2950 Niles Road, St. Joseph, MI 49085-9659, USA
269.429.0300 fax 269.429.3852 hq@asabe.org www.asabe.org

An ASABE Meeting Presentation

Paper Number:
[Click here to enter paper number]

Partial Least Squares - Discriminant Analysis (PLS-DA) of *Miscanthus x giganteus* by FT-NIR Spectroscopy

Daniel A. Williams[1], Mary-Grace C. Danao[1,2,*], Marvin R. Paulsen[1], Kent D. Rausch[1], and Stefan Bauer[3]

[1]Dept. of Agricultural and Biological Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[2]Energy Biosciences Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[3]Energy Biosciences Institute, University of California, Berkeley, CA 94720
*Corresponding author: gdanao@illinois.edu

Written for presentation at the
2013 ASABE Annual International Meeting
Sponsored by ASABE
Kansas City, Missouri
July 21 – 24, 2013

Abstract. *The objectives of this research were to describe the variation in chemical composition of Miscanthus x giganteus and to probe the potential of using Fourier transform near infrared (FT-NIR) spectroscopy in quantitatively analyzing the composition of Miscanthus and qualitatively classifying Miscanthus. Large variations in glucan (40.7 ± 2.37%), xylan (20.6 ± 1.20%), arabinan (1.83 ± 0.36%), acetyl (2.84 ± 0.28%), lignin (20.7 ± 1.35%), ash (2.60 ± 1.64%), and extractives (5.59 ± 0.86%) content were observed for 67 samples used in the calibration set that were collected from Miscanthus bales stored under a variety of conditions (indoors, under roof, outdoors with tarp cover, and outdoors without tarp cover) for a period of 1 to 24 months after harvest and baling. The composition of samples used for validation and model testing were comparable. Partial least squares (PLS) regression models based on the FT-NIR spectra of core samples collected from bales can be used to predict glucan, xylan, lignin, and ash contents with RPD values of 4.86, 4.08, 3.74, and 1.71, respectively. These models were used with linear discriminant analysis to classify the samples based on their glucan, lignin, and ash contents. The best classification results were based on the PLS-DA lignin model, which classified the samples into three groups, with small variations with each group. While the models developed in this study were based on a small sample size (less than 100 for calibration) and the small size contributed to some of the inaccuracy and imprecision in the predictions, the approach demonstrated that FT-NIR spectra and PLS-DA modeling can be used to rapidly screen Miscanthus samples at different stages of the supply chain, including after long-term storage.*

Keywords. *Miscanthus x giganteus*, composition, FT-NIR spectroscopy, partial least squares regression, linear discriminant analysis.

Introduction

Miscanthus x giganteus is a woody rhizomatous C4 grass species which is perceived as a promising high yielding lignocellulosic material for both energy and fiber production (Jones and Walsh, 2001). It consists of three main components – cellulose, hemicellulose, and lignin, which account for nearly 40, 20, and 20% (w/w) respectively, while the balance is composed of organic acids, ash, and extractives (Sanderson et al., 1996). Cellulose, hemicellulose, and lignin are structural polysaccharides contained in the plant cell walls (Hatfield, 1989). The cellulose and hemicellulose fractions can be converted into energy and chemicals through direction combustion, pyrolysis, or biological conversion. Since only a fraction of the biomass can be converted into chemical energy, bioethanol yields per unit mass of biomass are directly proportional to the composition of the biomass, which can vary from the averages due to age, stage of growth, growth conditions, and other factors (Perez et al., 2002).

Decision-making during plant breeding, crop management, harvest, transportation, preprocessing (e.g., size reduction, densification) and storage conditions needs to be guided by biomass conversion requirements (Vidal et al., 2011). Cost, quality, and volume of feedstocks determine the viability of commercial-scale bioenergy production. Conversion facilities would like to receive feedstocks that are consistent, or uniform, in quality in moisture content, ash content, and convertible carbohydrates so they can operate their chemical pretreatment and conversion processes efficiently (Kenney and Ovard, 2013). Recently, Tao et al. (2013) demonstrated that the variability in corn stover component composition strongly impacts the variability of the minimum ethanol selling price (MSEP) due to the variability in ethanol yield from the structural carbohydrates in the stover feedstocks. Therefore, it is advantageous to know the composition prior to conversion so that enzyme mixtures, yeast strains, and process control parameters can be adjusted accordingly to maximize yields. Knowing the composition at earlier stages of the supply chain can also help in the development of quality-based valuations which incentivize farmers and suppliers to implement best management practices to ensure a uniform and consistent supply system (Kenney et al., 2013). For example, biochemical conversion processes are sensitive to carbohydrate of content as the ratio of C5 to C6 sugars and accessibility of these sugars are important in optimizing pretreatment and fermentation conditions (Öhgren et al., 2007; Berlin et al., 2007). Lignin content in biomass represents the recalcitrance of cell walls to saccharification, particularly during enzymatic hydrolysis (Öhgren et al., 2007; Chen and Dixon, 2007). Ash is important to control as it displaces valuable carbohydrates (i.e., when ash content increases, the convertible carbohydrates content in biomass decrease) and decreases pretreatment efficiency (Bakker and Elberson, 2005). In thermochemical conversion, ash components impair catalysts and contribute to slag formation within the combustion processes (James et al., 2012).

Current wet chemistry methods for chemical characterization of biomass feedstock are not applicable for field or inline monitoring because they are expensive, labor-intensive, and cannot provide compositional information in real time for process control (Ye et al., 2008). One approach to reducing the time and cost of compositional analysis is the development of near infrared (NIR) spectroscopy and a good calibration based on multivariate analysis to provide either a quantitative or qualitative measure of composition.

NIR spectroscopy is a type of vibrational spectroscopy that utilizes the optical region ranging from 4000 to 10000 cm^{-1} (780 to 2500 nm). The energy absorption in this region corresponds to combinations of the fundamental vibrational transitions along with overtones associated with each bond (Blanco and Villaroya, 2002). Depending on which atoms are interacting, different anharmonicities arise giving each compound a unique fingerprint (Theander and Aman, 1984). When the NIR spectra are calibrated to reference values, e.g., results from a wet chemistry assay, using multivariate data analysis such as principal components analysis (PCA), principal components regression (PCR), or partial least squares (PLS) regression. The resulting models can be used to predict or classify the composition of complex samples using techniques such as soft independent modeling of class analogy (SIMCA) or linear discriminant analysis (LDA).

Several studies have shown NIR spectroscopy as a promising technique to assessing biomass composition. Sanderson et al. (1996) demonstrated individual carbohydrates can be estimated in woody and herbaceous feedstocks such as straw, corn stover, poplar, etc. using standard normal variate-detrend (SNV-D) preprocessing to correct the scatter in the NIR spectra collected and regression by PLS. Hames et al. (2003) reported NIR calibration models for corn stover feedstock and dilute acid pretreated corn stover. Pordesimo et al. (2005) later used the corn stover feedstock model to investigate the variability of stover composition with crop maturity at harvest. They took samples from corn plans from approximately two weeks before the corn grain reached physiological maturity to approximately one month after the grain was at a moisture content suitable for harvesting. Their results showed large decreases in the extractives content of the samples, with increases in both xylan and lignin content. The corn stover feedstock model was also used by Hoskinson et al. (2007) to provide compositional data for a study investigating the variation in quality and quantity of corn stover available under different harvesting scenarios. PLS models of NIR spectra was used to evaluate compositional

variation and sources of variability in 508 commercial hybrid corn stover samples collected from 47 sites in eight Corn Belt states after the 2001, 2002, and 2003 harvests (Templeton et al., 2009). Similarly, Haffner et al. (2013) demonstrated the use of PLS regression models of NIR spectra of 241 *Miscanthus x giganteus* samples harvested from seven sites in Illinois for fast monitoring of Miscanthus in plant breeding studies.

In this study, PLS regression models were developed using the Fourier transform near infrared (FT-NIR) spectra of 101 *Miscanthus x giganteus* samples harvested from two sites in Illinois and stored under a variety of conditions (indoors, under roof, outdoors with tarp cover, and outdoors without tarp cover) for a period of 1 to 24 months after harvest and baling. The PLS regression models were developed to provide a numerical estimate of the composition (glucan, xylan, arabinan, acetyl, lignin, ash, and extractives content) of Miscanthus. A linear discriminant analysis (LDA) based on the PLS regression models for glucan, lignin, and ash was also performed to classify the feedstocks into groups based on their composition. These classes or grouping may be useful in rapid screening of biomass at any stage of the supply chain and as it enters a conversion facility.

Materials and Methods

Sample collection, preparation, and compositional analysis

Bale core samples were collected using a hay probe bale sampler (Part No. BHP550C, Best Harvest, St. Petersburg, FL) from stacked bales that have been stored in Urbana, Griggsville, and Taylorville, Illinois for a period of 3 to 24 months. The bales stored in Urbana, IL were harvested at the senescent stage (December to January) from the Energy Biosciences Institute (EBI) farm at the University of Illinois in Urbana-Champaign in 2008 to 2011. The bales stored in Griggsville and Taylorville were harvested at the senescent stage in Pana, IL in December 2008 to January 2009. The bales were stored under different conditions from indoors (Taylorville), under roof (Urbana), outdoors with a tarp (Urbana and Griggsville), and outdoors without a tarp (Urbana). Each sample, approximately 40 g, was a collection of multiple core samples from the bale. Additional information regarding the 101 samples used in this study is provided in the Appendix.

Immediately after collection, the samples were dried at 60°C for 72 hr according to ASABE Standard S358.2 (1998) (Figure 1). The dried samples were milled using a cutting mill (SM 2000, Retsch, Inc., Haan, Germany) fitted with a 2 mm sieve. The dry milled samples were bagged and stored at room temperature prior to scanning with an FT-NIR spectrophotometer and sending to the Energy Biosciences Institute (EBI) Analytical Chemistry Laboratory for compositional analysis (glucan, xylan, arabinan, acetyl, lignin, ash, and extractives content). Compositional analysis was conducted in duplicates following standard procedures developed by the National Renewable Energy Laboratory (NREL) and is discussed in Haffner et al. (2013).

Scanning, preprocessing, and analyses of FT-NIR spectra

An FT-NIR spectrophotometer (Spectrum^{TM} One NTS, Perkin Elmer, Waltham, MA) was used to scan the dry milled samples (dry basis moisture contents were less than 2%). Approximately a 5 g subsample was poured in a near infrared reflectance accessory (NIRA) cup, leveled with a spatula, and placed in an automatic spinner (Figure 1). The spectrophotometer was set to collect an average of 32 scans from 4000 to 10000 cm^{-1} at a spectral resolution of 4 cm^{-1}.

Unscrambler® (Version 10.1, Camo Software Inc., Woodbridge, NJ) was used to preprocess and analyze the spectral data. The data were mean centered and either preprocessed using multiplicative scatter correction (MSC); Savitzy-Golay (SG) first derivative filtering using a second, third, or fourth order polynomial; or combination of MSC and SG. MSC was used to remove multiplicative scatter or interferences resulting from baseline shifts and the sample's particle size distribution. The SG derivative algorithm does smoothing and differentiation of the data by simplified least squares, where each point (i.e., the absorbance at a specific wavenumber) became the weighted average derivative of the points surrounding it.

PLS-DA modeling

PLS regression models for each component were developed in Unscrambler® using 67 samples for calibration, 24 samples for validation, and 10 samples for testing. The PLS models were compared and evaluated on the number of factors used and corresponding explained variance, coefficient of determination (R^2), root mean square error (RMSE), and ratio of performance to deviation (RPD).

The resulting PLS regression models for glucan, lignin, and ash were further used to develop qualitative discriminant analysis models that may be used for screening Miscanthus samples based on these components. The method used in discriminant analysis was linear discriminant analysis (LDA), which explicitly attempts to model the difference between the classes of data. The resulting PLS-DA models were evaluated on the

accuracy of the predicted classification and the uniformity of sample composition within each group. The means of each component across groups were compared using a Tukey's test to identify any difference between two means that is greater than the expected standard error. The Tukey's test was conducted using *R* (Version 2.15.2, 2012).

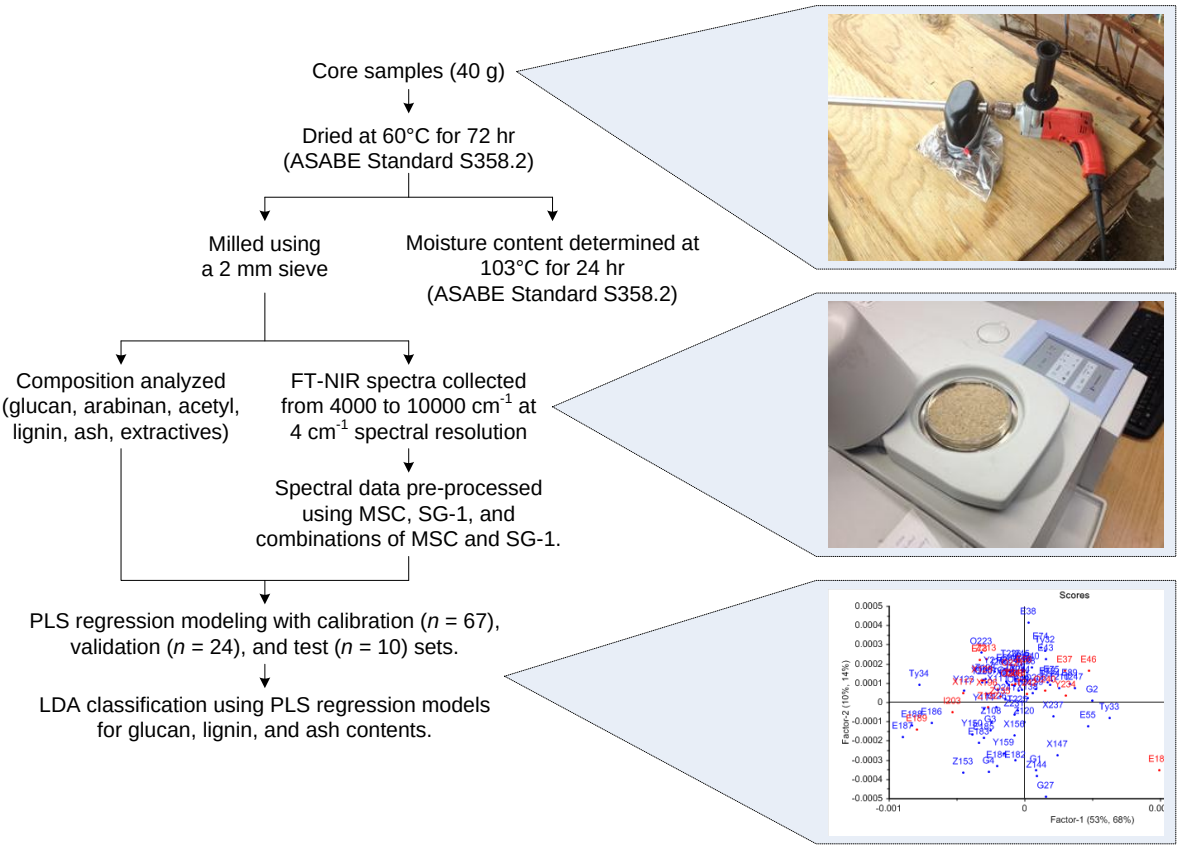

Figure 1. Preparation of Miscanthus samples for FT-NIR spectra collection and PLS-DA model development.

Results and Discussion

Compositional analysis

There was a large variation in composition of *Miscanthus x giganteus* samples from having been grown in multiple locations and growing seasons and stored under a variety of conditions for 3 to 24 months. A summary of the compositional analysis for the calibration, validation, and test sets are shown in Table 1. The calibration set was used in developing the PLS regression and PLS-DA classification models. The validation set was used in evaluating the prediction performance of the PLS regression (instrumental in the RPD calculation) and PLS-DA classification models. The test set is a small set of samples used to assess model performance separate from its development and initial evaluation.

Overall, the composition values of Miscanthus from the stored bales were comparable those reported by Haffner et al. (2013), who used Miscanthus samples that were manually cut above ground at either pre-senescent or at senescent stages of growth. In contrast, samples from this study came from mechanically harvested Miscanthus, cut at the senescent stage, baled, and stored. Haffner et al. (2013) reported glucan contents ranging from 36.3 to 45.3% whereas glucan contents found in this study has a wider range, from 25.8 to 44.0%. The lower ranges found with stored bales was attributable to dry matter loss that occur during storage. Xylan, arabinan, acetyl, and ash contents reported from the two studies were comparable. Lignin contents in the stored bales was as high as 26.5% while the highest lignin content reported by Haffner et al. (2013) was 23.2%. This difference was likely due to Haffner et al. (2013) focusing on pre-senescent or senescent Miscanthus while this study focused on stored baled samples. Additionally, the extractives contents in the stored bales were only as high as 9.10% -- as extractives tended to be lost with storage time (Wiselogel et al., 1996) -- whereas higher extractives, up to 12.2%, were found by Haffner et al. (2013).

Table 1. Minimum (min), maximum (max), mean and standard deviation (S.D.) of the composition of *Miscanthus x giganteus* samples.

	Calibration set (*n* = 67)				Validation set (*n* = 24)				Test set (*n* = 10)			
	Min (%)	Max (%)	Mean (%)	S.D. (%)	Min (%)	Max (%)	Mean (%)	S.D. (%)	Min (%)	Max (%)	Mean (%)	S.D. (%)
Glucan	26.5	44.0	40.7	2.37	25.8	43.6	39.8	3.70	36.9	43.7	40.5	2.49
Xylan	17.7	24.2	20.6	1.20	19.3	23.8	20.7	1.02	18.1	24.2	21.6	2.34
Arabinan	1.05	3.11	1.83	0.36	1.20	2.71	1.85	0.34	1.21	2.75	1.99	0.50
Acetyl	2.06	3.40	2.74	0.28	1.76	3.29	2.81	0.38	2.09	3.01	2.61	0.34
Klason lignin	17.7	26.5	20.7	1.35	18.2	25.7	20.5	1.61	17.5	22.2	20.1	1.48
Ash after extraction	0.59	5.16	2.60	1.64	0.59	4.47	2.53	0.91	0.59	4.25	2.26	1.46
Extractives	3.61	8.72	5.59	0.86	3.59	9.10	5.66	1.16	4.64	6.60	5.61	0.57

PLS regression modeling

PLS regression models using MSC, SG first and second derivatives, and combinations of MSC and SG second derivative models were developed to predict the glucan, xylan, arabinan, acetyl, lignin, ash, and extractives contents of Miscanthus. Since the classification will be based on glucan, lignin, and ash only, their PLS regression models, along with the PLS models for xylan, are presented in Table 2 while the models for the other components are included in the Appendix.

Table 2. PLS regression models for glucan, xylan, lignin, and ash contents of *Miscanthus x giganteus*.

	Glucan	Xylan	Lignin	Ash
Spectral data preprocessing	MSC+Savitsky-Golay (1,20,20,3)[1]	Savitsky-Golay (1,20,20,3)+MSC	MSC+Savitsky-Golay (1,65,65,4)	MSC+Savitsky-Golay (1,65,65,4)
No. of factors used in the model	2	5	4	7
R^2				
Calibration	0.93	0.87	0.87	0.84
Validation	0.95	0.86	0.93	0.64
Test	0.52	0.60	0.70	0.85
RMSE[2] (%)				
Calibration	0.60	0.38	0.39	0.33
Validation	0.76	0.25	0.43	0.53
Test	1.62	1.39	0.75	0.53
Bias[3]	0.10	-0.02	0.06	-0.20
RPD[4]	4.86	4.08	3.74	1.71
R/SEP[5]	22.9	17.4	16.6	7.76
Pearson's kurtosis of the calibration set	19.2	10.6	10.6	1.21

[1]Savitzky-Golay derivative parameters (derivative order, left points, right points, polynomial order)

[2]Root mean square error, $RMSE = \sqrt{\sum_{i=1}^{N}(y_i - \hat{y})^2/N}$, where y_i is the individual reference value (from wet chemistry assay), $\hat{y}$ is the NIR predicted value from the PLS regression model, and N is the total number of samples.

[3]$Bias = \sum_{i=1}^{N}(\hat{y} - y_i)/(N-1)$

[4]Ratio of performance to deviation, $RPD = \sigma_y/SEP$, where the standard deviation of the reference data is $\sigma_y = \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})/(N-1)}$, $\bar{y}$ is the average reference value, and the standard error of prediction or prediction (using the validation set) is $\sqrt{\sum_{i=1}^{N}(y_i - \hat{y} - Bias)/N}$.

[5]Range of the calibration data set to the standard error of performance, SEP.

The PLS regression models were assessed using the AACCI Guidelines for Model Development and Maintenance (AACCI Method 39-00, 1999). The standard specifies two different performance targets – RPD and R/SEP. When RPD for a regression model is greater than 2.5, the model is deemed suitable for screening samples; RPD values greater than 5 mean the models are acceptable for calibration for quality control; and RPD values greater than 8 are good for process control, development, and applied research. The resulting RPD models for glucan and lignin indicate they are at least suitable for screening samples. The PLS model developed for ash, despite having a calibration R^2 of 0.84 between predicted and the reference values, had an RPD of 1.71. Since ash is inorganic and does not absorb radiation in the near infrared region, any calibration model developed will be an indirect measure of the association of minerals with the organic compounds in the biomass (Clark et al., 1985).

The resulting R/SEP values, on the hand, suggested these models were suitable for good calibration for quantification of glucan, lignin, and ash. Per AACI Method 39-00, R/SEP values greater than 4 mean the model is suitable for screening; greater than 10, model may be used for calibration for quality control; and greater than 15 notes the model may be used for good calibration for quantification. While the PLS models for glucan, lignin, and ash had relatively low bias and relatively high R^2 for calibration (i.e., greater than 0.80), the calibration data sets had high kurtosis values ($K > 1$) which meant small standard deviations and high ranges led an overestimate of the R/SEP values. Overall, these PLS regression models can be improved by increasing the number of samples in the calibration and having kurtosis values near zero.

PLS-DA classification modeling

The models were originally grouped into their respective classes by, first, looking for natural breaks in the histograms and, secondly, considering the range of the main portion of the data set for each component. Since there were few natural breaks, the groups were largely based on the range.

Based on the histograms of the calibration sets, the maximum bin width was set at 2% and a minimum of three bins were used. For glucan, range was 8 percentage points, which led to 4 groupings of the samples (Figure 2) – Group G1 (less than 38%), Group G2 (38 to 40%), Group G3 (40 to 42%), and Group G4 (greater than 42%). Similarly, samples were split into three groups based on their lignin content (Group L1, less than 20%; Group L2, 20-22%; and Group L3, greater than 22%) and ash content (Group A1, less than 1.75%; Group A2, 1.75 to 3.25%; and Group A3, greater than 3.25%). Care was taken so the means of each group were different from each other (Table 3).

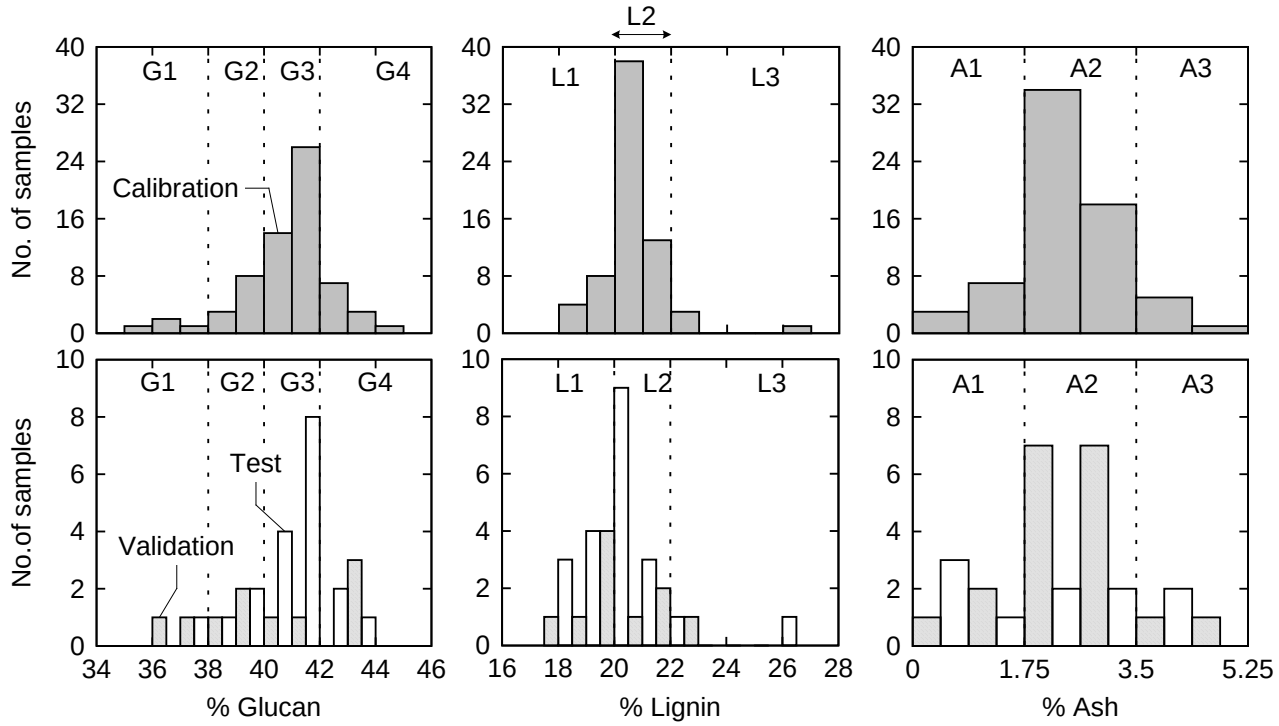

Figure 2. Classification of *Miscanthus x giganteus* samples based on glucan, lignin, and ash contents determined by wet chemistry assays.

Accuracy of the PLS-DA classification models

During classification, three samples were identified as outliers – Gr1 (from a bale stored outdoors with tarp cover for 24 months), E44 (from a bale stored outdoors without a tarp cover for 12 months), and E91 (from a bale stored under roof for 6 months). Using the PLS regression model for glucan for classification, 29 of 98 samples were misclassified into a neighboring group. The overall predictability of the model was 70%. A comparison of means showed that each predicted group, G1 to G4, was different from each other, whereas G3 and G4 were not. However each group was not different from the actual classification groups.

Table 3. Classification of samples based on FT-NIR spectra and PLS-DA models for glucan, lignin, and ash contents.

		Actual classification by reference (wet chemistry) values											
		PLS-DA Glucan Model				PLS-DA Lignin Model				PLS-DA			
		G1	G2	G3	G4		L1	L2	L3		A1	A2	A3
Predicted classfication by FT-NIR spectra	G1	**7**	1	0	0	L1	**19**	3	0	A1	**12**	1	0
	G2	2	**10**	8	0	L2	6	**54**	1	A2	5	**56**	4
	G3	0	5	**40**	5	L3	0	9	**6**	A3	0	8	**11**
	G4	0	1	7	**11**								
Means of actual classification groups (%)		34.5 aA	39.2 bB	41.2 cC	43.0 dD		19.1 aA	20.8 bB	23.4 cC		0.97 aA	2.47 bB	3.9 cC
Means of predicted classification groups (%)		34.9 aA	39.6 bB	41.2 cC	42.4 dC		19.4 aA	20.6 bB	22.2 cC		1.06 aA	2.40 bB	3.57 cC

[1]Values followed by the same lowercase letter in the same row, per PLS-DA model, are not significantly different ($p < 0.05$). Means values of actual and predicted classification followed by the same uppercase letter in the same column are not significantly different.

The classification and prediction results from the PLSR-DA lignin model showed that most of the samples were classified as Group L2. The model was able to classify 79 of 98 samples correctly. Samples with high lignin content tended to be classified correctly. The PLS-DA lignin model was able to classify 19 out of 25 samples correctly into Group L1; 54 out of 66 samples correctly into Group L2; and 6 out of 7 samples correctly in Group L3. When samples were misclassified, they fell into a neighboring group, i.e., no sample from Group L1 was misclassified into Group L3 and vice versa.

Similarly, the PLS-DA model for ash was able to classify 80 out of 98 samples correctly. Of the 18 samples that were misclassified they only fell one group away from their assigned group. This should be expected since the ash samples cover a continuous range but the samples are being classified into fixed groups. The PLS-DA ash model was able to classify 12 of 17 samples into Group A1; 56 of 65 samples into Group A2; and 11 of 15 into Group A3. A comparison of the means was run and all group means were significantly different from each other.

Uniformity of the samples after classification

Taking the predicted classes resulting from the PLS-DA modeling, the effect of this grouping on the other six components was evaluated. The box plots for each component showed the range and quartiles, with outliers designated by solid (laying 1.5 to 3 interquartile range distances from the mean) and hollow (laying more than 3 interquartile range distances from the mean) data points.

The PLS-DA glucan model would be most useful if it can also deliver a uniform feedstock within each grouping (Figure 3). In terms of xylan and arabinan content, the means for Groups G1 and G2 were not different from each other, similarly as the means for Groups G3 and G4. All four groups, however, did not differ in lignin and ash content. With the PLS-DA lignin model, while there were no difference in ash and extractives contents across Groups L1, L2, and L3, there were significant differences across the groups in terms of glucan and lignin contents and the variation within groups is small (Figure 4). The box plots for glucan, xylan, and lignin were tighter than the other groups for group 2 which took 63 of the 98 samples representing a relatively consistent sample while group 3 was wider, but was still relatively consistent. Even though the samples were classified into distinct groups when the PLS-DA for ash was applied, across groups, there was little variation in the other components (Figure 5). Group A2 had the least variation within its group. Based on these results, the best classification scheme was based on the PLS-DA lignin model, which classified the samples into three groups, with small variations with each group.

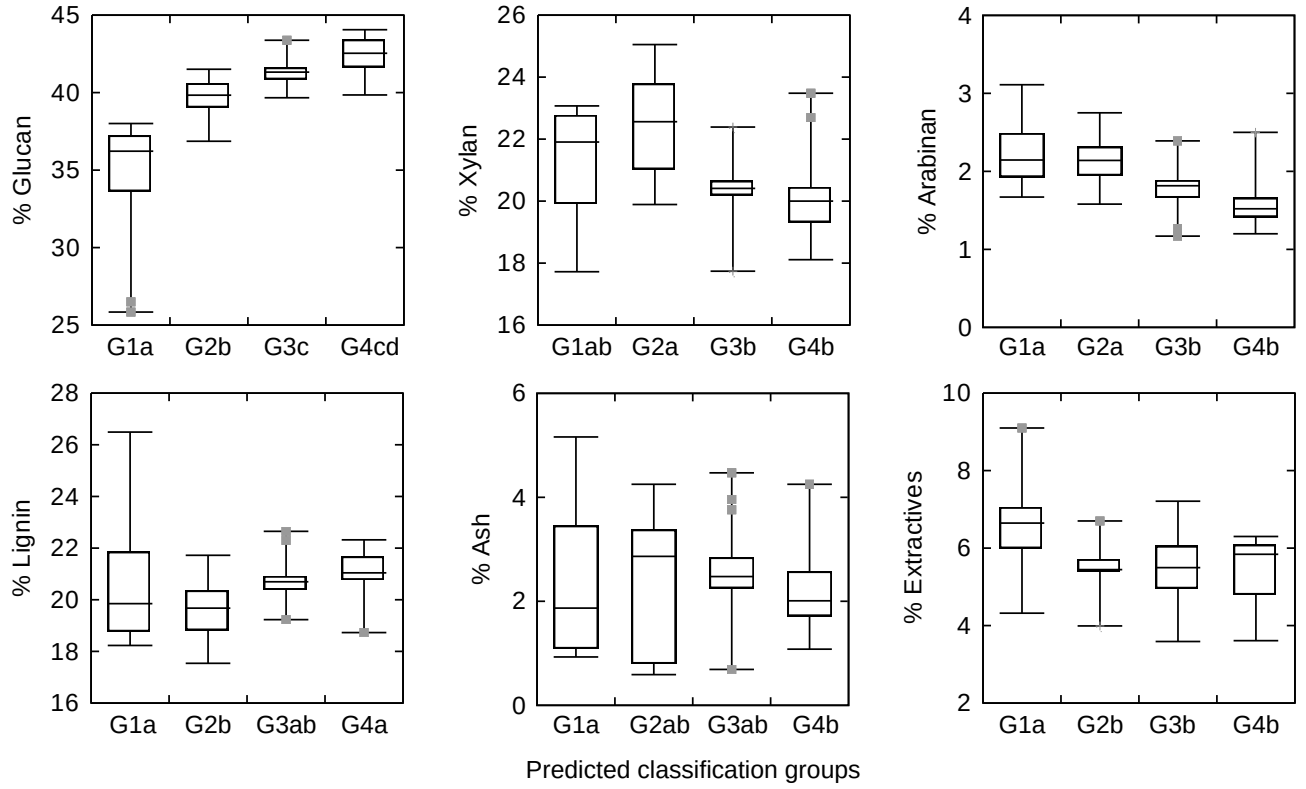

Figure 3. Variability in composition across groups as classified by the PLS-DA glucan model. Groups with the same letters, per component, are not significantly different (p = 0.05). Median-based box plots represent the minimum, maximum, interquartile range (IQR), outliers (■, which are defined as data lying at 1.5·IQR distance from the median), and extremes (+, which are defined as data lying at 3·IQR distance from the median).

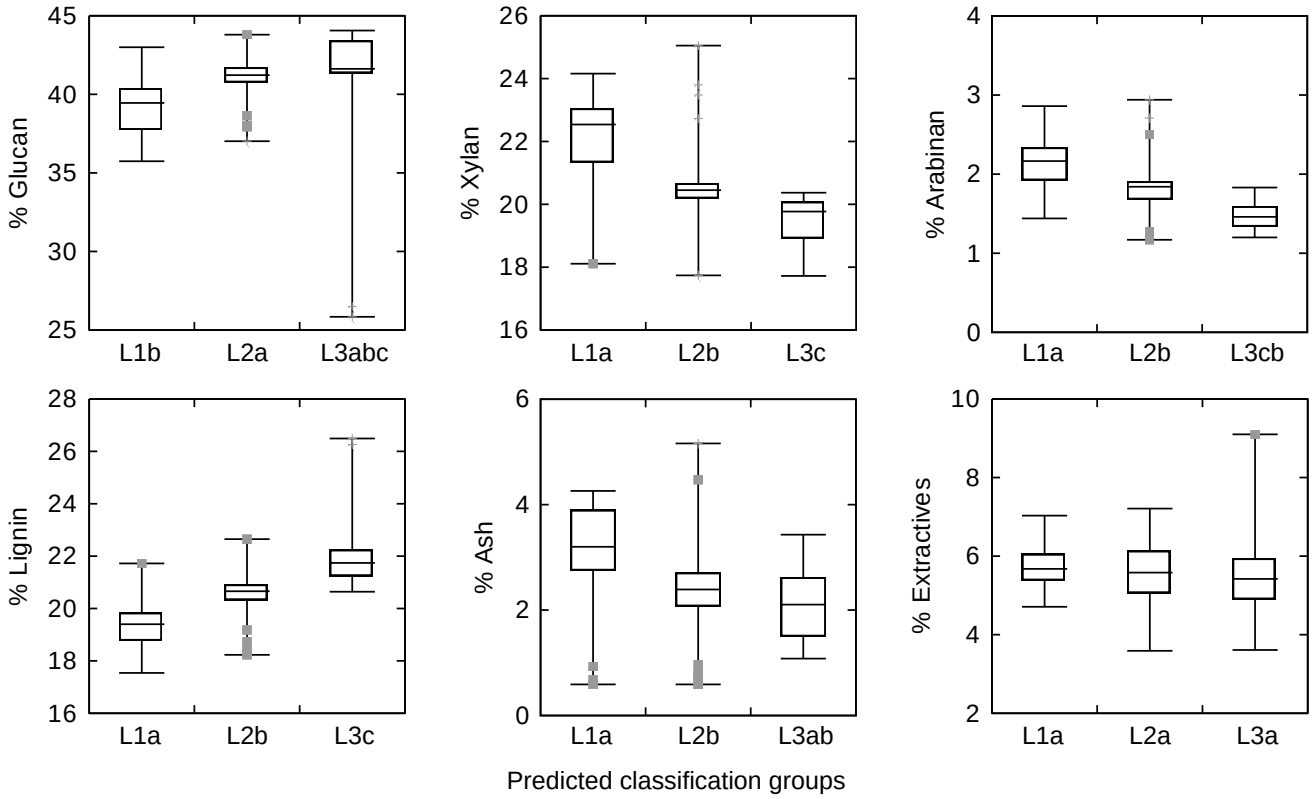

Figure 4. Variability in composition across groups as classified by the PLS-DA lignin model. Groups with the same letters, per component, are not significantly different (p = 0.05).

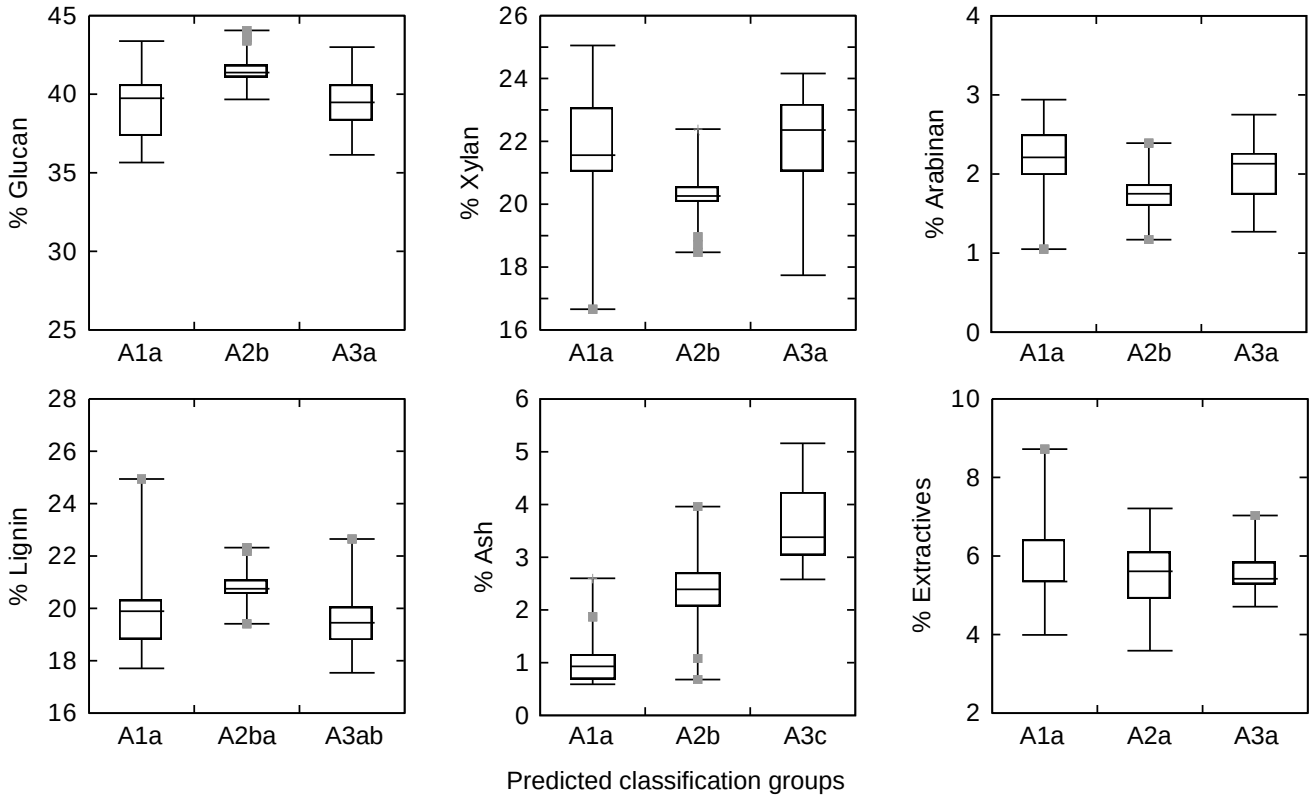

Figure 5. Variability in composition across groups as classified by the PLS-DA ash model. Groups with the same letters, per component, are not significantly different (p = 0.05).

Tao et al. (2013) evaluated the minimum ethanol selling price (MESP) compared to corn stover composition and determined that the carybohydrates content (%) can have drastic effects on the conversion cost and final selling point of the ethanol. The corn stover used in their analysis ranged in total carbohydrates from 53 to 64%, resulting in an MESP range of $2.05 to $2.50 per gallon. The main carbohydrates in Miscanthus are cellulose (glucan) and the hemicellulose (arabino-xylan) and, when these components and summed up and plotted across groups, there were no differences across Groups G1 to G3 with G4 having the lowest polysaccharides content (Figure 6). Assuming that the conversion process is identical to that used by Tao et al. (2013) in their analysis, these data suggest Miscanthus as a promising alternative to corn stover since the majority of the samples had total carbohydrates contents greater than 60%. In addition to ash not varying greatly between groups when the samples were classified using the PLS-DA models for glucan and lignin, overall, ash content after extraction were lower than those for corn stover samples (Templeton et al., 2009).

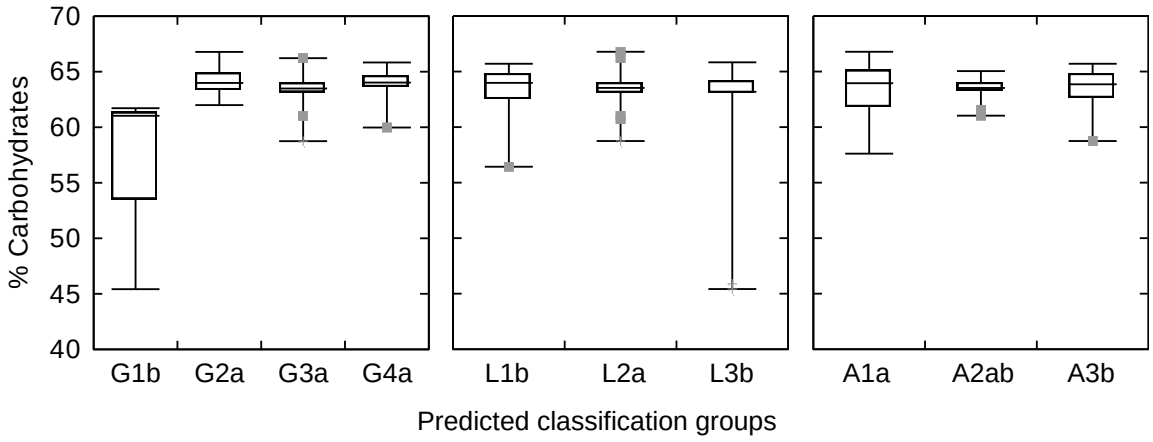

Figure 6. Carbohydrates (glucan + xylan + arabinan) content of the groups as classified by the PLS-DA models for glucan, lignin, and arabinan. Groups with the same letters, per PLS-DA model classification, are not significantly different (p = 0.05).

Conclusion

These data indicate that FT-NIR spectra can be used to predict the chemical composition of *Miscanthus x giganteus* and that the resulting PLS regression models can be used to predict glucan, xylan, and lignin with RPD values of 4.86, 4.08, and 3.74, respectively, making them suitable for screening Miscanthus samples. A PLS regression model for ash was also developed with a calibration R^2 = 0.84, RMSE = 0.33%, and RPD value of 1.71. The models were used with LDA to classify the samples based on their glucan, lignin, and ash contents. The best classification results were based on the PLS-DA lignin model, which classified the samples into three groups, with small variations with each group. While the models developed in this study were based on a small sample size (less than 100 for calibration) and the small size contributed to some of the inaccuracy and imprecision in the predictions, the approach demonstrated that FT-NIR spectra and PLS-DA modeling can be used to rapidly screen Miscanthus samples at different stages of the supply chain, including after long-term storage.

Acknowledgements

This work was funded by the Energy Biosciences Institute through the program titled, "Engineering Solutions for Biomass Feedstock Production." The authors would like to thank Xiangwei Chen, Joshua Jochem, Gary Letterly, and Tim Mies for their technical assistance.

References

AACCI International. 1999. Gudelines for model development and maintenance. Method 39-00. The American Association of Cereal Chemists, 10th Edition. The Association: St. Paul, MN.

ASABE. 2008. Moisture measurement – forages. Standard S358.2. ASABE: St. Joseph, MI.

Bakker, R.R. and H.W. Elberson. 2005. Managing ash content and quality in herbaceous biomass: an analysis from plant to product. In *Proceedings of the 14th European Biomass Conference.* pp. 17-21.

Berlin, A., V. Maximenko, N. Gilkes, and J. Saddler. 2007. Optimization of enzyme complexes for lignocellulose hydrolysis. *Biotechnology and Bioengineering* 97: 287-296.

Blanco, M. and I. Villarroya. 2002. NIR spectroscopy: a rapid-response analytical tool. *Trends in Analytical Chemistry* 21: 240-50-PII S0165-9936(02)00404-1.

Chen, F. and R.A. Dixon. 2007. Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology* 25: 759-761.

Clark, D.A., R.P. Adams, R.C. Lamb, and M.J. Andrews. 1985. Near infrared analysis of hydrocarbon producing plant species. *Biomass* 8: 1-11.

Haffner, F.B., V.D. Mitchell, R.A. Arundale, and S. Bauer. 2013. Compositional analysis of *Miscanthus giganteus* by near infrared spectroscopy. *Cellulose*. DOI 10.1007/s10570-013-9935-1.

Hames, B.R., S.R. Thomas, A.D. Sluiter, C.J. Roth, and D.W. Templeton. 2003. Rapid biomass analysis – new tools for compositional analysis of corn stover feedstocks and process intermediates from ethanol production. *Applied Biochemistry and Biotechnology* 105: 5-16.

Hatfield, R.D. 1989. Structural polysaccharides in forages and their degradability. *Agronomy Journal* 46: 39-46.

Hoskinson, R., D. Karlen, S. Birrell, C. Radtke, and W. Wilhelm. 2007. Engineering, nutrient removal, and feedstock conversion evaluations of four corn stover harvest scenarios. *Biomass and Bioenergy* 31: 126-136.

James, A.K., R.W. Thring, S. Helle, and H.S. Ghuman. 2012. Ash management review – applications of biomass bottom ash. *Energies* 5: 3856-3873.

Jones, M.B. and M. Walsh. 2001. *Miscanthus for energy and fibre.* London: James and James, pp. 35-37.

Kenney, K.L. and L.P. Ovard. 2013. Advanced feedstocks for advanced biofuels: transforming biomass to feedstocks. *Biofuels* 4: 1-3.

Kenney, K.L., W.A. Smith, G.L. Gresham, and T.L. Westover. 2013. Understanding biomass feedstock variability. *Biofuels* 4: 111-127.

Öhgren, K., R. Bura, J. Saddler, and G. Zacchi. 2007. Effect of hemicellulose and lignin removal on enzymatic hydrolysis of steam pretreated corn stover. *Bioresource Technology* 98: 2503-2510.

Perez, J., J. Muñoz-Dorado, T. de la Rubia, and J. Martinez. 2002. Biodegradation and biological treatments of cellulose, hemicellulose, and lignin: an overview. *International Microbiology* 5: 53-63.

Pordesimo, L.O., B.R. Hames, S. Sokhansanj, and W.C. Edens. 2005. Variation in corn stover composition and energy content with crop maturity. *Biomass and Bioenergy* 28: 366-374.

R Development Core Team. 2012. *R*: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Coputing. Retrieved from http://www.R-project.org.

Sanderson, M.A., F.A. Agblevor, M. Collins, and D.K. Johnson. 1996. Compositional analysis of biomass feedstocks by near infrared reflectance spectroscopy. *Biomass and Bioenergy* 11: 365-370.

Tao, L., D.W. Templeton, D. Humbird, and A. Aden. 2013. Effect of corn stover compositional variability on minimum ethanol selling price (MSEP). *Bioresource Technology*, http://dx.doi.org/10.1016/j.biortech.2013.04.083.

Templeton, D.W., A.D. Sluiter, T.K. Hayward, B.R. Hames, and S.R. Thomas. 2009. Assessing corn stover composition and sources of variability via NIRS. *Cellulose* 16: 621-639.

Theander, O. and P. Aman. 1984. Anatomical and chemical characteristics. In *Straw and Other Fibrous By-Products as Feed*. Amsterdam: Elsevier, pp. 45-78.

Unscrambler. 2011. Camo Software, Inc., Woodbridge, NJ.

Vidal B.C., B.S. Dien, K.C. Ting, V. Singh. 2011. Influence of feedstock particle size on lignocellulose conversion – a review. *Applied Biochemistry and Biotechnology* 164: 1405-1421.

Wiselogel, A.E., F.A. Agblevor, D.K. Johnson, S. Deutsch, J.A. Fennell, and M.A. Sanderson. 1996. Compositional changes during storage of large round switchgrass bales. *Bioresource Technology* 56: 103-109.

Ye, X.P., L. Liu, D. Hayes, A. Womac, K. Hong, and S. Sokhansanj. 2008. Fast classification and compositional analysis of corn stover fractions using Fourier transform near-infrared techniques. *Bioresource Technology* 99: 7323-7332.

Appendix

Miscanthus x giganteus bales and sample collection

Samples used in this study were collected from stacked bales that have been stored in Urbana, Griggsville, and Taylorville, Illinois for a period of 3 to 24 months (Table A1 and Figure A1). The bales stored in Urbana, IL were harvested at the senescent stage (December to January) from the Energy Biosciences Institute (EBI) farm at the University of Illinois in Urbana-Champaign in 2008 to 2011. The bales stored in Griggsville and Taylorville were harvested at the senescent stage in Pana, IL in December 2008 to January 2009.

Table A1. Sources of *Miscanthus x giganteus* samples used in this study.

Group No.	Storage			Sample Names	Count
	Location	Conditions	Period (months)		
1	Taylorville, IL	Indoors	24	Ty32, Ty33, Ty34	3
2	Griggsville, IL	Outdoors, with tarp[2]	24	Gr1, Gr2, Gr3, Gr4, Gr10, Gr14, Gr16, Gr21	9
3	Urbana, IL	Under roof[3]	24	E185, E186, E187, E188, E189, E190	6
4	Urbana, IL	Under roof	12	E37, E38, E39, E40, E52, E53, E55, E56, E59	9
5	Urbana, IL	Outdoors, without a tarp	12	E43, E44, E45, E46, E73, E74, E75, E76, E79, E272	10
6	Urbana, IL	Under roof	6	E89, E90, E91, E92, E93	5
7	Urbana, IL	Outdoors, without a tarp	6	E180, E181, E182, E183, E184	5
8	Urbana, IL	Outdoors, without a tarp	6	Z108, Z120, Z132, Z135, Z144, Z153, Z198, Z213, Z231	9
9	Urbana, IL	Outdoors, with tarp	6	Y114, Y123, Y126, Y141, Y150, Y159, Y200, Y216, Y234	9
10	Urbana, IL	Under roof	6	X111, X117, X129, X138, 147, X156, X196, X219, X237	9
11	Urbana, IL	Outdoors, without a tarp	3	O206, O207, O208, O221, O222, O223, O239, O240, O241	9
12	Urbana, IL	Outdoors, with tarp	3	T209, T210, T211, T224, T225, T226, T242, T243, T244	9
13	Urbana, IL	Under roof	3	I203, I204, I205, I227, I228, I229, I245, I246, I247	9

(a) Bale stack in Griggsville, IL. Bales were covered with a tarp for 12 months and ripped tarp for the next 12 months.

(b) Bale stack in Taylorville, IL. Bales were stored indoors for a period of 24 months.

(c) Bale stacks in the Energy Bioscience Institute (EBI) Farm in Urbana, IL. From left to right, bales were stored under roof, outdoors without tarp cover; and outdoors with tarp cover.

Figure A1. Sources of Miscanthus x giganteus core samples used in this study.

PLS regression modeling

PLS regression models using MSC, SG first derivative, and combinations of MSC and SG were developed to predict the arabinan, extractives, and acetyl contents of *Miscanthus x giganteus*. However, the resulting PLS models for extractives and acetyl had low RPD values and were not useful for prediction.

Table A.2. PLS regression models for arabinan, extractives, and acetyl contents of *Miscanthus x giganteus*.

	Arabinan	Extractives	Acetyl
Spectral data preprocessing	Savitsky-Golay[1] (1,50,1,4) + MSC	MSC+Savitsky-Golay (1,65,65,4)	Savitsky-Golay (1,40,40,4) + MSC
No. of factors	4	1	2
Explained variance (%), calibration			
R^2			
Calibration	0.85	0.04	0.47
Validation	0.81	N/A	0.16
Test	0.69	N/A	N/A
RMSE[2] (%)			
Calibration	0.13	0.90	0.20
Validation	0.14	1.04	0.33
Test	0.26	N/A	N/A
Bias[3]	-0.04	-0.002	0.06
RPD[4]	2.42	N/A[6]	1.15[6]
R/SEP[5]	11.61	N/A	4.25

[1]Savitzky-Golay derivative parameters (derivative order, left points, right points, polynomial order)

[2]Root mean square error, $RMSE = \sqrt{\sum_{i=1}^{N}(y_i - \hat{y})^2/N}$, where y_i is the individual reference value (from wet chemistry assay), $\hat{y}$ is the NIR predicted value from the PLS regression model, and N is the total number of samples.

[3]$Bias = \sum_{i=1}^{N}(\hat{y} - y_i)/(N-1)$.

[4]Ratio of performance to deviation, $RPD = \sigma_y/SEP$, where the standard deviation of the reference data is $\sigma_y = \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})/(N-1)}$, $\bar{y}$ is the average reference value, and the standard error of prediction or prediction (using the validation set) is $\sqrt{\sum_{i=1}^{N}(y_i - \hat{y} - Bias)/N}$.

[5]Range of the calibration data set to the standard error of performance, SEP.

[6]Both models for extractives and acetyl contents were deemed poor and not useful for prediction.