Multivariate Calibration Models for Sorghum Composition using Near-Infrared Spectroscopy

E. Wolfrum and C. Payne
National Renewable Energy Laboratory

T. Stefaniak and W. Rooney
Texas A&M University

N. Dighe
Monsanto

B. Bean
Texas Agrilife Research and Extension

J. Dahlberg
Kearney Research and Extension Center

Multivariate Calibration Models for Sorghum Composition using Near-Infrared Spectroscopy

E. Wolfrum and C. Payne
National Renewable Energy Laboratory

T. Stefaniak and W. Rooney
Texas A&M University

N. Dighe
Monsanto/Texas A&M University

B. Bean
Texas Agrilife Research and Extension

J. Dahlberg
Kearney Research and Extension Center

NREL is a national laboratory of the U.S. Department of Energy, Office of Energy Efficiency & Renewable Energy, operated by the Alliance for Sustainable Energy, LLC.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, Colorado 80401
303-275-3000 • www.nrel.gov

Abstract

We have developed calibration models based on near-infrared (NIR) spectroscopy coupled with multivariate statistics to predict compositional properties relevant to cellulosic biofuels production for a variety of sorghum cultivars. A robust calibration population was developed in an iterative fashion. The quality of models developed using the same sample geometry on two different types of NIR spectrometers and two different sample geometries on the same spectrometer did not vary greatly.

List of Acronyms

FT-NIR   Fourier transform near-infrared

HPLC   high performance liquid chromatography

NIR   near-infrared

PC   principal component

PLS   partial least squares

RMSECV   root mean standard error of cross-validation

sd   standard deviation

SNV   standard normal variate

UV/VIS   ultraviolet-visible spectroscopy

Table of Contents

1 Introduction

There is great interest in the development of sustainable biofuels to displace petroleum and other fossil fuels, and investigations into a variety of dedicated bioenergy feedstocks are underway [1-6]. Sorghum [*Sorghum* bicolor (L.) Moench] is one such dedicated bioenergy feedstock. Recent publications discuss the compositional variety and agronomic traits of different sorghum cultivars [7-12]. Rapid compositional analysis methods based on near-infrared (NIR) reflectance spectroscopy combined with multivariate statistics are well-established and widely used in agriculture [13-15]. Rapid compositional analysis methods have been developed for a number of different potential bioenergy feedstocks [16-18]. The goal of this work was to develop an NIR calibration model for sorghum as a rapid analysis tool for sorghum researchers.

2 Materials and Methods

2.1 Sample Selection

Sorghum samples were taken from various breeding and agronomic trials conducted by the Texas Agrilife Research Sorghum Breeding program. The samples were collected from tests located in several different locations in Texas and a detailed description of the samples and their compositional variety has recently been published [9, 10]. Once all the samples were collected, they were dried in a forced air drier at 45°C. Samples were then knife-milled to pass through a 2-mm screen and stored at ambient temperature in paper bags.

2.2 Biomass Analysis

Samples were analyzed using standard methods for summative biomass compositional analysis. A detailed review of the history of these methods and typical analytical uncertainties are available [19, 20]. Mass closures were close to 100%, with no systematic variation by sorghum type. A detailed discussion of the results of these analyses is presented elsewhere [9, 10]. The samples were analyzed in groups of five to eight over the course of several months. Table 1 shows the summary compositional analysis results for the sorghum samples.

One modification to the standard analytical methods described above was the determination of starch in the samples. We performed starch analysis on all samples as received using a standard assay procedure with HPLC rather than UV/VIS detection of the resulting glucose (Megazyme). We refer to this as-received value as the whole starch content. Any sample containing more than 2% whole starch was also analyzed after water/ethanol extraction to determine the remaining amount of structural starch. This was necessary because the analytical hydrolysis procedure described above, used to break down cellulose to glucose, will also hydrolyze any structural starch that is present to glucose. In the absence of starch, all glucose is reported as structural glucan. With starch present, the structural glucan value must be corrected for structural starch. Of the 155 samples, 113 had non-zero amounts of whole starch, and 64 had non-zero amounts of structural starch.

2.3 Near-Infrared Spectroscopy

Two types of NIR spectrometers were used in this work: a grating-monochrometer type (XDS, Foss North America, Eden Prairie, MN) and a Fourier transform (FT) type (Antaris, Thermo-

Fisher, Waltham, MA). The spectral range for the grating-monochrometer unit (hereafter referred to as the NIR unit) was 400–2,500 nm, although only the region of 1,100–2,500 nm was used for calibration model development. The spectral range of the Fourier transform unit (hereafter referred to as the FT-NIR unit) was 3,700–12,000 cm^{-1}, although only the region of 4,000–9,000 cm^{-1} was used for calibration model development. The spectral regions of the two instruments used for modeling are quite similar, and conversion between wavelength and wavenumber is straightforward (cm^{-1}=10^7/nm).

Knife-milled samples were placed in quartz sample cups and scanned in reflectance mode on both the NIR and FT-NIR instruments. Samples were also scanned in reflectance mode on the FT-NIR instrument using an automated sampling carousel and borosilicate sampling vials (Thermo-Fisher p/n 03-339-26C). All samples were scanned in either duplicate or triplicate. Replicate scans were averaged prior to building calibration models. Quality-control ("check") materials were scanned along with the experimental samples to ensure instrument stability; no anomalies were seen with the check scans.

2.4 Multivariate Analysis

Multivariate analyses were performed in Unscrambler 10.1 (Camo Inc., www.camo.com). Some data manipulation and statistical analyses were performed in R (R Project for Statistical Computing, www.r-project.org). The quality of a calibration model was assessed (and different calibration models were compared) based on the correlation coefficient R and the root mean standard error of cross-validation (RMSECV). Correlation coefficients were first converted to z-values using the Fisher z-transform, and then differences were compared using a pooled standard error $\sigma = \sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}$ to the critical z-value of 1.96. Because the RMSECV values are the square root of a variance measure (in this case the variance of predicted and calibration values), the ratios of the squares of these values are compared to a critical F-value.

Two types of partial least squares (PLS) calibration models were developed: grouped PLS-1 and PLS-2. The critical difference between the two types is that the grouped PLS-1 algorithm creates a calibration equation for each variable (in this case, for each constituent in sorghum) separately. The PLS-2 algorithm creates single calibration equations for the variable set at once; all constituent models are solved simultaneously.

3 Results and Discussion

3.1 Sample Selection

Multivariate calibration models are secondary analytical techniques in that they require primary analytical data for calibration. Thus, a robust multivariate calibration model requires a sufficient number of representative samples with primary analytical data. Sample selection for the calibration models discussed here proceeded in an iterative manner. The first group of sorghum samples analyzed and included in the model included only commercially-available cultivars. Preliminary calibration models were developed using these samples and were used to predict later populations that included more experimental germplasm. Samples from these later populations were selected for wet chemical analysis for a number of reasons: they represented unique agronomic properties, they were poorly predicted by the preliminary model, or they had

NIR spectra that were distinct from the original model population. Table 1 shows the summary statistics for the compositional analysis of the 155 sorghum samples used in this work.

3.2 Calibration Model Results

The average spectrum of the entire calibration set collected using each instrument assembly is shown in Figure 1. Table 2 shows the summary calibration statistics for PLS-2 models developed for each instrument assembly for the constituent glucan. We found that calibration models improved when the active wavelength ranges of the spectra were truncated to 4,000–9,000 cm^{-1} for the FT-NIR system and 1,100–2,500 nm for the NIR system. These are essentially the same ranges because 4,000–9,000 cm^{-1} is equivalent to 1,111–2,500 nm. Using derivatization and smoothing improved the models as well; we saw a broad maximum in model quality with respect to the size of the window for Savitsky-Golay derivative smoothing. The 21-point window was used to produce the models discussed here; this value was a compromise that provided smooth regression coefficients for each constituent as well as optimal RMSECV values. Although the RMSECV value for the NIR assembly is slightly higher than for the FT-NIR assemblies, statistical analysis of the data in Table 2 showed no differences in RMSECV or R^2 values for any of the models; the cross validation results suggest that all models are equivalent.
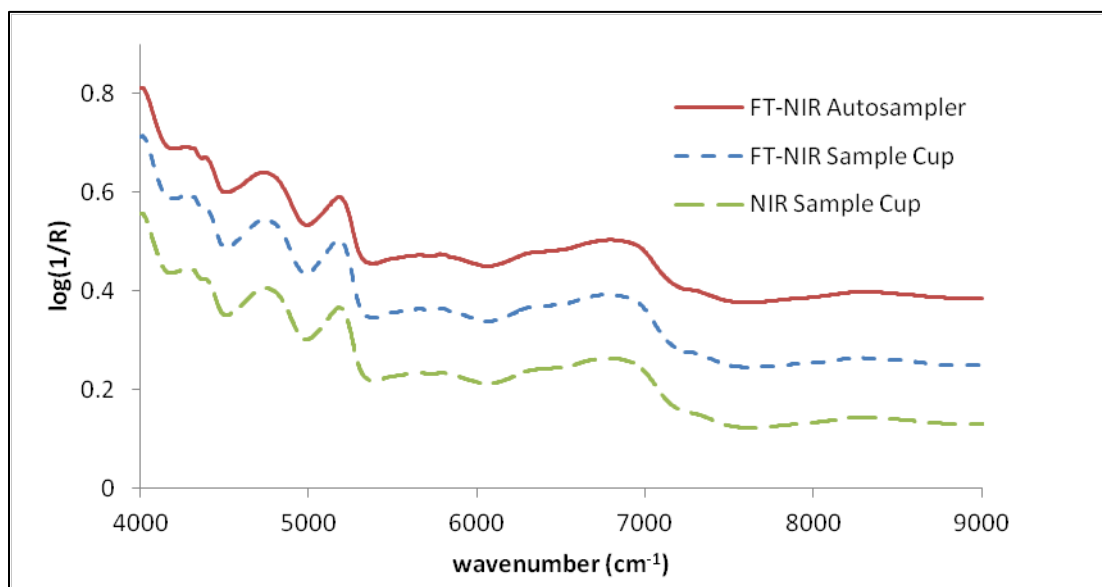

Figure 1. Average spectra taken with each spectrometer assembly (NIR sample cup, FT-NIR sample cup, FT-NIR autosampler). The NIR spectra abscissa for the diffraction-based instrument (NIR sample cup) was converted from wavelength to wavenumber for the plots

Table 1. Summary Statistics for the 155 Sorghum Samples Used to Develop Calibration Models
All constituent values are in dry weight percent. The samples were taken from a variety of agronomic field trials conducted at Texas Agrilife Research Stations.

Constituent	Mean	Max	Min	SD
Glucan	27.8	38.2	17.6	5.7
Xylan	16.0	21.2	10.8	2.6
Lignin	13.2	20.6	8.9	2.9
Starch	9.0	30.9	0.0	8.0
Extractives	27.2	49.9	13.3	8.7
Ash	6.8	11.3	2.3	2.0

Table 2. Summary Statistics for PLS-2 Calibration Models for Glucan Content in Sorghum Using Three Different Sets of NIR Spectra
Spectral ranges are expressed in wavenumbers (cm^{-1}) for the FT-NIR instrument and in wavelength (nm) for the diffraction NIR unit; the spectral ranges are essentially equivalent. The "Combined FT-NIR" model used both the sample cup and the autosampler spectra.

Instrument/Model	Spectral Range	RMSECV	R^2	PCs (#)	Samples (#)
NIR Sample Cup	1,100–2,500	1.48	0.93	7	130
FT-NIR Sample Cup	4,000–9,000	1.24	0.95	6	134
FT-NIR Autosampler	4,000–9,000	1.25	0.95	5	121
Combined FT-NIR	4,000–9,000	1.52	0.93	7	261

Table 3. Summary Statistics for PLS-2 and PLS-1 Calibration Models for the Major Constituents (Glucan, Xylan, Lignin, Starch, Total Extractives, and Ash) of Sorghum Using FT-NIR Sample Cup Spectra
Spectra underwent pretreatment prior to model-building. The PLS-2 model used six principal components (PCs) and had 134 samples. We saw no significant differences between PLS-1 and PLS-2 model statistics (p=0.05). Reference method uncertainty values are expressed as twice the standard deviation (sd) of replicate measurements and are taken from [20] except for starch, which is estimated as twice the sd of replicate measurement (unpublished results).

Constituent	*PLS-2 Model* RMSECV	R^2	*PLS-1 Model* RMSECV	R^2	PCs (#)	Samples (#)	Reference Method Uncertainty
Glucan	1.24	0.95	1.45	0.94	5	147	1.0
Xylan	0.85	0.89	0.81	0.90	4	147	0.6
Lignin	0.86	0.91	0.82	0.92	3	151	0.4
Starch	2.33	0.91	2.43	0.90	5	140	0.6
Extractives	2.29	0.91	2.33	0.93	5	146	1.2
Ash	0.88	0.81	0.78	0.85	6	148	0.4

Summary statistics for the grouped PLS-1 and PLS-2 models for the FT-NIR sample cup spectra are shown in Table 3 along with the uncertainties of the primary analytical methods [20], and plots of the predicted vs. measured constituent values for the constituents in the PLS-2 calibration model are shown in Figure 2. There are no significant differences between the grouped PLS-1 and the PLS-2 models; both algorithms provide equivalent results. The PLS-1 models require fewer principal components, but a separate model is required for each constituent.
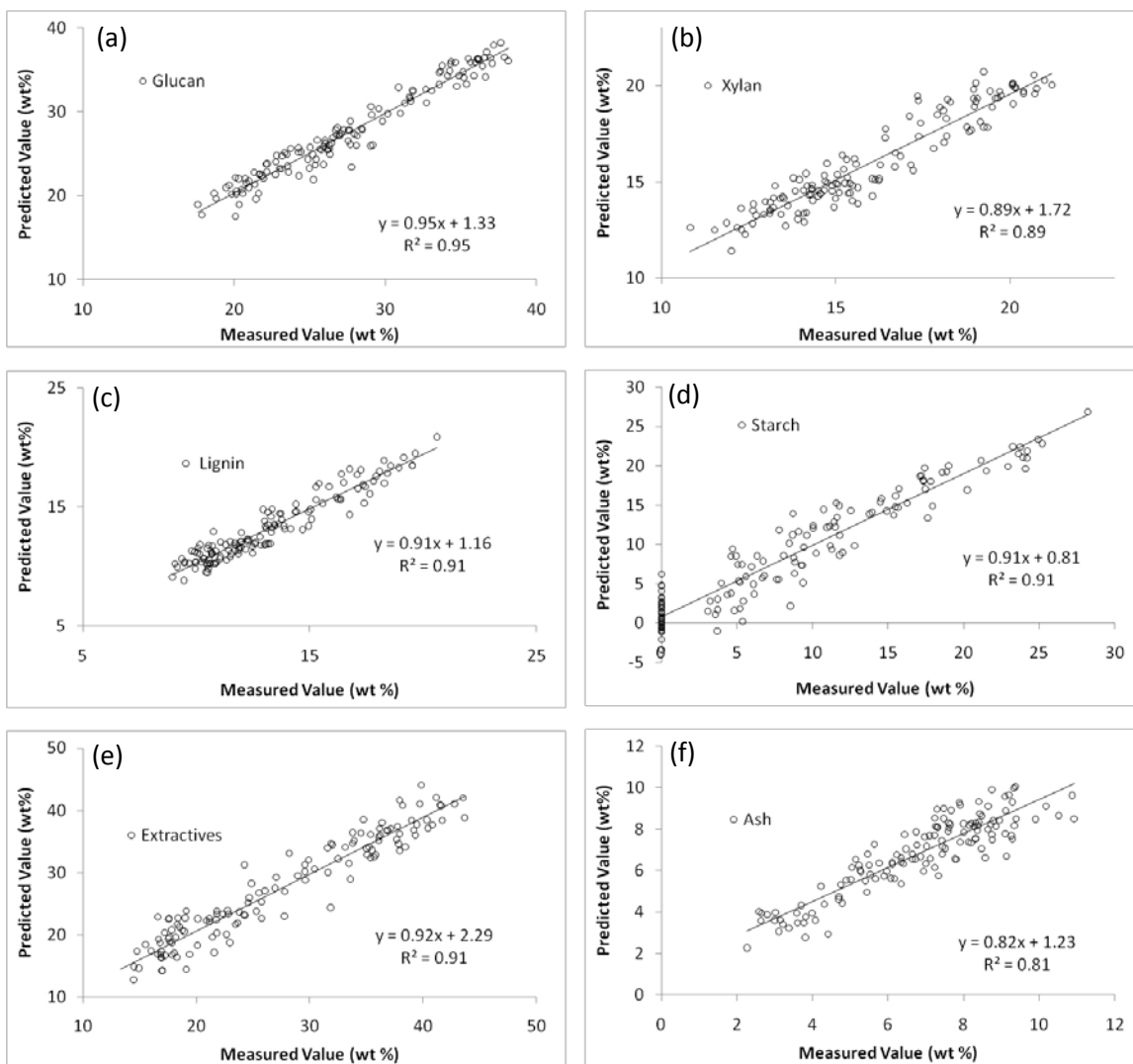

Figure 2. Predicted vs. measured composition for FT-NIR quartz cell calibration model: (a) glucan, (b) xylan, (c) lignin, (d) starch, (e) extractives, (f) ash. Some constituent predictions exhibit greater uncertainty than others.

5

As the data in Table 2 indicate, there is little difference between the calibration models using FT-NIR sample cup spectra and FT-NIR autosampler vial spectra. Each model could predict samples from the other model successfully, although there was approximately a two-fold higher uncertainty associated with the predictions (data not shown). A calibration model combining both the sample cup and autosampler vial spectra showed results similar to the individual models, although the RMSECV value is higher (p=0.05). These results suggest that with proper spectral pretreatment, a hybrid model consisting of samples taken with different sampling geometries may be possible. This is useful for two reasons. First, the borosilicate autosampler is a commodity item, available in large quantity very inexpensively. Second, samples can be stored in the autosampler vials immediately after milling, so the spectra of a given sample can be collected without additional handling of the sample. This minimizes operator effort and also reduces the opportunity for sample spillage or loss.

4 Conclusions

Robust multivariate calibration models using NIR spectroscopy coupled with chemometrics can be used as a rapid analysis tool for determining sorghum composition relevant for biofuels production. Sample selection is critical to building a robust model; multivariate calibration models must contain samples similar to those to be predicted. Standard spectral pretreatment methods reduce the number of principal components required for a model. Models developed on different types of NIR spectrometers, and with different sampling geometries, provided essentially equivalent results.

We intend to continue development of the calibration models presented here. Integration of new samples (particularly those with interesting agronomic or genetic traits) as well as samples poorly predicted by the current models will increase the robustness of the model. With proper laboratory quality control processes in place to ensure the stability of instrument response over time, new samples can routinely be added as they are identified and undergo wet chemical analysis.

References

1. Parrish, D.J.; Fike, J.H. "The Biology and Agronomy of Switchgrass for Biofuels." *Critical Reviews in Plant Sciences* (24), 2005; pp. 423-459.

2. David, K.; Ragauskas, A.J. "Switchgrass as an Energy Crop for Biofuel Production: A Review of its Ligno-Cellulosic Chemical Properties." *Energy & Environmental Science* (3:9), 2010; pp. 1182-1190.

3. Clifton-Brown, J.C.; Chiang, Y.-C.; Hodkinson, T.R. "Miscanthus: Genetic Resources and Breeding Potential to Enhance Bioenergy Production," Chapter 10. Vermerris, W., ed. *Genetic Improvement of Bioenergy Crops*. New York: Springer, 2008; pp. 295-308.

4. Yuan, J.S.; Tiller, K.H.; Al-Ahmad, H.; Stewart, N.R.; Stewart, C.N. Jr. "Plants to Power: Bioenergy to Fuel the Future." *Trends in Plant Science* (13:8), 2008; pp. 421-429.

5. Harper, R.J.; Sochacki, S.J.; Smettem, K.R.J.; Robinson, N. "Bioenergy Feedstock Potential from Short-Rotation Woody Crops in a Dryland Environment." *Energy & Fuels* (24:1), 2009; pp. 225-231.

6. Hoskinson, R.L.; Karlen, D.L.; Birrell, S.J.; Radtkea, C.W.; Wilhelm, W.W. "Engineering, Nutrient Removal, and Feedstock Conversion Evaluations of Four Corn Stover Harvest Scenarios." *Biomass and Bioenergy* (31:2-3), 2007; pp. 126-136.

7. Rooney, W.L.; Blumenthal, J.; Bean, B.; Mullet, J.E. "Designing Sorghum as a Dedicated Bioenergy Feedstock." *Biofuels, Bioproducts and Biorefining* (1:2), 2007; pp. 147-157.

8. Rooney, W.L. "Sorghum Improvement – Integrating Traditional and New Technology to Produce Improved Genotypes," Chapter 2. Sparks, D., ed. *Advances in Agronomy*. Vol. 83, Academic Press, 2004; pp. 37-109.

9. Dahlberg, J.; Wolfrum, E.; Bean, B.; Rooney, W.L. "Compositional and Agronomic Evaluation of Sorghum Biomass as a Potential Feedstock for Renewable Fuels." *Journal of Biobased Materials and Bioenergy* (5:4), 2011; pp. 507-513.

10. Stefaniak, T.R., et al. "Variation in Biomass Composition Components in Sorghums Grouped by End Use." *Biomass and Bioenergy*, submitted, 2011.

11. Murray, S.C.; Sharma, A.; Rooney, W.L.; Klein, P.E.; Mullet, J.E.; Mitchell, S.E.; Kresovich, S. "Genetic Improvement of Sorghum as a Biofuel Feedstock: I. QTL for Stem Sugar and Grain Nonstructural Carbohydrates." *Crop Science* (48:6), 2008; pp. 2165-2179.

12. Murray, S.C.; Rooney, W.L.; Mitchell, S.E.; Sharma, A.; Klein, P.E.; Mullet, J.E.; Kresovich, S. "Genetic Improvement of Sorghum as a Biofuel Feedstock: II. QTL for Stem and Leaf Structural Carbohydrates." *Crop Science* (48:6), 2008; pp. 2180-2193.

13. Shenk, J.S.; Workman, J.J. Jr.; Westerhaus, M.O. "Application of NIR Spectroscopy to Agricultural Products," Chapter 16. Burns, D.A. and Ciurczak, E.W., eds. *Handbook of Near-Infrared Analysis*. 2nd Edition, New York: Marcel Dekker, 2001.

14. Roberts, C.A., Workman, J.J., and Reeves, J.B.I., eds. *Near-Infrared Spectroscopy in Agriculture (Agronomy)*. Madison, WI: American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, 2004.

15. Williams, P. and Norris, K., eds. *Near-Infrared Technology in the Agricultural and Food Industries*. 2nd edition, St. Paul, MN: American Association of Cereal Chemists, 2001.

16. Hames, B.R.; Thomas, S.R.; Sluiter, A.D.; Roth, C.J.; Templeton, D.W. "Rapid Biomass Analysis." *Applied Biochemistry and Biotechnology* (105:1), 2003; pp. 5-16.

17. Wolfrum, E.; Sluiter, A. "Improved Multivariate Calibration Models for Corn Stover Feedstock and Dilute-Acid Pretreated Corn Stover." *Cellulose* (16:4), 2009; pp. 567-576.

18. Sanderson, M.A.; Agblevor, F.; Collins, M.; Johnson, D.K. "Compositional Analysis of Biomass Feedstocks by Near Infrared Reflectance Spectroscopy." *Biomass and Bioenergy* (11:5), 1996; pp. 365-370.

19. Sluiter, J.B.; Ruiz, R.O.; Scarlata, C.J.; Sluiter, A.D.; Templeton, D.W. "Compositional Analysis of Lignocellulosic Feedstocks. 1. Review and Description of Methods." *Journal of Agricultural and Food Chemistry* (58:16), 2010; pp. 9043-9053.

20. Templeton, D.W.; Scarlata, C.J.; Sluiter, J.B.; Wolfrum, E.J. "Compositional Analysis of Lignocellulosic Feedstocks. 2. Method Uncertainties." *Journal of Agricultural and Food Chemistry* (58:16), 2010; pp. 9054-9062.