscopic methods were proposed for quantification of methoxyl group in protolignin and technical lignin, which will reduce cost, time and chemicals used.

RESULTS AND DISCUSSION

The methoxyl contents of technical and proto lignin from non-wood lignocellulosic materials calculated by the Zeisel–Viebock–Schwappach approach are presented in Table 2.

FT-IR spectroscopic data of transmittance (%) against wavenumber (cm^{-1}) are depicted in Figure 1 ranging from 4000 to 715 cm^{-1}.

Principal Component Analysis (PCA)

The FT-IR spectroscopic data of lignin samples were analyzed using Principal Component Analysis (PCA), and the resulting score plot is presented in Figure 2. The plot reveals that the samples are scattered and distributed randomly, without any discernible pattern. This suggests the absence of autocorrelation among the samples, indicating their suitability for developing multivariate models.[31]

Table 2
Methoxyl contents of technical and proto lignin from non-wood lignocellulosic materials

Technical lignin			Proto lignin		
No.	Raw materials	Methoxyl content (wt%)	No.	Raw materials	Methoxyl content (wt%)
TL 1	Jute stick	15.31	PL 1	Jute stick	18.76
TL 2	Sugarcane bagasse	15.88	PL 2	Sugarcane bagasse	18.08
TL 3	Mulberry	25.89	PL 3	Mulberry	21.2
TL 4	Banana pseudo-stem	16.72	PL 4	Banana pseudo-stem	14.18
TL 5	Corn stalk	20.62	PL 5	Corn stalk	20.19
TL 6	Chia	24.01	PL 6	Chia	16.15
TL 7	Rice straw	19.09	PL 7	Rice straw	16.47
TL 8	Kash	18.64	PL 8	Kash	17.87
TL 9	Zara grass	18.98	PL 9	Zara grass	12.78
TL 10	Bamboo	15.96	PL 10	Bamboo	16.31
TL 11	Coconut husk	19.12	PL 11	Coconut husk	15.99
TL 12	Jute fiber	17.98	PL 12	Jute fiber	11.43
TL 13	Hogla	16.91	PL 13	Hogla	16.5
TL 14	Bringle	15.82	PL 14	Bringle	12.08
TL 15	Dhaincha	17.96	PL 15	Dhaincha	10.52

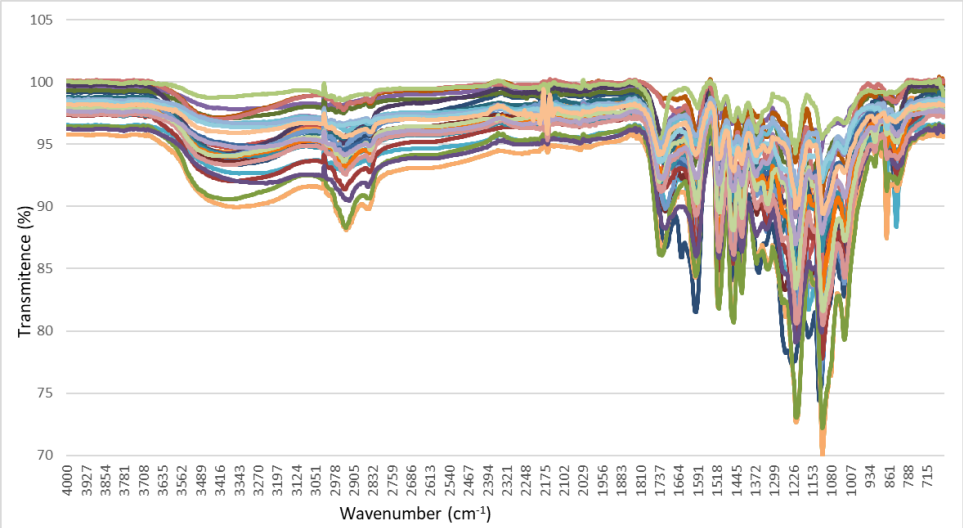

Figure 1: Presentation of FT-IR spectral data

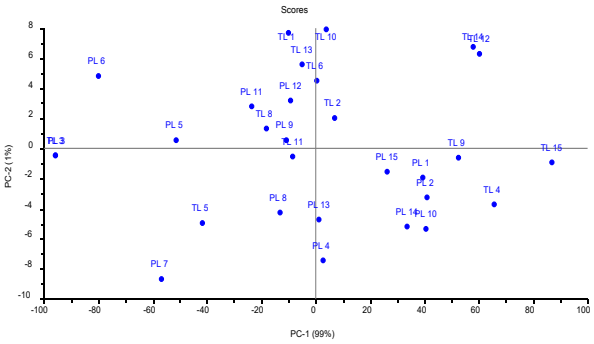

Figure 2: Scores plot of first two principal components of PCA

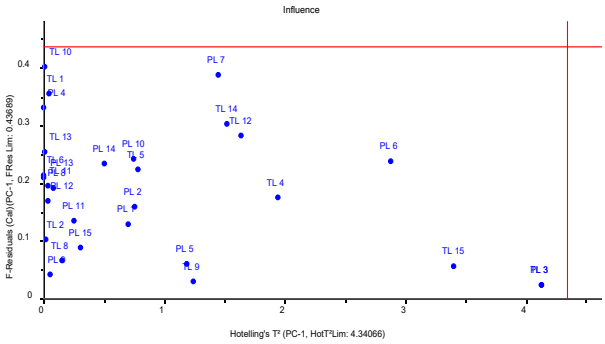

Figure 3: Influence plot of PCA

Principal Components (PCs) are newly derived variables created by transforming a large set of original variables into a smaller set, while retaining most of the original information. These components are arranged so that the first Principal Component (PC1) captures the largest possible variation in the data, followed by the second (PC2), and so on. In this case, the first seven components collectively account for 99.99% of the total variation in the dataset. These components are selected by Root Mean Square Error of Cross-Validation (RMSECV) within 1 standard error (SE) of the minimum because it favors parsimony and reduces overfitting as addition of any components does not significantly reduce prediction error.

To detect any potential outliers in the sample set that could adversely affect the efficiency of the model, influence plots were utilized within the PCA framework. The PCA influence plot displays F-residuals plotted against Hotelling's T^2 values. Diagnostic analysis of this plot confirms that no outliers are present in the dataset, as shown in Figure 3.

Calibration models

Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) are widely recognized as effective and reliable calibration techniques for the analysis of lignocellulosic materials.[9,10,32,33] Model performance was assessed by using two model efficiency criteria: the Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2). Here, model parameters (Slope and Offset) and model efficiency parameters (RMSE and R^2) were evaluated through the study. The results in terms of values and lines are presented both for calibration (blue) and validation (red) datasets in Figures 4-7.

In this study, though the first two principal components expressed almost all the variables among the samples (99.99%) in PCA, Principal Component Regression (PCR) has not showed satisfactory results (R^2=11.4%). That means,

spectroscopic data are not responding to the experimental data through PCR. So, two other calibration technique, Support Vector Regression (SVR) and Partial Least Squares Regression (PLSR) have been used here. Next, SVR and PLSR models were developed for the quantification of methoxyl content using raw FT-IR spectroscopic data of lignin samples.

Support Vector Regression (SVR) is not a suitable candidate for prediction with raw spectral data (R^2=26%) with raw FT-IR data. However, PLSR produces better results among these techniques (R^2=82.50% for calibration and 53.49% for validation dataset). So, in the quest of a better predictive model, pretreatment and selection of a better informative range of spectral data becomes necessary.

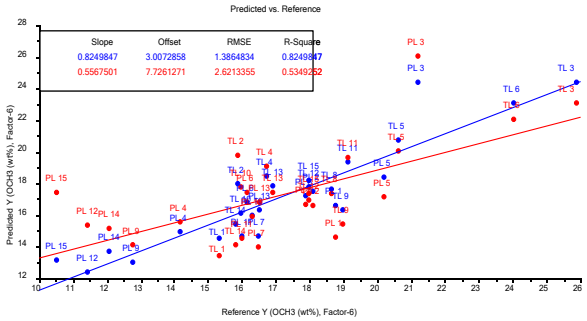

Figure 4: PLSR model for predicting OCH_3 with raw data of the spectral range 3600-3100 with calibration (blue) and validation (red) datasets

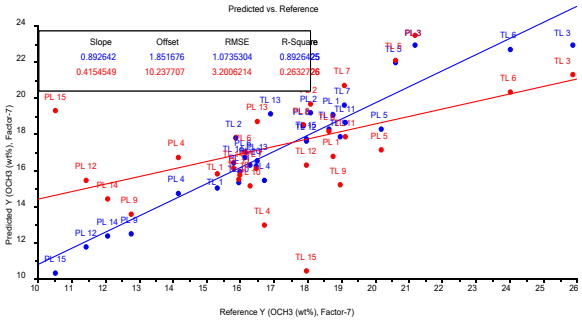

Figure 5: PLSR with MSC of spectral data of the range 3600-3100

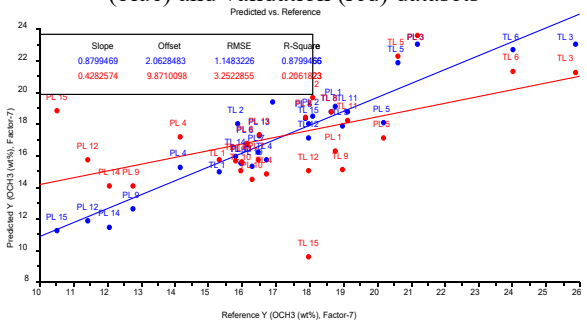

Figure 6: PLSR with SNV of spectral data of the range 3600-3100

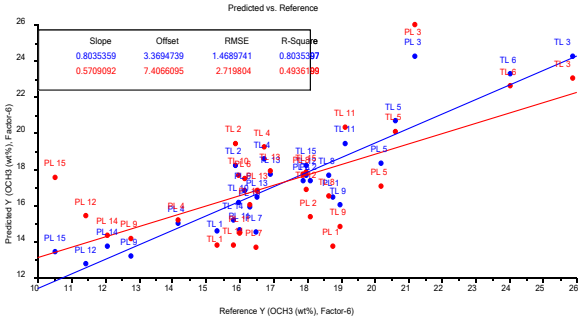

Figure 7: PLSR with Moving Average of spectral data of the range 3600-3100

Table 3
Efficiencies of models for predicting OCH_3 (wt%) with different ranges of FT-IR spectral raw data

Dataset	Full range		2900-2800 cm^{-1}		3600-3100 cm^{-1}	
	SVR	PLSR	SVR	PLSR	SVR	PLSR
Calibration	0.300	0.487	0.219	0.689	0.256	0.825
Validation	0.000	0.112	0.000	0.434	0.018	0.535

Effective spectral range selection

FT-IR spectral data in its full range might not be equally contributing for developing the calibration model. So, it might become necessary to find out the most effective part of the spectral data. Different segments of FT-IR spectra have been used for developing SVR and PLSR models. The most efficient spectral ranges 2900-2800 cm^{-1} and 3600-3100 cm^{-1} show better predictive results than others, and the results are presented in Table 3.

Effects of pretreatment of spectral data

The region 3600–3100 cm^{-1} is dominated by O–H stretching (broad alcohol/phenolic/carboxylic acid bands), not methoxy C–O or CH_3 vibrations. However, in chemometric modeling, each sample gives spectroscopic data, which act as a finger print of that sample. That is why, the spectral range of FTIR shows differently in the modelling.

Among different spectral ranges, 3600-3100 cm^{-1} gives the best results. The spectral values of this range have been treated with different de-

noising techniques. Among them, Normalization, Moving Average, Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) show better predictive efficiencies than other pretreatment techniques of spectral data.

The PLSR model and its efficiency parameters have been depicted in Figure 5, where the model efficiency is satisfactory (R^2=89.26%) with calibration data, but not at all satisfactory (R^2=26.33%) with validation data when the FT-IR data of the range 3600-3100 were pretreated with MSC.

Similarly, the PLSR model SNV-transformed FT-IR data of the range 3600-3100 are illustrated in Figure 6. Here, the model shows satisfactory efficiency with the calibration data (R^2 = 87.99%), its performance with the validation data is notably poor (R^2 = 20.62%).

However, the PLSR model based on SNV-transformed FT-IR data in the range of 3600–3100 cm^{-1} is presented in Figure 7. While the model demonstrates satisfactory performance with the calibration data (R^2 = 80.35%), its performance on the validation data is moderate (R^2 = 49.362%).

Two calibration models, namely, Support Vector Regression (SVR) and Partial Least Square Regression (PLSR) have been explored with raw and pretreated FT-IR spectral data of the range 3600-3100 cm^{-1}, and presented in Table 4. The predictive performance of PLSR by far outweighed that of SVR in all cases.

The highest predictive efficiency has been shown when the spectral data are pretreated with Multiplicative Scatter Correction (MSC) in calibration (R^2=89.3%) and validation (26.3%). It indicates that though the model has high predictive efficiencies, it is not stable to use universally. On the contrary, the predictive efficiency of the PLSR model with raw spectral data is high (R^2=82.5%) in the calibration stage, and the model is stable as well (53.5%) in the validation stage for quantification of OCH_3 in lignin from non-wood samples. In a published study, the PLSR model efficiency is higher (R^2=94.84%) (18) with wood samples than in our study.

Table 4
Efficiencies of models for predicting OCH_3 (wt%) with 3600-3100 cm^{-1} range of FT-IR spectral data by different pretreatment techniques

Dataset	Raw data		Normalization		Moving Average (MA)		Standard Normal Variate (SNV)		Multiplicative Scatter Correction (MSC)	
	SVR	PLSR	SVR	PLSR	SVR	PLSR	SVR	PLSR	SVR	PLSR
Calibration	0.256	0.825	0.442	0.780	0.254	0.804	0.229	0.880	0.257	0.893
Validation	0.018	0.535	0.174	0.379	0.014	0.494	0.023	0.206	0.023	0.263

Table 5
Predicted values of OCH_3 (wt%) by developed PLSR models with FT-IR spectral data in 3600-3100 cm^{-1} range, and their deviation from fractional method

Sample ID	Values of OCH_3 (calculated by Zeisel–Viebock–Schwappach method)	Predicted values of OCH_3 by developed models (Deviations from values of Zeisel–Viebock–Schwappach method)	
		Raw data	MSC
PS1	20.62	20.73 (0.11)	21.93 (-1.31)
PS2	15.96	16.17 (0.21)	15.29 (0.67)
PS3	16.15	16.84 (0.69)	16.72 (-0.57)
PS4	12.78	13.23 (0.30)	12.45 (0.33)

Prediction efficiency of the development models

Finally, with the best performing PLSR model of raw spectral data and FT-IR data pretreated by Multiplicative Scatter Correction (MSC), the methoxyl group contents in lignin of four samples were predicted and the results are presented in Table 5. These predicted values of methoxyl group are very close to the values calculated with the PLSR model with raw FT-IR spectral data of the range 3600-3100 cm^{-1}, which indicates very good predictive efficiency of the models developed in the study.

CONCLUSION

Initially, SVR and PLSR models were developed using different segments of FT-IR spectral data to quantify the methoxyl group in non-wood lignin from thirty lignin samples derived

from various non-wood sources. In all cases, the PLSR model outperformed the SVR model.

The best performance of the PLSR model was observed when calibrated with spectral data in the range of 3600–3100 cm^{-1}, achieving an R^2 of 82.5% in calibration and 45% in validation. Subsequently, various mathematical preprocessing techniques, including Normalization, Moving Average, Standard Normal Variate (SNV), and Multiplicative Scatter Correction (MSC), were applied. Among these, MSC preprocessing yielded the highest predictive efficiency, with R^2 values of 89.3% in calibration and 26.3% in validation. However, despite its high predictive accuracy, the model's low validation performance suggests limited generalizability. In contrast, the PLSR model using raw spectral data exhibited strong predictive efficiency (R^2 = 82.5%) during calibration and demonstrated better stability (R^2 = 53.5%) in validation.

In the end, the methoxyl group was successfully quantified using the PLSR model with raw FT-IR spectral data in the 3600–3100 cm^{-1} range, as the predicted values closely matched the measured values. This approach allows for the quantification of the methoxyl group in non-wood lignin samples solely based on FT-IR spectral data, eliminating the need for chemical reagents or labor-intensive procedures. As a result, this method provides a non-destructive, rapid, and cost-effective alternative for determining the methoxyl group in lignin from similar sources.

ACKNOWLEDGEMENTS: Authors wish to thanks Bangladesh Council of Scientific and Industrial Research for providing necessary funds to carry out the research.

REFERENCES

[1] Anon., *The Global Market for Biobased Packaging 2025–2035*, 2024, available at: https://www.futuremarketsinc.com/the-global-market-for-biobased-packaging-2025-2035/ (accessed April 17, 2025)

[2] V. L. Chiang and M. Funaoka, *Holzforschung*, **44**, 309 (1990), https://doi.org/10.1515/hfsg.1990.44.4.309

[3] F. Kacík, J. Durkovic and D. Kacíková, "Lignin: Structural Analysis, Applications in Biomaterials and Ecological Significance", Nova Science Publishers, Hauppauge, NY, 67–89, 2014

[4] S. Shimizu, T. Akiyama, T. Yokoyama and Y. Matsumoto, *J. Wood Chem. Technol.*, **37**, 451 (2017), https://doi.org/10.1080/02773813.2017.1340957

[5] G. F. Zakis, "Functional Analysis of Lignins and Their Derivatives", TAPPI Press, Atlanta, GA, USA, 1994

[6] H. Li, X. S. Chai, M. Liu and Y. Liu, *J. Agric. Food Chem.*, **60**, 5307 (2012), https://doi.org/10.1021/jf300455g

[7] I. Sumerskii, T. Zweckmair, H. Hettegger, G. Zinovyev, M. Bacher *et al.*, *RSC Adv.*, **7**, 22974 (2017), https://doi.org/10.1039/c7ra00690j

[8] M. N. Uddin, M. M. Rahman, M. N. A. Likhon and M. S. Jahan, *Nord. Pulp Pap. Res. J.*, **40**, 71 (2025), https://doi.org/10.1515/npprj-2024-0060

[9] M. N. Uddin, S. Ahmed, S. K. Ray, A. H. Quadery and M. S. Jahan, *Nord. Pulp Pap. Res. J.*, **34**, 1 (2019), https://doi.org/10.1515/npprj-2018-0018

[10] M. N. Uddin, S. K. Ray, M. S. Islam, J. Nayeem and M. S. Jahan, *J. Sci. Technol. For. Prod. Process*, **6**, 22 (2017)

[11] M. N. Uddin, T. Ferdous, Z. Islam, M. S. Jahan and M. A. Quaiyyum, *J. Bioresour. Bioprod.*, **5**, 196 (2020), https://doi.org/10.1016/j.jobab.2020.07.005

[12] M. N. Uddin, J. Nayeem, M. S. Islam and M. S. Jahan, *Biomass Conv. Bioref.*, **9**, 585 (2019), https://doi.org/10.1007/s13399-019-00383-8

[13] X. S. Chai, J. Y. Zhu and J. Li, *J. Pulp Pap. Sci.*, **27**, 165 (2001)

[14] A. M. Saariaho, B. Hortling, A. S. Jaaskelainen, T. Tamminen and T. Vuorinen, *J. Pulp Pap. Sci.*, **29**, 363 (2003)

[15] V. Hoang, N. K. Bhardwaj and K. L. Nguyen, *Carbohyd. Polym.*, **61**, 5 (2005), https://doi.org/10.1016/j.carbpol.2004.12.007

[16] M. Monrroy, R. Mendonça, J. Baeza, J. Ruiz, A. Ferraz *et al.*, *J. Near Infrared Spectrosc.*, **16**, 121 (2008), https://doi.org/10.1255/jnirs.766

[17] M. N. Uddin, T. Ferdous, Y. Jin, M. M. Rahman and M. S. Jahan, *Anal. Sci. Adv.*, **6**, e70005 (2025), https://doi.org/10.1002/ansa.70005

[18] M. Jablonsky, M. Botkova and J. Adamovska, *Cellulose Chem. Technol.*, **49**, 165 (2015), https://www.cellulosechemtechnol.ro/pdf/CCT2(2015)/p.165-168.pdf

[19] Y. Jiao, Z. Li, X. Chen and S. Fei, *J. Chemom.*, **34**, e3306 (2020), https://doi.org/10.1002/cem.3306

[20] A. A. Al-Mbaideen, *J. Anal. Chem.*, **74**, 686 (2019), https://doi.org/10.1134/s1061934819070013

[21] S. D. Brown, R. Tauler and B. Walczak, *Anal. Bioanal. Chem.*, **396**, 551 (2009), https://doi.org/10.1007/s00216-009-3284-9

[22] W. Windig, J. Shaver and R. Bro, *Appl. Spectrosc.*, **62**, 1153 (2008), https://doi.org/10.1366/000370208786049097

[23] M. R. Maleki, A. M. Mouazen, H. Ramon and J. De Baerdemaeker, *Biosyst. Eng.*, **96**, 427 (2007), https://doi.org/10.1016/j.biosystemseng.2006.11.014

[24] M. Awad and R. Khanna, "Efficient Learning Machines", Apress, Berkeley, CA, 2015, pp. 67–80, https://doi.org/10.1007/978-1-4302-5990-9_4

[25] F. Zhang and L. J. O'Donnell, "Machine Learning", Elsevier, 2020, pp. 123–140, https://doi.org/10.1016/B978-0-12-815739-8.00007-9
[26] R. Kramer, "Chemometric Techniques for Quantitative Analysis", CRC Press, 1998, https://doi.org/10.1201/9780203909805
[27] S. Wold, M. Sjöström and L. Eriksson, *Chemom*. *Intell*. *Lab*. *Syst*., **58**, 109 (2001), https://doi.org/10.1016/S0169-7439(01)00155-1
[28] I. T. Jolliffe, "Principal Component Analysis", Springer Series in Statistics, Springer-Verlag, New York, 2002, pp. 338–372, https://doi.org/10.1007/978-1-4757-1904-8_8
[29] C. J. Colares, T. Pastore, V. T. Coradin, J. A. Camargos, A. C. Moreira *et al*., *J*. *Braz*. *Chem*. *Soc*., **26**, 1297 (2015), https://doi.org/10.5935/0103-5053.20150096
[30] W. He and H. Hu, *J. Wood Chem*. *Technol*., **33**, 52 (2013), https://doi.org/10.1080/02773813.2012.731463
[31] X. Li, C. Sun, B. Zhou and Y. He, *Sci*. *Rep*., **5**, 17210 (2015), https://doi.org/10.1038/srep17210
[32] Q. Luo and J. Zhu, *J. Sci*. *Technol*. *For*. *Prod*. *Process*., **5**, 6 (2015)
[33] S. Shukla, S. Shashikala and M. Sujatha, *J. Near Infrared Spectrosc*., **29**, 168 (2021), https://doi.org/10.1177/0967033521999118