Carbohydrate Polymers 87 (2012) 1124–1130

Contents lists available at SciVerse ScienceDirect

Carbohydrate Polymers

journal homepage: www.elsevier.com/locate/carbpol

Quantitative determination of cellulose dissolved in 1-ethyl-3-methylimidazolium acetate using partial least squares regression on FTIR spectra

Michael FitzPatrick[a], Pascale Champagne[a,b,*], Michael F. Cunningham[a]

[a] Department of Chemical Engineering, Queen's University, Kingston, ON, Canada K7L 3N6
[b] Department of Civil Engineering, Queen's University, Kingston, ON, Canada K7L 3N6

ARTICLE INFO

Article history:
Received 28 February 2011
Accepted 23 August 2011
Available online 12 September 2011

Keywords:
Cellulose
Ionic liquids
IR spectroscopy
PLS regression
Biotechnology

ABSTRACT

Rapid and quantitative measurements of cellulose concentrations in ionic liquids (ILs) are difficult. In this study, FTIR operated in attenuated total reflectance (ATR) mode was investigated as a tool to measure cellulose concentration in 1-ethyl-3-methylimidazolium acetate ([emim][OAc]) and the spectra were subjected to partial least squares (PLS) regression for the quantitative determination of cellulose content. Additionally, the spectra were subjected to 7 data preprocessing methods to reduce physical effects in the spectra. Peak normalization was found to be the technique that most improved the prediction of dissolved cellulose in [emim][OAc]. When peak normalization was used for data preprocessing, a model for the quantitative estimation of cellulose content between 0 wt.% and 4 wt.% with an error of 0.53 wt.% was generated. The methods described here provide the basis for a rapid and facile technique for the determination of dissolved cellulose content in [emim][OAc].

1. Introduction

Ionic liquids are a class of organic salts that exist as liquids at low temperatures; often well below 100 °C. They have low vapour pressures, good thermal stability, electroconductivity, tunable physico-chemical properties and unique solubility properties (FitzPatrick, Champagne, Cunningham, & Whitney, 2010; Fort et al., 2007; Kiefer et al., 2008; Lee, Doherty, Linhardt, & Dordick, 2008; Yang & Pan, 2005). Since Swatloski, Spear, Holbrey, & Rogers (2002) first published their findings on cellulose dissolution in 1-butyl-3-methylimidazolium chloride, there has been a great deal of interest in the use of ILs in a variety of studies involving biomass, including the dissolution of cellulose, lignin and complete biomass samples, as well as the *in situ* processing and modification of cellulose (Cao et al., 2009; Dadi, Varanasi, & Schall, 2006; Fort et al., 2007; Heinze, Schwikal, & Barthel, 2005; Kilpeläinen et al., 2007; Lateef, Grimes, Kewcharoenwong, & Feinberg, 2009; Lee et al., 2008; Pu, Jiang, & Ragauskas, 2007; Sun et al., 2009; Swatloski et al., 2002; Wu et al., 2004; Zhang, Du, Qian, & Chen, 2010; Zhao et al., 2009). However, the development of novel ILs and IL-biomass dissolution and processing systems is limited by a general lack of simple and reliable quantitative characterization tools and techniques (Joglekar, Rahman, & Kulkarni, 2007).

There is a long history involving the use of infrared spectroscopy biomass sample analysis. It has been used to study the crystallinity and type of cellulose crystal lattice, as well as to distinguish different structural characteristics between cellulose sources (Hurtubise & Krassig, 1960; Nelson & O'Connor, 1964a,b). Infrared spectroscopy has been used to characterize phenolics (including lignin) in cell wall samples and to quantify lignin content (Hatfield & Fukushima, 2005). Fourier transform infrared (FTIR) spectroscopy has also been used in plant cell wall analysis, where the different polysaccharide structures and compositions are identified based on their unique spectral band positions (Kačuráková, Capek, Sasinková, Wellner, & Ebringerová, 2000), as well as for the compositional analysis of potential biomass feedstocks, including corn stover and woody species (Hames, Thomas, Sluiter, Roth, & Templeton, 2003; Moore & Owen, 2001). Lignin and reducing sugar content in hydrolysates from dilute acid pretreatment have been determined using FTIR analysis (Tucker et al., 2001). Furthermore, developments in the use of real time FTIR for process monitoring have enabled *in situ* monitoring of liquid phase sucrose hydrolysis and the online monitoring of ethanol fermentation (Pintar, Batista, & Levec, 2002; Veale, Irudayaraj, & Demirci, 2007). Recently, FTIR spectroscopy has been used to monitor the changes in cellulose crystallinity index and crystallite size after being subjected to NaOH and CO_2 treatments (Oh et al., 2005). FTIR spectroscopy, operated

* Corresponding author at: Department of Civil Engineering, Queen's University, Kingston, ON, Canada K7L 3N6. Tel.: +1 613 533 3053; fax: +1 613 533 2128.
E-mail address: champagne@civil.queensu.ca (P. Champagne).

0144-8617/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carbpol.2011.08.086

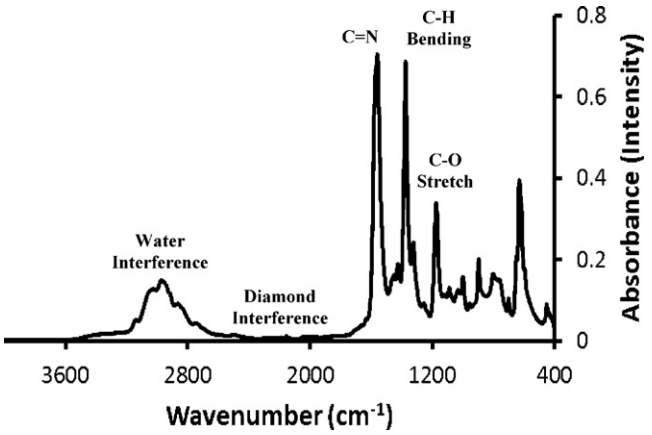

Fig. 1. Representative ATR FTIR spectra (4 wt.% cellulose in [emim][OAc]).

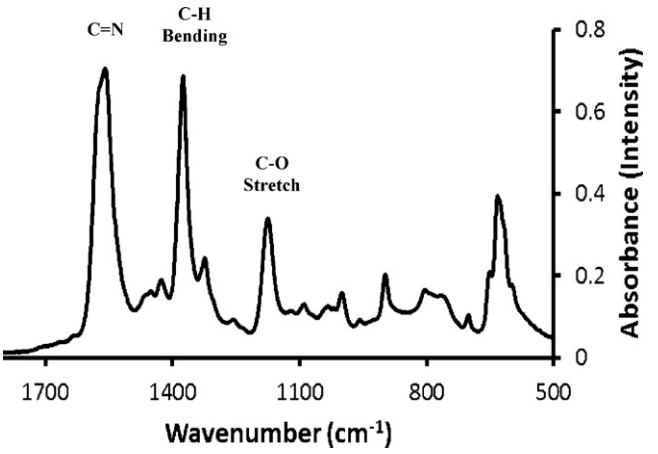

Fig. 2. Representative truncated FTIR spectra (4 wt.% cellulose in [emim][OAc]) demonstrating extended fingerprint region.

in ATR mode, has also been used to quantitatively determine α-D-glucose (from 0 mass% to 20 mass%) dissolved in [emim][OAc] (Kiefer et al., 2008).

Determination of the dissolved cellulose content in ILs, while simple in concept, is often experimentally difficult and/or time consuming (Sievers et al., 2009). Herein, the development of a quantitative FTIR model for predicting dissolved cellulose concentration in 1-ethyl-3-methylimidazolium acetate ([emim][OAc]), at 0–4 wt.% loadings, through the application of partial least squares (PLS) regression is reported. The model allows rapid predictions of dissolved cellulose content in ILs from FTIR spectra. Enhancement of the PLS regression model to improve the accuracy of the predicted cellulose wt.% via data processing techniques was also investigated.

2. Materials and methods

The ionic liquid, [emim][OAc], (≥90%, BASF) was used as received for all dissolution studies. Avicel PH-101 (Sigma–Aldrich), a model microcrystalline cellulose compound, was used as received. Ultra high purity nitrogen was supplied by Praxair.

2.1. IR spectroscopy

FTIR spectra with a resolution of 4 cm^{-1} were obtained from a Nicolet 6700 spectrometer operating in ATR FTIR mode and equipped with a diamond crystal. Thirty-two scans were run per sample between 400 cm^{-1} and 4000 cm^{-1}. Each sample (calibration and validation) was analyzed in triplicate and the resulting spectra were averaged using OMNIC software. The resulting spectra were truncated to an extended fingerprint region (500–1800 cm^{-1}) to eliminate interference from the diamond crystal, nitrogen, and the signature of water in [emim][OAc], while maintaining the C=N peak observed at 1565 cm^{-1}. Representative FTIR spectra, for the full and the truncated fingerprint region, are shown in Figs. 1 and 2, respectively.

2.1.1. Calibration set

To develop the predictive FTIR model for cellulose content in [emim][OAc], a 25-sample calibration set was generated as per the ASTM Standard E 1655-05 (ASTM Committee E13, 2005). The calibration set contained samples that varied between 0 and 5 wt.% cellulose per 1 mL of [emim][OAc] dissolved under a 20 psi nitrogen atmosphere.

2.1.2. Validation set

To assess the predictive capability of the FTIR calibration model for cellulose content in [emim][OAc], a validation set of 10 samples ranging between 0 and 6 wt.% cellulose dissolved under a 20 psi nitrogen atmosphere was employed. The cellulose concentrations used for all calibration and validation sample sets are shown in Table 1.

2.1.3. FTIR model

Multivariate PLS regression was conducted on the ATR FTIR spectra to allow for quantitative analysis and validation of cellulose content dissolved in [emim][OAc]. Each spectrum was subjected to mean centering and scaling of the data to center the data about the origin (Boysworth & Booksh, 2008). Additionally, the spectra were processed via different data preprocessing techniques to correct for physical effects and improve the predictive capabilities of the ATR FTIR model using the Unscrambler X by CAMO Software.

2.1.3.1. Savitzky–Golay differentiation: first and second derivative. Savitzky–Golay differentiation computes derivatives based on a polynomial approximation of a portion of the curve to be regressed. This algorithm fits a polynomial to each successive curve segment, resulting in the replacement of the original curve values with values generating a smoothed curve (Gorry, 1990; Savitzky & Golay, 1964). For the generation of the curve segment to be smoothed, 13 right side and 13 left side points in addition to the starting point were chosen, resulting in a 27-point curve employed for each correction. These 27 points were chosen to ensure the curve exceeded

Table 1
Cellulose concentrations for calibration and validation test sets. Samples beginning with C are calibration set samples and samples beginning with V are validation set samples.

Sample	Cellulose wt.%	Sample	Cellulose wt.%	Sample	Cellulose wt.%
C1	4.00	C13	3.68	C25	2.96
C2	2.00	C14	1.99	V1	2.95
C3	3.97	C15	1.13	V2	3.98
C4	2.00	C16	2.96	V3	3.67
C5	0.98	C17	1.99	V4	6.02
C6	0.00	C18	0.00	V5	1.01
C7	1.00	C19	3.99	V6	1.00
C8	2.01	C20	1.98	V7	0.00
C9	0.00	C21	1.04	V8	2.95
C10	1.58	C22	3.02	V9	2.01
C11	2.49	C23	3.99	V10	2.04
C12	4.92	C24	0.00		

the width of the largest spectral band (at 1565 cm^{-1}) at half of the maximum absorbance (full width half maximum). Additionally, for each correction a second order polynomial was selected to fit the data points.

The first derivative treatment typically results in the generation of a new spectrum that is composed of the slope of the spectral curve at each wavenumber. Generally, this treatment can be used to successfully reduce the effect of baseline offset and linear baseline shift in the spectra. The main drawback associated with this treatment is that it causes a shift in the characteristic peaks, making interpretation of the resulting spectra more difficult.

The second derivative treatment calculates the change in the slope of the curve at each data point and generates a new spectrum. Like the first derivative treatment, this technique is particularly effective in correcting baseline offset and linear baseline shift in spectra. Furthermore, the second derivative treatment does not shift, but rather preserves the location of characteristic peaks, unlike the first derivative treatment, potentially making it a more useful treatment than the first derivative.

2.1.3.2. Multiplicative scatter correction (MSC). Multiplicative scatter correction (MSC) is a data transformation technique that is used to overcome additive and/or multiplicative effects seen in spectra (Dhanoa, Lister, Sanderson, & Barnes, 1994). Using this technique, two effects often seen in spectra, signal amplification and offset, are removed from the data yielding a clearer observation of the actual response. This data correction consists of calculating two correction coefficients, a (offset) and b (slope), from the average spectrum in the data set and using these two coefficients to correct a regression line for each individual spectrum, which is also computed as a function of the average value observed for each wavelength (Helland, Næs, & Isaksson, 1995). As this technique uses the mean spectrum for the data set, its success depends on how well the calculated mean spectrum represents the true mean spectrum and, as such, is more representative when larger sample sets are employed.

2.1.3.3. Standard normal variate (SNV). The SNV transform centers and scales each individual spectrum. The SNV correction results in the removal of multiplicative interferences of scatter and size effects from spectra. It only uses data contained within a given spectrum for standardization (of that spectrum) as opposed to the mean spectrum of the data set (Dhanoa et al., 1994). Practically, this correction is very similar to MSC, with the exception of the application of vertical scaling possible using MSC. This correction can be used in conjunction with detrending as a means to reduce multicollinearity, baseline shift and curvature in spectroscopic data.

2.1.3.4. Linear baseline correction. Baseline correction is a spectral treatment that adjusts a spectral offset by either adjusting the data to a minimum point common to all spectra or by making a linear correction based on pre-defined points (Brereton, 2003). In this study, a linear baseline correction was employed using the first and last truncated spectral points as the endpoints for the correction. These two points were automatically re-defined as 0 and the rest of the spectrum was linearly transformed according to this new baseline. The application of a linear baseline correction forces a sloped baseline into a horizontal baseline.

2.1.3.5. Detrending. Detrending is a correction that is applied to remove nonlinear trends in spectroscopic data by calculating a baseline function as a least squares fit of a polynomial to each individual spectrum (Barnes, Dhanoa, & Lister, 1989). Detrending corrections are applied to an individual spectrum, which is different from many other data treatments that operate at a specific wavelength across the entire spectral set. The order of the polynomial used in the detrending correction determines the baseline effects that are removed. In this study, a third order polynomial was used, allowing for the removal of baseline offset, slope and curvature. Additionally, this correction can be used in conjunction with SNV, where SNV can remove the multiplicative effect of baseline shift and detrending can remove baseline curvature (using a 2nd order or higher polynomial for detrending). Furthermore, the detrending correction does not change the shape of the data, which commonly occurs with derivative-based corrections.

2.1.3.6. Normalization. Normalization describes a series of transformations, computed on each sample individually, to convert all data to approximately the same scale. Specifically, peak normalization was used in this study. This transformation normalizes a sample using a selected data point, identical for both the calibration and validation sets. It is assumed that the spectral point selected does not vary with concentration between different sample spectra. Therefore any change at this spectral point is attributed to an increase or decrease in sample path length, thus this technique corrects for any variation in path length. In this study, the first data point in each spectrum was used as the basis for the peak normalization correction.

3. Results and discussion

When applying non-linear regression techniques to spectral data for quantitative analysis, as well as when applying data preprocessing techniques to reduce baseline effects, it is imperative that a reliable and consistent metric for model performance be selected. In this study, the root mean squared error (RMSE), after application of PLS on the calibration sample set, was used as the decision metric to determine the number of factors to be used in generating a predictive PLS model (Kjeldahl & Bro, 2010). Additionally, the number of correctly predicted validation set samples was determined, for each model, to assess performance. These two measures of model performance were selected for their "concrete" nature, as opposed to more relative indices of model performance such as R^2 (Kjeldahl & Bro, 2010). The percentage of explained variance, Hotelling T^2 statistic and the predicted cellulose wt.% residuals vs. predicted cellulose wt.% were also used, in addition to the RMSE of the actual cellulose wt.% vs. predicted cellulose wt.%, to determine how well the model described the calibration set.

In general, an increase in the explained variance represents an increase in the trend identified by the model. However, such a relative comparison is insufficient to compare the effect of data preprocessing techniques (Kjeldahl & Bro, 2010). The Hotelling T^2 statistic is used to identify sample outliers through the application of hypothesis testing. The predicted cellulose wt.% residuals vs. predicted cellulose wt.% provides a quick graphical representation of whether or not a trend in the residuals exists that is not explained by the model. Here, if the predicted cellulose wt.% residuals vs. predicted cellulose wt.% are randomly distributed, there is no apparent trend and the residuals represent noise, whereas if a relationship exists between the predicted cellulose wt.% residuals and the predicted cellulose wt.%, there is trend in the residuals, suggesting that the model may be inadequate. Finally, the RMSE of the actual cellulose wt.% vs. predicted cellulose wt.% provides an indication of how close the predicted cellulose values were to the actual cellulose values. In this study, a lower RMSE represented a better model prediction of the actual cellulose wt.% in the sample.

For internal model validation, cross-validation was employed using a segmented approach. Each model was subjected to a cross-validation consisting of 10 segments containing either 2 or 3 randomly selected samples, corresponding to roughly 10% of all the samples. This approach aimed to reduce the

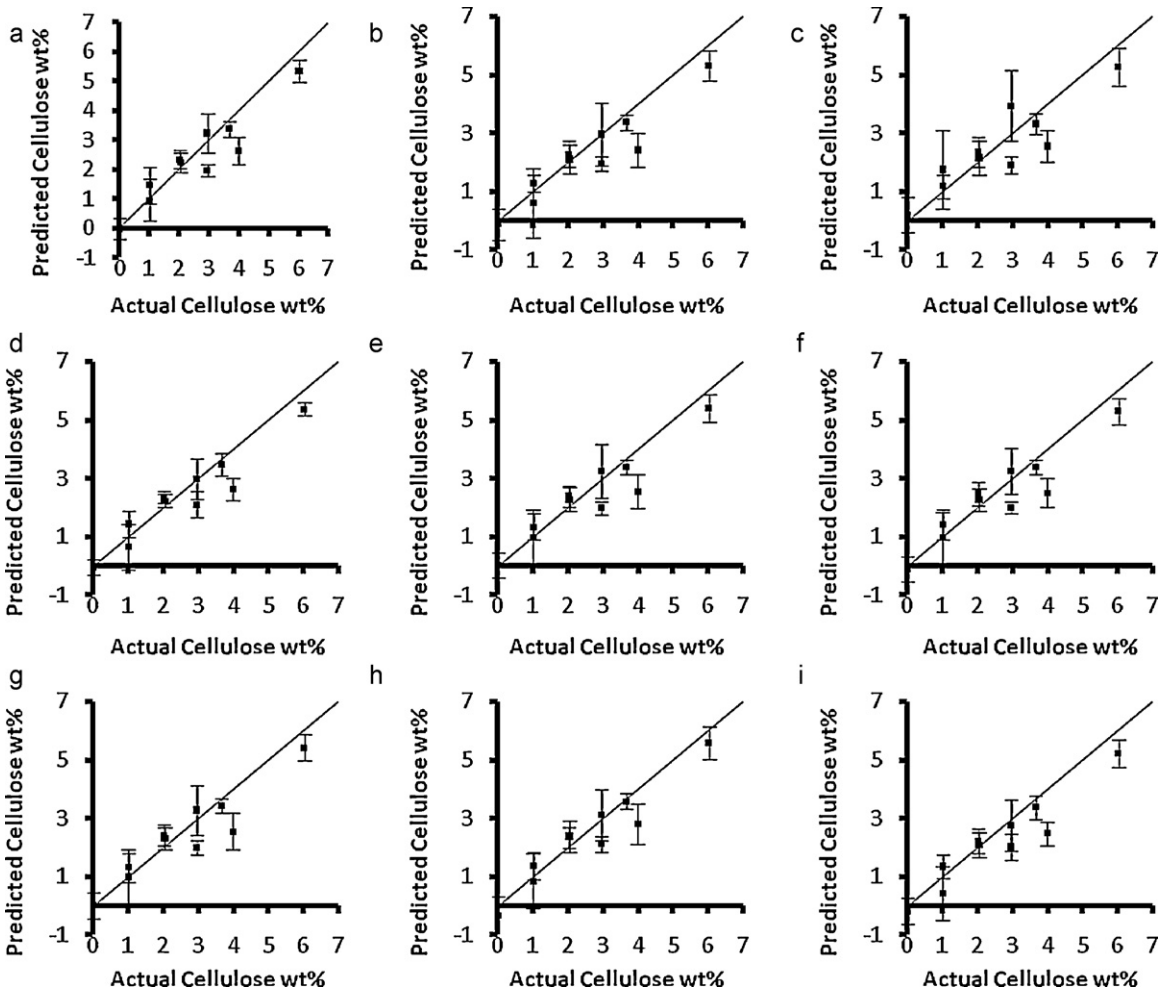

Fig. 3. Predicted vs. actual cellulose wt.% for the following data preprocessing FTIR models: (a) untreated, (b) Savitzky–Golay differentiation – 1st derivative, (c) Savitzky–Golay differentiation – 2nd derivative, (d) MSC, (e) SNV, (f) baseline correction, (g) detrending, (h) normalization, and (i) SNV-detrending. Error bars represent a 95% confidence interval.

bias seen in leave-one-out cross-validation (Kjeldahl & Bro, 2010). Furthermore, cross-validation was used to ascertain the appropriate number of PLS regression factors through the selection of the number of factors for each model that minimized the RMSE (Ghasemi & Niazi, 2001). The resulting PLS models were compared for their prediction capability using the 10 samples in the independent validation set. The root mean squared error of prediction (RMSEP) and the number of samples, within a 95% confidence interval, predicted by the model were used as comparison metrics. The predicted vs. actual cellulose wt.% for the validation standards of the PLS models using each of the data preprocessing techniques is shown in Fig. 3.

PLS regression on the 500–1800 cm^{-1} region of the original (untreated) FTIR calibration spectra set yielded a 3-factor model with an RMSE (for the internal validation) of 0.337. These three factors explained approximately 95% of the variance for both the calibration and validation sets. This model predicted the wt.% of 5 of the 10 validation standards, as seen in Fig. 3a. Finally, the RMSEP for this technique was 0.612 (Table 2). Since we did observe a linear trend in the residuals vs. predicted plot (Supplemental material), the original FTIR data were subjected to eight different data preprocessing treatments to address this noted trend and the results were compared with the untreated data model to assess preprocessing performance. A summary of the model prediction results, including the untreated case, is shown in Table 2.

3.1. Comparison of the data preprocessing techniques

To qualitatively assess whether or not an appropriate number of factors were employed to develop the PLS model, qualitative evaluation metrics, namely explained variance, Hotelling's T^2 statistic and the *Y*-residuals vs. predicted *Y*, were employed in this study (Supplemental material). Generally, a higher number of factors employed in PLS regression corresponds to an increase in the variance explained by the model, however, the use of too many factors in model development can lead to error being explained by the model and the model can, in such cases, over-fit the data. Chen et al. (2010) applied PLS on FTIR spectra to predict cellulose, hemicellulose and lignin content in both hard- and softwood samples. Although the authors were able to develop a model that could explain over 99% and 78% of the variance in the calibration and internal validation sets, respectively, and could predict cellulose content with an RMSEP of 0.96 wt.%, their use of a 9-factor PLS model might suggest a model that over-fit the data (Ghasemi & Niazi, 2001). In this study ≤4 factors were used for the PLS regression model (Table 2). Also, the explained variance for the calibration and internal validation sets was well over 90% for each PLS regression model, which is considered an "excellent data explanation" (Chen et al., 2010). This demonstrates that a high percentage of the data were explained for both the calibration and internal validation sets by each model, while minimizing the potential error explained by the model.

Table 2
Summary of the effects of different data preprocessing treatments on PLS FTIR models to predict cellulose wt.%.

Technique	Number of factors in PLS model	RMSE (wt.%)	Correct number of validation predictions	RMSEP (wt.%)
Untreated	3	0.337	5	0.612
Savitzky–Golay first derivative	4	0.319	6	0.658
Savitzky–Golay second derivative	4	0.407	7	1.80
MSC	2	0.330	6	0.592
SNV	3	0.308	5	0.618
Baseline correction – linear	4	0.308	5	0.652
Detrending	3	0.332	6	0.619
Normalization	4	0.332	8	0.533
SNV-detrending	2	0.353	7	0.662

Ghasemi and Niazi (2001) applied Savitzky–Golay first and second derivative treatments to electronic absorption spectra to determine the concentration of Co^{2+} and Ni^{2+} in binary solution mixtures. Similar to the [emim][OAc]/cellulose system, there is a high degree of overlap in the spectra of Co^{2+} and Ni^{2+}, and application of derivative treatments to the spectra led to a magnification of noise and an increase in the error of prediction (Chen et al., 2010). Thus, the poor performance in terms of predictive capability (Fig. 3b and c) and high RMSEP (Table 2) of the two derivative methods used in this study was also attributed to magnification of noise in the spectra. This effect of derivative preprocessing on spectra also was also observed by Carlini, Massantini, & Mencarelli (2000). Furthermore, the original characteristic spectral peaks (Fig. 2) were shifted after application of the first derivative correction to the calibration spectra. The changes in the characteristic spectra upon application of first and second derivative treatments were also noted by Ghasemi and Niazi (2001).

The use of MSC to correct for signal amplification and offset is a widely used data preprocessing technique (Ghasemi & Niazi, 2001). Johansson, Pettersson, & Folestad (2005) used MSC for data preprocessing in addition to PLS regression on Raman spectra for the quantitative assessment of active pharmaceutical ingredients in pharmaceutical immediate release tablets. They noted that MSC preprocessing prior to PLS regression resulted in a lower RMSE than both second derivative and SNV techniques, as well as assessment of the active pharmaceutical ingredient without data preprocessing and PLS regression. In this study, application of MSC preprocessing to the spectra also resulted in an improved (lower) RMSEP (Table 2), because this technique uses an average spectrum from the dataset to perform corrections (Carlini et al., 2000). However, since MSC depends on the accuracy of the mean spectrum of the data set, it is likely that with an increase in the number of samples in the calibration set, MSC treatment may yield better results than those observed in this study.

The relatively poor performance (low number of correctly predicted samples and high RMSEP) of the linear baseline correction (Table 2) suggests that there was a nonlinear trend affecting the spectra. This was supported by the predictive enhancement afforded by the detrending preprocessing technique (when compared to the linear baseline correction) before PLS regression since detrending is designed to remove nonlinear trends from spectroscopic data. The detrending preprocessing treatment prior to PLS regression resulted in an RMSEP of 0.619 and the correct prediction of 6 of the 10 cellulose wt.% validation standards (Fig. 3g and Table 2). Garrido Frenich, Martínez Galera, Martínez Vidal, & Gil García (1996) applied both linear baseline and SNV-type centering preprocessing techniques, individually, to HPLC chromatograms to enhance the simultaneous determination of several pesticides via multivariate methods. It was shown that applying a linear baseline to the highly overlapped chromatogram peaks did not enhance the predictive capabilities of PLS, demonstrating that there was a stable baseline between samples. However the application of the SNV-type preprocessing technique enhanced the calibration of the PLS modeling, removing multiplicative interference effects from the chromatograms (Garrido Frenich et al., 1996). In this study, models generated from PLS regression on the overlapped calibration spectra (Fig. 1) subjected to both SNV and detrending preprocessing treatments had a lower RMSEP than the linear baseline correction treatment (Table 2). Thus, the calibration spectra in this study likely had stable baselines between samples, but were affected by nonlinear trends in the physical spectral structure. However, both the SNV and detrending preprocessing techniques offered little improvement over the untreated data (Table 2).

When using infrared light for analysis, it is often assumed that the path length of the light is identical for all sample measurements, however slight differences between samples can readily cause a change in length and alter spectra (Kadam, van der Windt, Daudey, & Kramer, 2010). In their experiments dissolving α-lactose monohydrate in water, Kadam et al. (2010) demonstrated that variability in light path length could be eliminated by normalization of the spectra and a good correlation between the spectral data and the concentration could be achieved. In this study, peak normalization was applied to the first data point in each spectrum in the calibration set (4000 cm^{-1}). Since this point contained no signature from either [emim][OAc] or cellulose, the goal of data preprocessing to minimize the physical (and not chemical) differences between sample spectra (Carlini et al., 2000) was achieved, while preserving the characteristic information contained within the original spectra. Here, the lowest RMSEP (0.533 wt.%) and largest number of correct cellulose wt.% validation predictions (8) was achieved by applying a normalization preprocessing treatment to the spectra (Table 2 and Fig. 3h). This indicated the presence of a variable infrared light path length in the spectra that was subsequently minimized via the normalization preprocessing treatment.

Analysis of the Hotelling T^2 statistic indicated that for each of the PLS regression models, the 25 calibration set samples were within a hypothesis test set at a 95% level, implying that no samples within the calibration set were considered to be outliers. Furthermore, there was good agreement between the predicted and actual cellulose wt.% for both the calibration and validation data sets. The *Y*-residuals vs. predicted *Y* (where *Y* is wt.% cellulose) were plotted to determine if any trend could be noted in the residuals, which would indicate the potential for model improvement. Only the models created using the two detrending techniques had no trend in the *Y*-residuals vs. predicted *Y* suggesting that, with the exception of MSC and normalization, the models generated using derivative, SNV and linear baseline data treatments may not be explaining sufficient trend. This is also reflected in the poor to moderate predictive capabilities of these models (Table 2). The relatively good predictive capabilities of the models using MSC and normalization preprocessing data treatments, coupled with the trend in the *Y*-residuals vs. *Y* predicted would imply that these preprocessing techniques are good candidates to provide model enhancement. This improvement, especially for the MSC treatment,

could be achieved through an increase in the size of the calibration set and potentially followed by a change in the number of PLS factors used to build the model.

4. Conclusions

In this study, the RMSEP and number of correct validation samples predicted were used to assess the precision and accuracy of each data preprocessing technique. The RMSEP provides an indication of the level of uncertainty associated with the predicted cellulose wt.% from each model. The number of correctly predicted samples in the validation set provides an indication of the accuracy of the model as it assesses the model capability in repeatedly predicting an actual value (cellulose wt.% in this study). As such, a good model would demonstrate a high accuracy (high number of validation samples predicted) and high precision (small RMSEP); a moderate model would demonstrate either high accuracy or high precision but not both; and a poor model would demonstrate neither high accuracy nor high precision.

Other model evaluation metrics, including the explained variance, Hotelling's T^2 statistic, and the Y-residuals vs. predicted Y, give a qualitative indication of how well the model explains the trend in the raw data. These were used to assess if the model over-fit the data or if there were trends unexplained by the model, which was indicative of whether or not an appropriate number of factors had been selected.

It was demonstrated that the application of a normalization data preprocessing treatment to PLS regression on truncated ATR FTIR spectra enabled the quantitative determination of cellulose loading in [emim][OAc] for samples between 0 wt.% and 4 wt.%, with an RMSEP of 0.533 wt.%. Additionally, 8 of the 10 validation standards were correctly predicted using this method. The treatment of the spectral data via spectral peak normalization allowed for a nearly 13% decrease in the RMSEP compared to the untreated data case. It is believed that an increase in the number of samples used to build both the calibration and validation sets could further enhance these results. Additionally, optimization of the number of factors after PLS regression using the most promising data preprocessing techniques for cellulose dissolved in ILs (normalization and MSC) for a larger calibration set is recommended to further improve the predictions achieved by the technique described in this study.

Acknowledgements

The authors would like to thank the Ontario Graduate Scholarship fund, the Natural Sciences and Engineering Research Council of Canada, the Ontario Ministry of Research and Innovation (Early Researcher Award and Ontario Research Chair in Green Chemistry and Engineering Programs) and Queen's University for funding for this project. The authors would also like to thank Drs. Brian Amsden and Dale Marecak, Department of Chemical Engineering, Queen's University, for assistance with the IR spectrometer.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.carbpol.2011.08.086.

References

ASTM Committee E13 on Molecular Spectroscopy and Chromatography. (2005). *Standard practices for infrared multivariate quantitative analysis. E 1655-05* (pp. 1–29).

Barnes, R. J., Dhanoa, M. S. & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*, 772–777.

Boysworth, M. K. & Booksh, K. S. (2008). Aspects of multivariate calibration applied to near-infrared spectroscopy. In D. A. Burns, & E. W. Ciurczak (Eds.), *Handbook of near-infrared analysis* , pp. 209–239.

Brereton, R. G. (2003). *Chemometrics: data analysis for the laboratory and chemical plant*. Etobicoke, ON, Canada:: John Wiley & Sons Canada Ltd.

Cao, Y., Wu, J., Zhang, J., Li, H., Zhang, Y. & He, J. (2009). Room temperature ionic liquids (RTILs): a new and versatile platform for cellulose processing and derivatization. *Chemical Engineering Journal*, *147*, 13–21.

Carlini, P., Massantini, R. & Mencarelli, F. (2000). Vis-NIR measurement of soluble solids in cherry and apricot by PLS regression and wavelength selection. *Journal of Agricultural and Food Chemistry*, *48*, 5236–5242.

Chen, H., Ferrari, C., Angiuli, M., Yao, J., Raspi, C. & Bramanti, E. (2010). Qualitative and quantitative analysis of wood samples by Fourier transform infrared spectroscopy and multivariate analysis. *Carbohydrate Polymers*, *82*, 772–778.

Dadi, A. P., Varanasi, S. & Schall, C. A. (2006). Enhancement of cellulose saccharification kinetics using an ionic liquid pretreatment step. *Biotechnology and Bioengineering*, *95*, 904–910.

Dhanoa, M. S., Lister, S. J., Sanderson, R. & Barnes, R. J. (1994). The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *Journal of Near Infrared Spectroscopy*, *2*, 43–47.

FitzPatrick, M., Champagne, P., Cunningham, M. F. & Whitney, R. A. (2010). A biorefinery processing perspective: treatment of lignocellulosic materials for the production of value-added products. *Bioresource Technology*, *101*, 8915–8922.

Fort, D. A., Remsing, R. C., Swatloski, R. P., Moyna, P., Moyna, G. & Rogers, R. D. (2007). Can ionic liquids dissolve wood? Processing and analysis of lignocellulosic materials with 1-n-butyl-3-methylimidazolium chloride. *Green Chemistry*, *9*, 63–69.

Garrido Frenich, A., Martínez Galera, M., Martínez Vidal, J. L. & Gil García, M. D. (1996). Partial least-squares and principal component regression of multianalyte high-performance liquid chromatography with diode-array detection. *Journal of Chromatography A*, *727*, 27–38.

Ghasemi, J. & Niazi, A. (2001). Simultaneous determination of cobalt and nickel comparison of prediction ability of PCR and PLS using original, first and second derivative spectra. *Microchemical Journal*, *68*, 1–11.

Gorry, P. A. (1990). General least-squares smoothing and differentiation by the convolution (Savitzky–Golay) method. *Analytical Chemistry*, *62*, 570–573.

Hames, B. R., Thomas, S. R., Sluiter, A. D., Roth, C. J. & Templeton, D. W. (2003). Rapid biomass analysis. *Applied Biochemistry and Biotechnology*, *105*, 5–16.

Hatfield, R. & Fukushima, R. S. (2005). Can lignin be accurately measured? *Crop Science*, *45*, 832–839.

Heinze, T., Schwikal, K. & Barthel, S. (2005). Ionic Liquids as Reaction Medium in Cellulose Functionalization. *Macromolecular Bioscience*, *5*, 520–525.

Helland, I. S., Næs, T. & Isaksson, T. (1995). Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, *29*, 233–241.

Hurtubise, F. G. & Krassig, H. (1960). Classification of fine structural characteristics in cellulose by infared spectroscopy use of potassium bromide pellet technique. *Analytical Chemistry*, *32*, 177–181.

Joglekar, H., Rahman, I. & Kulkarni, B. (2007). The path ahead for ionic liquids. *Chemical Engineering and Technology*, *30*, 819–828.

Johansson, J., Pettersson, S. & Folestad, S. (2005). Characterization of different laser irradiation methods for quantitative Raman tablet assessment. *Journal of Pharmaceutical and Biomedical Analysis*, *39*, 510–516.

Kačuráková, M., Capek, P., Sasinková, V., Wellner, N. & Ebringerová, A. (2000). Spectroscopic study of plant cell wall model compounds: pectic polysaccharides and hemicelluloses. *Carbohydrate Polymers*, *43*, 195–203.

Kadam, S. S., van der Windt, E., Daudey, P. J. & Kramer, H. J. M. (2010). A comparative study of ATR-FTIR and FT-NIR spectroscopy for in-situ concentration monitoring during batch cooling crystallization processes. *Crystal Growth and Design*, *10*, 2629–2640.

Kiefer, J., Obert, K., Bösmann, A., Seeger, T., Wasserscheid, P. & Leipertz, A. (2008). Quantitative analysis of alpha-D-glucose in an ionic liquid by using infrared spectroscopy. *Chemphyschem: A European Journal of Chemical Physics and Physical Chemistry*, *9*, 1317–1322.

Kilpeläinen, I., Xie, H., King, A., Granstrom, M., Heikkinen, S. & Argyropoulos, D. S. (2007). Dissolution of wood in ionic liquids. *Journal of Agricultural and Food Chemistry*, *55*, 9142–9148.

Kjeldahl, K. & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal of Chemometrics*, *24*, 558–564.

Lateef, H., Grimes, S., Kewcharoenwong, P. & Feinberg, B. (2009). Separation and recovery of cellulose and lignin using ionic liquids: a process for recovery from paper-based waste. *Journal of Chemical Technology and Biotechnology*, *84*, 1818–1827.

Lee, S. H., Doherty, T. V., Linhardt, R. J. & Dordick, J. S. (2008). Ionic liquid-mediated selective extraction of lignin from wood leading to enhanced enzymatic cellulose hydrolysis. *Biotechnology and Bioengineering*, *102*, 1368–1376.

Moore, A. & Owen, N. (2001). Infrared spectroscopic studies of solid wood. *Applied Spectroscopic Reviews*, *36*, 65–86.

Nelson, M. L. & O'Connor, R. T. (1964a). Relation of certain infrared bands to cellulose crystallinity and crystal lattice type. Part II. A new infrared ratio for estimation of crystallinity in celluloses I and II. *Journal of Applied Polymer Science*, *8*, 1325–1341.

Nelson, M. L. & O'Connor, R. T. (1964b). Relation of certain infrared bands to cellulose crystallinity and crystal latticed type. Part I. Spectra of lattice types I, II, III and of amorphous cellulose. *Journal of Applied Polymer Science*, *8*, 1311–1324.

Oh, S. Y., Yoo, D. I., Shin, Y., Kim, H. C., Kim, H. Y., Chung, Y. S., et al. (2005). Crystalline structure analysis of cellulose treated with sodium hydroxide

and carbon dioxide by means of X-ray diffraction and FTIR spectroscopy. *Carbohydrate Research*, *340*, 2376–2391.

Pintar, A., Batista, J. & Levec, J. (2002). In situ Fourier transform infrared spectroscopy as an efficient tool for determination of reaction kinetics. *Analyst*, *127*, 1535–1540.

Pu, Y., Jiang, N. & Ragauskas, A. J. (2007). Ionic liquid as a green solvent for lignin. *Journal of Wood Chemistry and Technology*, *27*, 23–33.

Savitzky, A. & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*, 1627–1639.

Sievers, C., Valenzuela-Olarte, M. B., Marzialetti, T., Musin, I., Agrawal, P. K. & Jones, C. W. (2009). Ionic-liquid-phase hydrolysis of pine wood. *Industrial and Engineering Chemistry Research*, *48*, 1277–1286.

Sun, N., Rahman, M., Qin, Y., Maxim, M. L., Rodríguez, H. & Rogers, R. D. (2009). Complete dissolution and partial delignification of wood in the ionic liquid 1-ethyl-3-methylimidazolium acetate. *Greem Chemistry*, *11*, 646–655.

Swatloski, R. P., Spear, S. K., Holbrey, J. D. & Rogers, R. D. (2002). Dissolution of cellulose with ionic liquids. *Journal of the American Chemical Society*, *124*, 4974–4975.

Tucker, M. P., Nguyen, Q. A., Eddy, F. P., Kadam, K. L., Gedvilas, L. M. & Webb, J. D. (2001). Fourier transform infrared quantitative analysis of sugars and lignin in pretreated softwood solid residues. *Applied Biochemistry and Biotechnology*, *91*, 51–61.

Veale, E. L., Irudayaraj, J. & Demirci, A. (2007). An on-line approach to monitor ethanol fermentation using FTIR spectroscopy. *Biotechnology Progress*, *23*, 494–500.

Wu, J., Zhang, J., Zhang, H., He, J., Ren, Q. & Guo, M. (2004). Homogeneous acetylation of cellulose in a new ionic liquid. *Biomacromolecules*, *5*, 266–268.

Yang, Z. & Pan, W. (2005). Ionic liquids: green solvents for nonaqueous biocatalysis. *Enzyme and Microbial Technology*, *37*, 19–28.

Zhang, Y., Du, H., Qian, X. & Chen, E. Y. (2010). Ionic liquid-water mixtures: enhanced K_w for efficient cellulosic biomass conversion. *Energy and Fuels*, *4*, 2410–2417.

Zhao, H., Jones, C. L., Baker, G. A., Xia, S., Olubajo, O. & Person, V. N. (2009). Regenerating cellulose from ionic liquids for an accelerated enzymatic hydrolysis. *Journal of Biotechnology*, *139*, 47–54.