Journal of Cereal Science 50 (2009) 381–387

Contents lists available at ScienceDirect

Journal of Cereal Science

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

Prediction of maize seed attributes using a rapid single kernel near infrared instrument

Jasper G. Tallada [a,*], Natalia Palacios-Rojas [b], Paul R. Armstrong [a]

[a] *Engineering and Wind Erosion Research Unit, USDA ARS, Grain Marketing and Production Research Center (GMPRC), 1515 College Ave, Manhattan, KS 66502, USA*
[b] *International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico*

ARTICLE INFO

Article history:
Received 13 February 2009
Received in revised form
29 May 2009
Accepted 10 August 2009

Keywords:
Corn
NIR
Partial least squares regression
PLSR

ABSTRACT

Non-destructive measurements of seed attributes would significantly enhance breeder selection of seeds with specific traits, and could potentially improve hybrid development. A single kernel near infrared reflectance (NIR) instrument was developed for rapidly predicting maize grain attributes, which would enable plant breeders to quickly select promising individual seeds. With the overall goal being to develop spectrometric calibrations, absorbance spectra from 904 to 1685 nm were collected from 87 maize samples, with 30 kernels of each sample (2610 kernels total), representing a wide variability in the essential amino acids tryptophan and lysine, crude protein, oil and soluble sugar contents. Average sample spectra were matched to bulk reference values. Partial least squares regression (PLSR) calibration models with cross-validation were developed for both relative (% dry matter) and absolute (mg kernel^{-1}) constituent contents. Similarly, models using bagging PLSR were developed. The best model obtained was for relative crude protein content, with an R^2p of 0.75 and a SEP of 0.47%. Kernel mass was also highly predictable (R^2p=0.76, SEP=0.03 g). Tryptophan, lysine and oil were less predictable, but showed good potential for segregating individual seeds using NIR. Soluble sugar contents produced poor model statistics. Bagging PLSR yielded models with similar levels of prediction.

Published by Elsevier Ltd.

1. Introduction

Maize breeding programs evaluate and select numerous genetic traits while focusing on a few important desirable traits. While analytical reference methods exist that can accurately quantify the composition of bulk seeds, they are often destructive and require a fairly large amount of material for analysis. Because there can be significant variability in composition between seeds, even those coming from the same breeding line, a non-destructive technique that can sort individual kernels using various compositional traits would be very useful for breeding research.

Screening methods, such as near infrared reflectance spectroscopy (NIRS), are currently used for rapid non-destructive measurement of bulk constituent contents for a number of crops such as wheat, maize, sorghum and soybeans. An early study by Orman and Schumann (1991) compared NIR calibration methods for predicting protein, oil and starch contents in both whole and ground maize samples. They worked within a spectral range of 1100–2500 nm for reflectance and 680–1235 nm for transmittance modes. While the best models were obtained from ground grain reflectance data, they suggested that the transmittance mode for whole grains might be more useful because of its greater speed of analysis. Single kernel research by Cogdill et al. (2004) explored the transmittance mode with hyperspectral imaging and obtained a working prediction for moisture. However, they suggested that further work was needed to obtain an acceptable prediction for oil content. Similarly, Baye et al. (2006) explored near infrared transmittance and reflectance spectroscopy methods to predict protein, oil and starch contents from individual kernels of maize of several different genotypes. They found that the transmittance mode was not suitable for predicting kernel composition, whereas the reflectance mode gave good predictive power when the absolute amount of constituents per kernel was predicted. Calibrations for maize vitreousness and dry matter degradability using NIRS in the

Abbreviations: CP, Crude protein; CV, Coefficient of variation; LYS, Lysine; MC, Moisture content; MSC, Multiplicative scatter correction; Nc, Number of samples in calibration set; Nf, Number of factors; NIR, Near infrared; Np, Numbers of samples in validation set; OIL, Oil contents; PLSR, Partial least squares regression; R^2cv, Coefficient of determination for calibration with cross-validation; R^2p, Coefficient of determination for validation; RMSE, Root mean square of error; RPD, Ratio of standard error to standard deviation; SD, Standard deviation; SEC, Standard error of calibration; SECV, Standard error of calibration with cross-validation; SEP, Standard error of prediction; SNV, Standard normal variate; SSC, Soluble solids contents; TRP, Tryptophan.

* Corresponding author. Tel.: +1 785 776 2728; fax: +1 785 537 5550.
E-mail address: jasper.tallada@ars.usda.gov (J.G. Tallada).

0733-5210/$ – see front matter Published by Elsevier Ltd.
doi:10.1016/j.jcs.2009.08.003

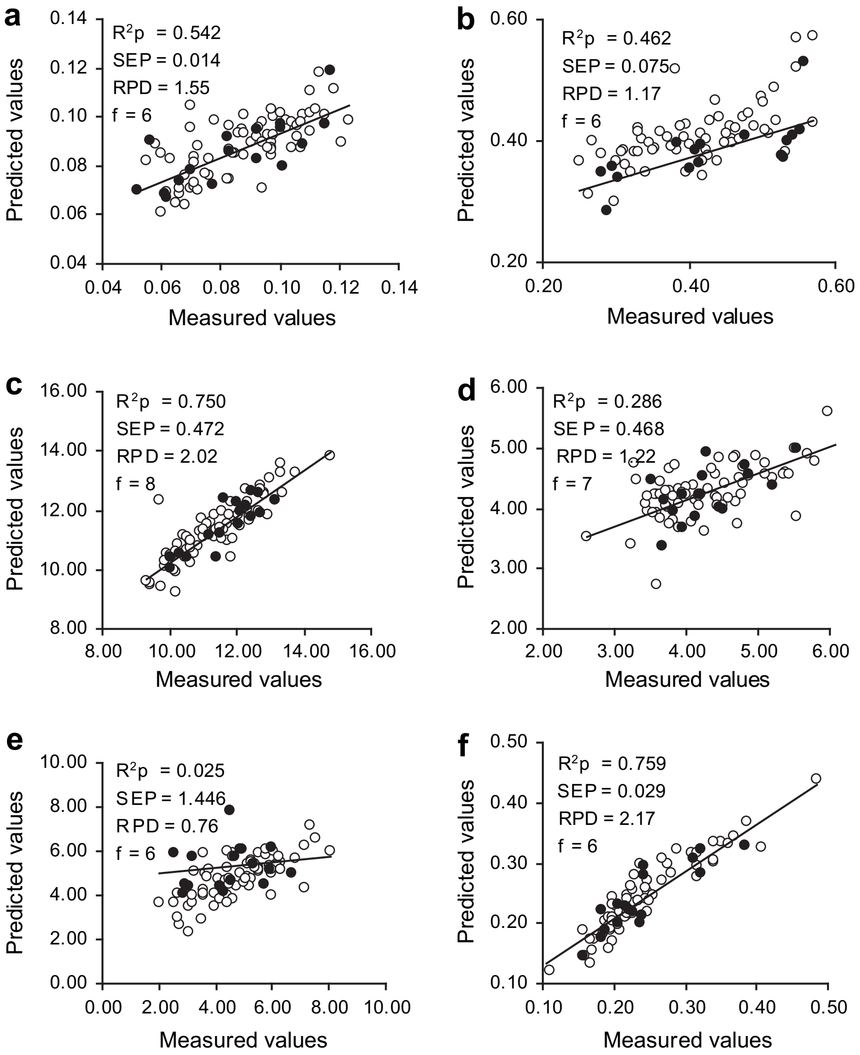

Fig. 1. Prediction plots for the relative contents of (a) tryptophan, (b) lysine, (c) crude protein, (d) oil, (e) soluble sugar; and (f) kernel mass; ○-samples from calibration set; ● – samples from prediction set.

400–2498 nm range were investigated by Ngonyamo-Majee et al. (2008). They concluded that NIRS can be used as a screening tool to develop maize hybrids in large-scale breeding programs. Armstrong (2006) developed a rapid single kernel NIR sorting instrument for maize and soybean. Prediction models for moisture of both seed types, and protein contents for soybeans were developed utilizing a spectrometric range from 906 to 1683 nm. Janni (2007) and Janni et al. (2008) described other patented NIR methods for seed analysis.

PLS regression models are commonly based on cross-validation statistics to determine the factor levels for a model derived from a single set of data. Bagging or "bootstrap aggregating" techniques aim to reduce the variance of predictors by re-sampling the training dataset with replacement into a number of learning sets, and calculating a calibration model for each set (Viscarra Rossel, 2007a). This has been shown to be useful for unstable regression models and neural networks through the aggregation of a number of versions of predictors (by averaging) over different model versions (Breiman, 1996). Wehrens et al. (2000) discussed that cross-validation, aside from being the most used method in chemometrics, may be prone to large variability of prediction, particularly when a small number of samples is used. In these cases, bootstrapping techniques would be most useful since parameters are computed from their central tendency estimates. For this reason, both methods of model development were investigated and compared herein.

The main objective of this study was to develop single kernel constituent calibration models from reference bulk analyses of tryptophan, lysine, crude protein, oil and soluble sugar contents for maize from a wide range of breeding lines. Furthermore, we aimed to evaluate the performance effectiveness of the rapid single kernel NIR sorting instrument, and to compare standard partial least squares regression (PLSR) with cross-validation and bagging PLSR for model development.

2. Experimental

2.1. Maize samples

Eighty-seven maize breeding samples, representing a wide range of size, shape, color and constituent compositions, were obtained from the International Maize and Wheat Improvement Center (CIMMYT, Texcoco, Mexico). Each maize breeding sample was composed of 30 kernels, thus the experimental set consisted of 2610 kernels. A few samples did not have reference values for a particular constituent, and were likewise removed from analysis of that constituent. All samples were allowed to equilibrate under laboratory conditions (20 °C, 50% relative humidity) for at least seven days. Afterward, images of the kernels were taken for documentation. Mean kernel mass was determined by weighing the bulk sample in an analytical balance and dividing by the number of kernels.

2.2. NIR instrument and spectra acquisition

The instrument consists of an NIR spectrometer, a light tube assembly, a control circuit and a computer, as shown in Fig. 1 The CDI Spectrometer (Control Development, Inc., South Bend, IN) has a thermoelectrically cooled InGaAs sensor with a spectral range of 904–1685 nm. It is controlled using a Microsoft Visual C++ 6.0 program using CDI's dynamically linked program library. The light tube assembly has 48 miniature tungsten light bulbs, arranged equidistantly in six rows along the tube periphery. A fiber-optic switch is used to monitor the passage of the kernels in the upper portion of the instrument. A complete description of the construction and operation of the assembly is provided in the paper by Armstrong (2006). One difference between the instrument used by Armstrong (2006) and that used in this study is that a bifurcated fiber (BIF600 Vis-NIR; 600 μm core diameter; Ocean Optics, Dunedin, FL, USA) collected spectra from both ends of the tube rather than the single fiber previously used. These new optical fibers are shielded from direct light since their axes are situated along the longitude of the tube and perpendicular to the light bulbs. Thus, this arrangement ensures that the fibers collect diffuse light coming from the sample seeds.

Prior to collection of the NIR absorbance spectra, the instrument was allowed to warm to its operating conditions for at least an hour. At the start of a batch of 10 samples, a dark spectrum (mean of 10 spectra) was taken by shutting off power to the light source and room lights followed by a reference spectrum taken from a thin white disk of Spectralon diffuse reflectance standard (Labsphere Inc., North Sutton, NH) positioned in the center of the tube. The reference spectrum was taken after allowing the lamps to re-warm for a minimum of three min. The exposure or integration time was fixed at 43 ms for all spectra acquisitions.

Each sample was placed in a vibratory feeder that dropped the kernels individually through the instrument. As a kernel moved down the light tube, its presence was sensed by the fiber-optic switch at the top, which triggered a small electronic time delay. At the end of the delay, the spectrometer was triggered to acquire a spectrum, and to send the data to the computer. Absorbance values were automatically computed and stored for each kernel.

2.3. Reference values and data analysis

Mean absorption spectra were computed by averaging the spectra of the kernels within each sample. A calibration file for each constituent was prepared by merging the spectra and the reference values, using either the relative percentage or computed absolute amount per kernel. The five constituents: crude protein, lysine, tryptophan, oil and soluble sugar contents were previously analyzed using appropriate reference methods. Briefly, these methods were: oil, AOAC Method 7.044 (AOAC, 1975); protein determined by Technicon Autonalyzer II – Industrial method #334–74, 1977; tryptophan was determined by a colorimetric method based on glyoxilic acid; lysine was determined by the colorimetric method given by Tsai et al. (1972). All represent bulk reference values expressed on a dry matter basis (relative contents), and were provided by CIMMYT. The absolute constituent amount for each sample was computed by multiplying the mean kernel mass by the relative contents. In quantifying the absolute constituent amount, it was assumed that the samples had approximately the same level of equilibrium moisture content. It is important to note that this is technically not an absolute amount, since relative contents are typically expressed on some standard moisture contents (MC). Moisture determination was not possible due to the small sample size, and the need to preserve the sample.

Eighty percent of the total samples were randomly selected for the calibration set, and the remaining 20% were used for the validation set. PLSR with cross-validation was used to develop models for each of the constituents from the calibration set. This was accomplished using the ParLeS software (Viscarra Rossel, 2007a), with the best models selected based on the value of the root mean square error (RMSE) approaching a minimum as factors were increased. Spectral pre-processing procedures using mean centering to zero, multiplicative scatter correction (MSC) and standard normal variate (SNV) were explored in the analysis. The spectral range was limited to 950–1650 nm because low signal-to-noise ratios are usually found at the edges of the spectral range. While the cross-validation procedure systematically removed one sample at a time (leave-one-out) in the PLSR analysis to calculate the beta coefficients, the capabilities of the models for prediction were evaluated using the validation set.

Bagging PLSR has been previously shown to improve a model's accuracy in predicting constituent values, the details of which are

Table 1
Descriptive statistics of the constituent composition of samples used for calibration and validation.

	Calibration set						Validation set					
	Mean	SD	CV	Min	Max	N	Mean	SD	CV	Min	Max	*N*
Relative contents (%)												
TRP	0.09	0.02	20	0.06	0.12	70	0.08	0.02	25	0.05	0.12	17
LYS	0.41	0.08	20	0.25	0.57	62	0.43	0.10	24	0.28	0.56	16
CP	11.35	1.17	10	9.32	14.79	70	11.67	0.97	8	10.00	13.13	17
OIL	4.32	0.76	18	2.62	6.08	70	4.29	0.56	13	3.50	5.55	17
SSC	4.73	1.26	27	2.03	7.50	68	4.81	1.64	34	2.65	8.08	17
Absolute contents (mg kernel^{-1})												
TRP	0.21	0.07	35	0.09	0.43	70	0.19	0.07	35	0.06	0.32	17
LYS	0.96	0.36	38	0.42	2.20	62	1.01	0.28	28	0.54	1.60	16
CP	26.55	7.94	30	13.45	47.68	70	30.44	11.30	37	18.09	55.77	17
OIL	10.17	3.63	36	3.81	23.11	70	11.80	6.43	55	3.37	29.23	17
SSC	10.90	3.09	28	4.85	19.00	68	10.96	3.45	31	6.18	16.85	17
Mass	0.24	0.07	29	0.10	0.48	70	0.24	0.06	25	0.16	0.38	17

TRP: Tryptophan; LYS: Lysine; CP: Crude Protein; OIL: Oil content; SSC: Soluble sugar content; Mass: Kernel mass, g SD: Standard deviation; CV: Coefficient of variation=SD/Mean×100%; Range: Minimum−Maximum; *N*: No. of samples.

Table 2
Pearson correlation coefficient matrix for the five reference constituents.

	Relative contents					Absolute contents				
	TRP	LYS	CP	OIL	SSC	TRP	LYS	CP	OIL	SSC
TRP	1.00					1.00				
LYS	0.86	1.00				0.95	1.00			
CP	−0.01	0.22	1.00			0.76	0.80	1.00		
OIL	0.21	0.27	0.01	1.00		0.79	0.79	0.86	1.00	
SSC	0.04	−0.12	−0.21	−0.10	1.00	0.33	0.26	0.33	0.42	1.00
Mass	−0.01	0.08	0.14	0.42	−0.46	0.79	0.80	0.96	0.92	0.42

TRP: Tryptophan; LYS: Lysine; CP: Crude Protein; OIL: Oil contents; SSC: Soluble sugar contents; Mass: Kernel mass.

discussed by Viscarra Rossel (2007b) and Breiman (1996). Twenty-five bootstrap sets were prepared by randomly selecting samples with replacement. This sampling procedure likely resulted in some samples being selected two, three or more times within a set. The number of samples in each set was about 66% of the calibration set available for a constituent. Each set was analyzed using PLSR with cross-validation as described above. The final prediction model was derived from the average of the beta coefficients of the 25 sets with the factor level the same for each set. The models were similarly validated using the same validation set.

3. Results and discussion

3.1. Description of samples

A statistical summary of characteristics for tryptophan (TRP), lysine (LYS), crude protein (CP), oil and soluble sugar contents (SSC) for the calibration and prediction sets are shown in Table 1. There was a fairly large constituent variation in the samples. For relative contents, SSC had the highest coefficient of variation (CV) while CP had the least. Using their absolute contents, the CV values significantly increased, except for soluble sugar contents which essentially remained constant.

To achieve a robust prediction model for a target constituent, a broad range of reference values are needed to avoid predictions beyond what are established by the calibration. By using a wide range of genetic material, this condition was easily achieved. Additionally, the levels of the constituents should be large enough to affect the absorbance of light energy in the spectrum.

Table 2 shows the Pearson correlation coefficient matrices for the relative and absolute contents between the five constituents, which were computed using the entire sample set. The relative contents of TRP and LYS had a reasonably high correlation that significantly increased when their absolute contents were considered. This suggests that any calibration conducted with one of these constituents would allow for prediction of the constituent value for the other. These amino acids are building blocks for protein complexes, but a correlation between them and crude protein for relative contents was practically absent. The correlations were slightly significant when absolute contents were considered. The same observations were found between OIL and CP contents. SSC did not correlate with any of the other constituents in both relative and absolute contents.

Interestingly, the mean kernel mass correlated well with the absolute amounts of CP and OIL constituents. Mean kernel mass correlated less with the amino acid absolute contents, and was uncorrelated with SSC. The relative contents of crude protein and oil both had a narrow range of values (low CV). The low CV caused the computation of the absolute contents of these constituents to be dependent on the size of the kernels that had a similar endosperm-to-germ ratio. However, what we typically want to achieve is the ability to predict relative constituent levels in order to identify individual kernels that will advance to the next stages in the breeding program and save fieldwork and costs in screening material.

The mean spectral profile of the 87 samples used in the study had the usual significant absorption peaks between 975–1025 nm and at 1450 nm for moisture, and 1175–1225 nm for protein. The scattering of the spectra could be easily resolved by applying a mean centering to zero as a pre-processing option. Multiplicative

Table 3
Calibration statistics for the five reference constituents using relative contents (%) and absolute contents (mg kernel^{-1})for spectra in the 950–1650 nm range, and partial least squares regression with cross-validation.

Relative contents (%)							Absolute contents (mg kernel^{-1})					
Constituent	Nf	SECV	R^2cv	SEP	R^2p	RPD	Nf	SECV	R^2cv	SEP	R^2p	RPD
Mean centered spectra												
TRP	6	0.02	0.28	0.02	0.41	1.38	3	0.04	0.73	0.05	0.40	1.30
LYS	6	0.08	0.13	0.07	0.56	1.15	6	0.20	0.68	0.18	0.60	1.58
CP	7	1.00	0.40	0.70	0.65	1.42	6	2.88	0.87	3.64	0.89	3.15
OIL	6	0.72	0.15	0.58	0.07	1.02	9	2.41	0.57	2.49	0.88	2.53
SSC	7	1.16	0.32	1.28	0.03	0.80	6	2.88	0.16	2.71	0.34	1.31
Mass							6	0.02	0.90	0.02	0.88	2.98
Mean centered spectra + MSC												
TRP	6	0.02	0.31	0.01	0.54	1.54	4	0.04	0.70	0.06	0.43	1.20
LYS	6	0.07	0.19	0.08	0.47	1.01	4	0.21	0.66	0.21	0.48	1.40
CP	8	0.70	0.64	0.47	0.75	1.74	6	3.40	0.82	4.08	0.87	2.33
OIL	7	0.71	0.17	0.46	0.29	1.13	9	2.32	0.60	2.98	0.82	2.14
SSC	6	1.08	0.37	1.45	0.02	0.77	5	2.90	0.14	2.90	0.25	1.21
Mass							6	0.03	0.81	0.03	0.76	2.13
Mean centered spectra + SNV												
TRP	6	0.02	0.31	0.01	0.54	1.55	4	0.04	0.69	0.06	0.43	1.17
LYS	6	0.07	0.19	0.08	0.46	1.17	4	0.21	0.65	0.20	0.48	1.38
CP	8	0.70	0.64	0.47	0.75	2.02	6	3.42	0.81	4.48	0.87	2.56
OIL	7	0.72	0.17	0.47	0.29	1.22	9	2.33	0.59	2.98	0.82	2.14
SSC	6	1.08	0.37	1.45	0.02	0.76	5	2.90	0.14	2.90	0.25	1.22
Mass							6	0.03	0.81	0.03	0.76	2.17

TRP: Tryptophan; LYS: Lysine; CP: Crude Protein; OIL: Oil contents; SSC: Soluble sugar contents; Mass: Kernel mass, g; SECV, %: Standard error of calibration with cross-validation; SEP: Standard error of prediction, %; R^2cv: Coefficient of determination for calibration with cross-validation; R^2p: Coefficient of determination for validation/prediction; RPD: Ratio of standard deviation to standard error of prediction; Nf: No. of PLSR factors; MSC: Multiplicative scatter correction; SNV: Standard normal variate.

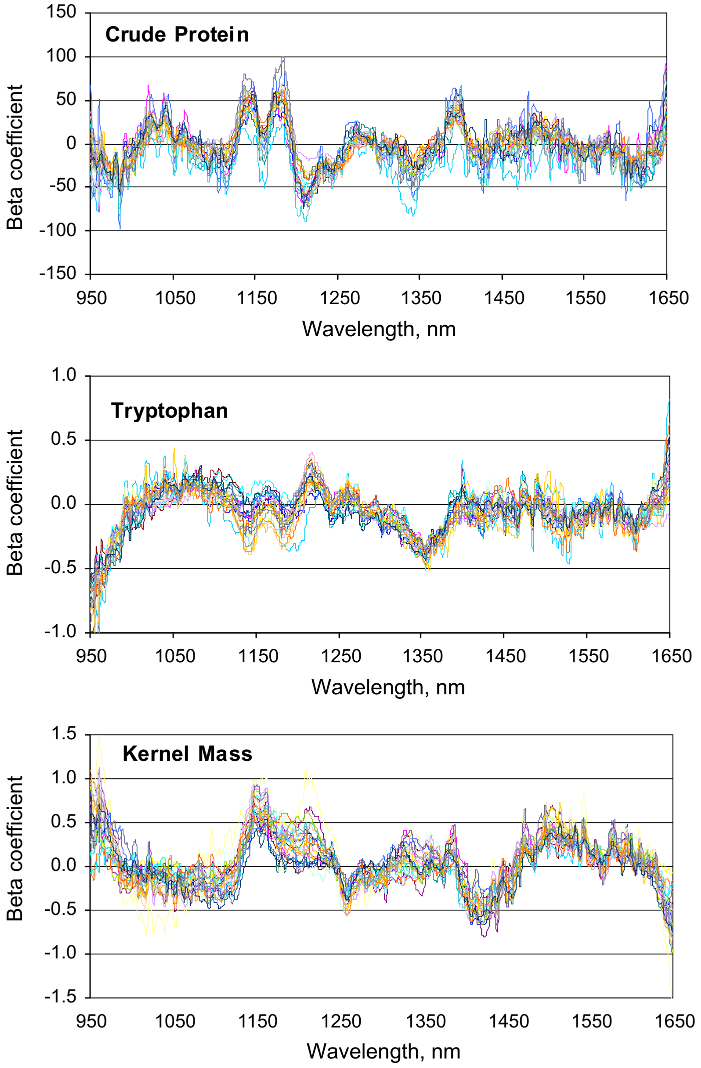

Fig. 2. Beta coefficient spectra of the 25 bagging PLSR calibration models for relative contents of crude protein and tryptophan, and kernel mass.

scattering correction (MSC) was also explored along with the standard normal variate (SNV) pre-treatment in the analysis. MSC was shown by Armstrong (2006) and Cogdill et al. (2004) to be able to remove spectral noise.

Calibrations between the spectrum of individual kernels and their reference values would be ideal. However, because of limitations in the sample size and the use of reference methods that require a large amount of sample for reliable measurements, bulk values were matched to the bulk spectra of the samples. In the case of this study, the best estimator of the bulk spectra would be the mean of the spectra of the individual kernels comprising the sample (Delwiche and Hruschka, 2000). Some levels of accuracy might be lost in the procedure. However, the predictive performance of the sorter instrument, which employs a novel technique in measuring the NIR spectra, was a prime consideration in this study.

3.2. Calibration models for relative constituent contents

Table 3 shows the results of a partial least squares regression for the relative amount of the constituents. Consistent with the experience of Baye et al. (2006), it was difficult to obtain useful prediction models when relative constituent contents of the samples were considered in the calculations. Following the guidelines for interpretation of modeling results that were given by Williams and Norris (2001), Williams (2005), and using the coefficient of determination (R^2) as cited by Kovalenko et al. (2006), the best model obtained for CP is suitable for sample screening with an R^2 of 0.75 for the prediction set and a standard error of prediction (SEP) of 0.47 when corrections for spectral scattering (MSC or SNV) were applied with mean spectral centering to zero. Apparently, models for TRP appear suitable for very rough sample screening. However, because of significant differences between the cross-validation and prediction R^2 values, the models may seem to be less stable but nevertheless show a possibility for calibration. In contrast, models for LYS have shown greater differences between the two R^2 values, implying that an accurate prediction is unlikely. Kovalenko et al. (2006) noted that the greatest challenge for developing calibration models for amino acids is how “to exceed the correlation between amino acids and protein concentrations”, which this study has shown to be possible, particularly for TRP.

Calibration models that used either MSC or SNV spectral preprocessing procedures produced better statistics than mean centering-to-zero alone. Both MSC and SNV helped to reduce the scattering of sample spectra due to the operation of the NIR sorting instrument. As kernels tumble down through the instrument’s light tube, aside from interrogating multiple points on the surface by the two bifurcated probes, the motion creates variability in the distance from the points to the probes and the spectral data. MSC is less attractive to use, because it requires a characteristic reference spectrum from the calibration set against which individual sample spectra are regressed upon (Maleki et al., 2006). On the other hand, SNV implements a simpler algorithm that scales the spectra and minimizes sample presentation effects (Barnes et al., 1989), which gave performance levels similar to those seen with MSC. The prediction plots of the constituents are shown in Fig. 2 for the models using mean centering and SNV.

3.3. Calibration models for absolute constituent contents

Table 3 also shows the calibration results for the absolute contents of the five constituents for three pre-processing data treatments. There was significant improvement in the calibration using absolute contents data compared to the relative contents data particularly for CP and OIL. CP still showed the highest R^2p (0.87) followed by the OIL contents, then the amino acids (TRP and LYS). SSC remained poorly predicted. Comparing the R^2cv and R^2p values, the TRP, LYS, and OIL calibration models do not appear to be stable. The model statistics can easily vary with the random selection of samples for calibration and validation sets.

The study of Baye et al. (2006) on maize concluded that single kernel near infrared spectroscopy most likely reports the absolute contents of constituents that account for the mass of individual kernels. They further explained that regressions based on relative contents would be less accurate because any changes in the compositional amount of a constituent are “not translated into differences within the NIR spectra.” Gergely and Salgo (2007) made the same conclusions in their study on wheat kernel compositional changes at different stages of development.

As demonstrated in Table 3, a calibration model suitable for rough sample screening application of kernel mass was found better when mean centering-to-zero alone is used than when using either MSC or SNV pre-treatments. This could be explained by the fact that the size and geometry of the kernels can affect the spectra as the kernels tumble down the light tube. This finding may have dual implications. First, this could eliminate a separate step for measuring individual kernel mass or estimating average kernel mass from bulk weights. Second, one can easily convert between relative and absolute contents of the constituents. However, the correlation matrix (Table 2) suggests a possible complication in the prediction of absolute contents of CP, TRP, LYS and OIL contents because of their seemingly high correlation with kernel mass. The predictions for these absolute constituent contents are most likely estimated indirectly from the kernel mass. Therefore, future studies should carefully account for the correlations existing between the absolute contents of the constituents and kernel mass.

3.4. Bagging PLSR

The results of bagging PLSR analysis are shown in Table 4 for relative contents of TRP and CP, and kernel mass, using mean centered NIR spectra. Viscarra Rossel (2007b) found bagging PLSR to be more robust than the normal PLSR with cross-validation, less prone to over fitting, improved prediction and provides statistical measures of the model. For this case, the two methods of model development yielded statistically similar models. The case for being more robust with future samples is not readily apparent. An interesting feature of bagging PLSR is that if the number of factors is increased, the model statistics continue to improve significantly, whereas for PLSR with cross-validation, no real improvement is realized beyond the level suggested by the RMSE-factor plot. It is not readily apparent how much the bagging model may be over-fit to the data for these cases. Some examples of beta coefficient spectra using the bagging PLSR technique are shown in Fig. 2. Note that there was a substantial range of beta coefficient values at a wavelength. However, they tend to aggregate into a central value, such that a characteristic spectrum may be derived, and can serve as an unbiased estimate of the true beta coefficient spectrum. This would then give a better calibration model for a constituent, reduce potential experimental errors due to calibration sample selection, which PLSR is particularly sensitive to, and improve the robustness of predictions in the long run.

Table 4
Model statistics for the selected three significant constituents using relative contents data (%) and Bagging PLSR with cross-validation.

Constituent	Nc	Nf	SEC	R^2c	Np	SEP	R^2p	RPD
TRP	70	6	0.014	0.649	17	0.016	0.440	1.34
CP	70	7	0.775	0.679	17	0.657	0.679	1.48
Mass	70	6	0.018	0.965	17	0.020	0.890	2.98

TRP: Tryptophan; CP: Crude Protein; Mass: Kernel mass; SEC: Standard error of calibration (%); R^2c: Coefficient of determination for calibration; SEP: Standard error of prediction (%); R^2p: Coefficient of determination for validation/prediction; RPD: Ratio of standard deviation to standard error of prediction; Nc: No. of samples for calibration; Np: No. of samples for validation/prediction; Nf: No. of PLSR factors.

4. Conclusion

Single kernel NIR spectroscopy can be used to predict the compositional levels of some constituents for maize using the rapid single kernel NIR sorting instrument and partial least squares regression. While the calibration models using either MSC or SNV for relative contents of crude protein and kernel mass were found to be suitable for rough sample screening based on their coefficients of determination, the models are less useful for all other constituents investigated. Further validation of the models is needed. Because of relatively high levels of correlation between most constituents and kernel mass, the calibration models for absolute contents were found to be less promising, even though they show good modeling statistics. Finally, bagging PLSR model accuracy was comparable to that of PLSR with cross-validation.

References

Association of Official Analytical Chemists (AOAC), 1975. Official Methods of Analysis, 12th ed. AOAC, Washington, DC.

Armstrong, P.R., 2006. Rapid single kernel NIR measurement of grain and oil-seed attributes. Applied Engineering in Agriculture 22, 767–772.

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.

Baye, T.M., Pearson, T.C., Settles, A.M., 2006. Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. Journal of Cereal Science 43, 236–243.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Cogdill, R.P., Hurburgh Jr., C.R., Rippke, G.R., Bajic, S.J., Jones, R.W., McClelland, J.F., Jensen, T.C., Liu, J., 2004. Single-kernel maize analysis by near-infrared hyperspectral imaging. Transactions of the American Society of Agricultural Engineers 47, 311–320.

Delwiche, S.R., Hruschka, W.R., 2000. Protein content of bulk wheat from near-infrared reflectance of individual kernels. Cereal Chemistry 77, 86–88.

Gergely, S., Salgo, A., 2007. Changes in protein content during wheat maturation – what is measured by near infrared spectroscopy? Journal of Near Infrared Spectroscopy 15, 49–58.

Janni, J.A., 2007. Non-destructive derivation of weight of single seed or several seeds. U.S Patent No 7,274,457 B2.

Janni, J.A., Weinstock, B.A., Hagan, L., Wright, S., 2008. Novel near-infrared sampling apparatus for single seed analysis of oil content in maize. Applied Spectroscopy 62, 423–426.

Kovalenko, I.V., Rippke, G.R., Hurburgh, C.R., 2006. Determination of amino acid composition of soybeans (*Glycine max*) by near-infrared spectroscopy. Journal of Agricultural and Food Chemistry 54, 3485–3491.

Maleki, M.R., Mouazen, A.M., Ramon, H., De Baerdemaker, J., 2006. Multiplicative scatter correction during online-line measurement with near infrared spectroscopy. Biosystems Engineering 96, 427–433.

Ngonyamo-Majee, D., Shaver, R.D., Coors, J.G., Sapienza, D., Correa, C.E.S., Lauer, J.G., Berzaghi, P., 2008. Relationships between kernel vitreousness and dry matter degradability for diverse corn germplasm, I. Development of near-infrared reflectance spectroscopy calibrations. Animal Feed Science and Technology 142, 247–258.

Orman, B.A., Schumann Jr., R.A., 1991. Comparison of near-infrared spectroscopy calibration methods for the prediction of protein, oil and starch in maize grain. Journal of Agricultural and Food Chemistry 39, 883–886.

Tsai, C.Y., Hansel, L.W., Nelson, O.E., 1972. A colorimetric method of screening maize seeds for lysine content. Cereal Chemistry 49, 572–579.

Viscarra Rossel, R.A., 2007a. ParLeS: software for chemometric analysis of spectroscopic data. Chemometrics and Intelligent Laboratory Systems 90, 72–83.

Viscarra Rossel, R.A., 2007b. Robust modeling of soil diffuse reflectance spectra by bagging-partial least squares regression. Journal of Near Infrared Spectroscopy 15, 39–47.

Wehrens, R., Putter, H., Buydens, L.M.C., 2000. The bootstrap: a tutorial. Chemometrics and Intelligent Laboratory Systems 54 (35), 52.

Williams, P., Norris, K., 2001. Near-Infrared Technology in the Agricultural and Food Industries, second ed. American Association of Cereal Chemists Inc., St. Paul, MN.

Williams, P.C., 2005. Near-infrared Technology – Getting the Best Out of Light. PDK Projects, Inc, Nanaimo, British Columbia, Canada.