Single-Kernel Near-Infrared Protein Prediction and the Role of Kernel Weight in Hard Red Winter Wheat

T. Bramble, F. E. Dowell, T. J. Herrman

Abstract. *A near-infrared single-kernel protein calibration for hard red winter wheat (*Triticum aestivum *L.) was developed to support research mapping the variance structure of single-kernel protein in commercial wheat fields. The hierarchical sampling design used to map the variance structure included fields, plots, rows, plants, heads, spikelet, and kernels from 47 fields containing the cultivars Jagger, 2137, Ike, or TAM 107. Each kernel was evaluated for protein content using an automated single-kernel NIR system. Five hundred kernels were selected as the model development set and reference protein content was determined using a combustion nitrogen analyzer. The resulting 11 factor PLS model had a standard error of prediction based on a cross validation (SECV) of 1.21% and r^2 = 0.84. Application of a kernel weight correction improved model performance statistics (SECV = 0.40%, r^2 = 0.89). A moderate negative correlation was observed (r = -0.55) between kernel weight and protein content. Previous research exploring single-kernel protein had not documented this relationship. The partial least squares model containing a kernel weight adjustment was most accurate with Jagger kernels (SECV = 0.32%, r^2 = 0.92) and least accurate with TAM 107 kernels (SECV = 0.51%, r^2 = 0.82). The application of the weight correction factor resulted in a lower SECV than previous research. Currently, single-kernel protein analysis instruments have not included a kernel weight apparatus, which represents a constraint in accurately predicting single-kernel protein using NIR technology.*

Keywords. *Automation, Grading, Near-infrared, NIR, Protein, Single-kernel, Wheat.*

Protein content is an important indicator of bread quality (Bushuk et al., 1969), with loaf volume generally increasing as protein content increases. However, protein content varies within wheat (*Triticum aestivum* L.) fields, plots, rows, plants, heads, spikelets, and position within spikelets (Bramble et al., 2002). Measuring protein content in bulk samples will not give an accurate indication of the protein content variation since protein content varies from kernel to kernel. Thus, technology and calibrations are needed to automatically measure single-kernel protein content.

Near-infrared (NIR) spectroscopy has been commonly used to measure protein content of wheat bulk samples (Hunt et al., 1977), but only more recently for single-kernel protein content. The original single-kernel work involved NIR transmittance, or passing energy through a kernel, to measure protein content (Delwiche, 1995); however, there has been a shift toward measuring reflected energy because it is easier to automate this process (Wang et al., 1999). The pioneering work by Delwiche (1995) involved precise kernel placement in front of the spectrometer that collected the kernel transmittance spectra from 740 to 1139 nm at 2-nm intervals, but was limited to 500 kernels per class. Using a 14-factor partial least squares (PLS) model, Delwiche obtained a standard error of prediction (SEP), or standard deviation of the differences in the measured and predicted protein values, of 0.83% and r^2 = 0.85. Delwiche (1998) investigated the use of NIR reflectance as a means of determining single wheat kernel protein content. In this work, kernels were hand-placed on the end of a revolving tube. The 1100- to 1798-nm range was used in the model development. Each kernel's spectrum was the average of 32 scans. This methodology, using an 11-factor PLS model, resulted in a SEP of 0.59% .

In research to automate single wheat kernel protein content measurements, Dowell et al. (1997) integrated a DA-7000 diode-array spectrometer with a Single Kernel Characterization System (SKCS 4100) (Perten Instruments, Stockholm, Sweden). This system, referred to as a SKCS 4170, automatically feeds single kernels to a viewing area where visible and NIR spectra are collected over the wavelength range of 400 to 1700 nm from randomly oriented kernels. When using kernels randomly placed by the automated system, a 19-factor PLS protein prediction model resulted in a SEP based on a cross-validation of 0.75% and r^2 = 0.94. Hand-placed kernels with fixed orientation yielded a SEP based on a cross-validation of 0.60% and r^2 = 0.96, indicating that random placement only slightly affects prediction accuracy.

The work reported here describes the development of a single-kernel protein prediction model for use in related research to study the single-kernel protein variance structure

Submitted for review in November 2004 as manuscript number BE 5613; approved for publication by the Biological Engineering Division of ASABE in August 2006.

The authors are **Tod Bramble,** Bakery Flour Sales Manager, King Arthur Flour Company, Norwich, Vermont; **Floyd E.** Dowell, ASABE Member Engineer, Research Leader, USDA ARS Grain Marketing and Production Research Center, Manhattan, Kansas; and **Timothy J. Herrman,** Director, Office of the Texas State Chemist, Texas A&M University, College Station, Texas. **Corresponding author:** Timothy J. Herrman, Office of the Texas State Chemist, Texas A&M University, College Station, TX 77841; phone: 979-845-1121; fax: 979-845-1389; e-mail: tjh@otsc.tamu.edu.

in commercial wheat fields in western Kansas as reported by Bramble et al. (2002). In the current study, single kernels were sampled using a hierarchical sampling method such that protein variance could be estimated for seven components in a commercial wheat production system in southwest Kansas. These components included variance among fields, among plots in a single field, among rows in plots, between two adjacent plants in a row, among heads on a plant, among spikelets on a head, and between two kernels in a spikelet. The nature of this work did not allow for bulk measurement of protein, as individual kernels were uniquely identified by their specific location within the sampling hierarchy. Protein content was measured for each individual kernel in this research requiring use of the SKCS 4170.

Therefore, a protein prediction model was developed for this unique sample set of data using the SKCS 4170. In view of the large variation in kernel protein content and weight resulting from the hierarchical sampling of commercial wheat fields, the objective of this research involved measuring the accuracy of the SKCS 4170 for predicting protein content of individual kernels. This sample set likely provides a better assessment of single-kernel protein NIR measurement compared to past studies and will help identify constraints in single-kernel NIR technology.

Materials and Methods

Individual Wheat Kernel Samples

Single kernels of hard red winter wheat (HRW) were collected from 47 fields under commercial production in Stanton and Kearney counties in southwest Kansas. The four cultivars in the study included Jagger (13 fields), 2137 (14 fields), Ike (11 fields), and TAM 107 (8 fields).

Sampling of the physiologically mature kernels occurred just prior to harvest. Individual kernels were obtained from three plots within each of the study fields. Individual heads were sampled at random from each plant, kept intact, and uniquely identified by field, plot, row, and plant location. The sample collection procedure is described in Bramble et al. (2002).

Collection of Individual Kernel Spectral Data

Spectral data for each kernel were collected using the Perten SKCS 4170 within three months of sample collection. The spectrometer in this system measures reflected energy at 400 to 1700 nm. The spectrometer measures reflected energy every 7 nm from 400 to 1100 nm, and every 11 nm from 1100 to 1700 nm. Kernels were hand-placed without regard to orientation (crease up or down) resulting in an almost random placement pattern in the sample bucket. Baseline readings were run every 100 to 200 samples. The system sampled 8 spectra per kernel and stored the average.

Protein Model Development

A total of 500 kernels from the sample set were chosen randomly, but in a manner such that every plot from each sampled field was represented by at least one kernel. These kernels, comprising the model development set, were analyzed for nitrogen content (N*5.7) by combustion using the LECO Model FP-428 nitrogen analyzer (St. Joseph, Mich.) according to AACC Approved Method 46-30 (AACC, 2000). Samples were analyzed for protein content within a few weeks after collecting spectra. This instrument was calibrated prior to analysis, with the model development set being run as a continuous lot and periodic test samples run throughout.

The stored spectral data and the LECO combustion nitrogen results were used to develop the single-kernel protein prediction model using a partial least squares (PLS) regression and a cross-validation. In a cross-validation, a sample is removed and a calibration developed without it. The removed sample is then predicted from the resulting calibration and the residual recorded. The sample is then replaced and another sample removed and a new calibration is developed. The process is repeated until each sample has been removed and predicted. The standard deviation of the residuals is the SEP. All data were mean centered before analysis. The number of factors chosen for the PLS model was based on the minimum residual sum of squares.

The resulting protein prediction model was then applied to the spectral data collected for the complete sample set, and individual kernel protein contents were predicted. Model performance was evaluated and reported as multiple coefficient of determination (r^2), the standard error of cross-validation (SECV), and the ratio of the sample standard deviation of the reference data to the standard error of cross-validation (RPD) as described by Williams (2001). Individual kernel data were analyzed using the descriptive statistics function ($\alpha = 0.05$) in Excel© (Microsoft Inc., Redmond, Wash.).

Results and Discussion

Figure 1 illustrates performance of the protein prediction model by plotting model results against protein reference values determined by the LECO combustion nitrogen analyzer. The SECV for the 11 factor PLS model was 1.21% with an $r^2 = 0.84$. The prediction model for individual protein content provided good results in the middle protein range (10% to 20%); however, there is an apparent departure from the ideal slope at protein contents greater than 20%. It appears that the protein prediction model is under-predicting kernel protein when the content is greater than 20%, although there are too few kernels to validate this observation (fig. 1).

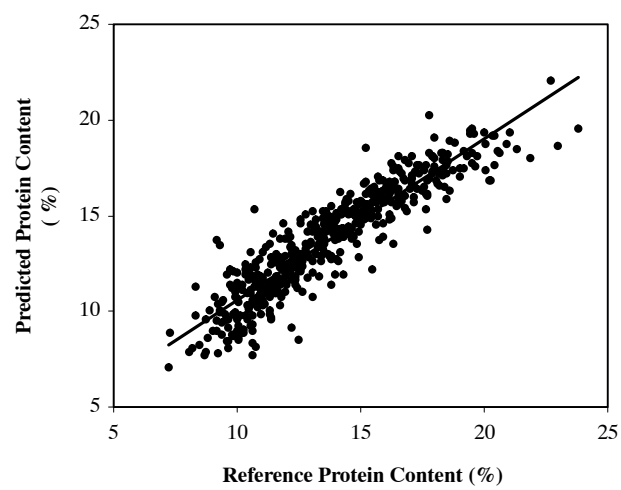

Figure 1. Protein content prediction model performance: reference protein content plotted against predicted protein content.

The protein prediction model developed here had a higher error than previous models reported by Delwiche (1998) and Dowell et al. (1997). Hruschka (2001) identified nearly 40 sources of error attributable to NIR spectroscopy. Of these, sample variables and operational variables are the two major sources of error affecting measurement accuracy. Sample variables including kernel size, density, and color all affect spectral measurement by influencing the absorption or reflection of a wheat sample (Watson et al., 1977). The repeatability and reproducibility of the LECO protein measurement method are about 0.15% and 0.27%, respectively (Bicsak, 1993). There was a moderate negative correlation between kernel weight and protein content (n = 500, r = -0.55, fig. 2). The correlation appears strongest for kernels with the highest weights. Other researchers did not find a kernel weight to protein correlation (Delwiche, 1995; Dowell et al., 1997; Delwiche, 1998) in previous single-kernel prediction model development. In those studies, kernels were obtained from commercial samples taken from lots from multiple locations throughout the Midwest. However, Wilkins et al. (1993) showed that protein content is inversely related to kernels size, particularly for non-irrigated wheat. For development of the model reported here, 10,150 kernels were sampled randomly from fields in southwest Kansas prior to mechanical harvesting. The 500 kernels included for use in model development were randomly obtained from this sample set and included without regard to size, color, or condition (table 1). Individual kernels displayed a range in protein content from 7.2% to 23.8% and kernel weight ranged between 5 to 43 mg. The kernel population displayed a normal distribution for weight and protein measurements for each cultivar, with the exception cultivar 2137 protein content, which displayed a positive skewness at a 95% confidence level. The individual kernel properties included in this study may not have been masked by mixing with samples from other growing regions as were the kernels used in previous studies. This enabled detection of the influence of kernel weight. This could be one possible explanation for the kernel weight correlation and the slightly higher SECV for our prediction model over previously reported models.

Thus, based on the findings by Wilkins et al. (1993) the model performance could be improved by using a weight-corrected average protein value. To achieve this end, the product of individual kernel weight and protein measures were multiplied by a weight-corrected protein factor calculated in equation 1:

$$P_{average} = \Sigma p_i m_i / \Sigma m_i \quad (1)$$

where Paverage is the weight-corrected protein and p and m represent the measured protein and kernel mass for the individual kernels (i). The difference between the predicted and actual individual weight-protein corrections were used to calculate the SECV. The weight correction yielded a plot much closer to the ideal slope (fig. 3). The SECV improved from 1.21% to 0.40% and the r^2 was slightly improved from 0.84 to 0.89.

To further evaluate this transformation, scatter plots of the residual errors for the original model (no weight correction) and predicted protein values were plotted (figs. 4, 5). The weight-corrected transformation had some effect, as evidenced by a slight flattening of the residual scatter plot, and the modest improvement in the performance statistics.

The ratio of the sample standard deviation to the standard error of cross-validation (RPD) for the protein prediction model was 2.58 without the kernel weight correction. Delwiche (1995) reported RPD's ranging from 2.88 for soft red winter wheat to 4.72 for hard white winter wheat. While just outside the lower end of their range, our RPD of 2.58 indicates an adequate standard deviation of the prediction model relative to the amount of variability found in the

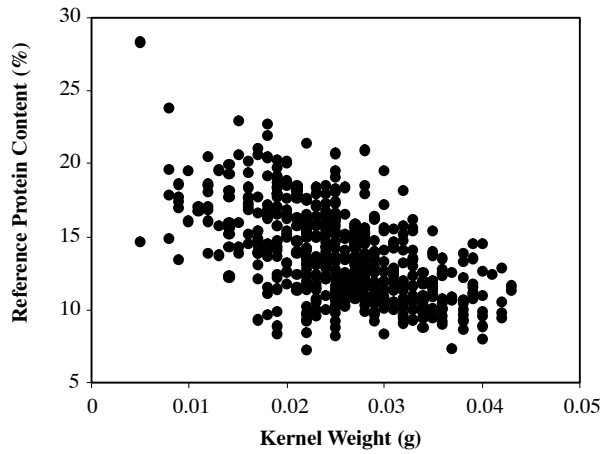

Figure 2. Protein content vs. kernel weight correlation.

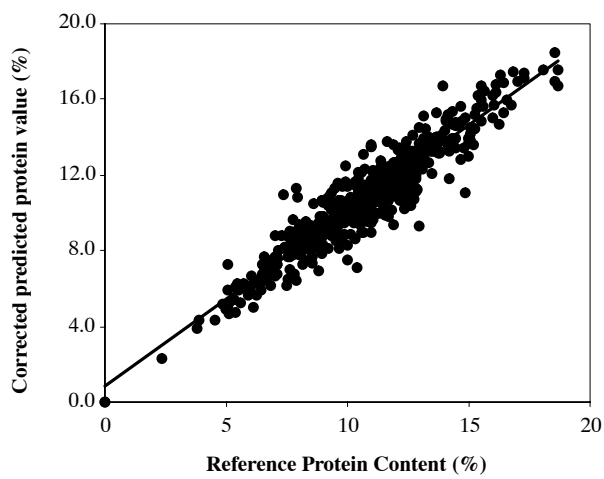

Figure 3. Weight-corrected protein content prediction model performance.

Table 1. Individual kernel protein and weight by cultivar as used for model development.

		Protein			Weight		
Cultivar	N	Mean (%)	Std. Dev.	Skewness	Mean (mg)	Std. Dev.	Skewness
Jagger	139	14.2	2.9	0.09	25.0	0.7	0.06
Ike	121	13.3	2.5	0.58	26.9	0.7	0.04
2137	152	13.1	3.2	0.94	25.4	0.7	-0.10
TAM 107	88	14.7	3.3	0.02	24.6	0.8	0.09

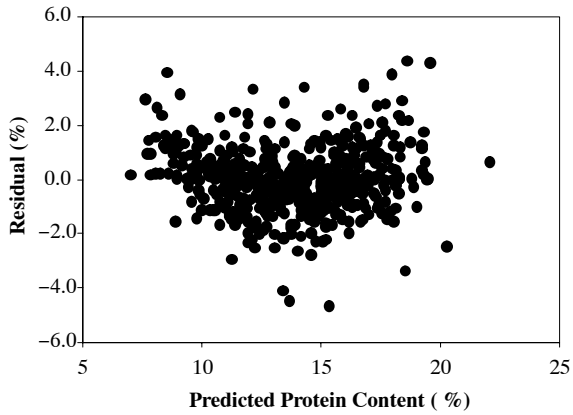

Figure 4. Residual scatter plot.

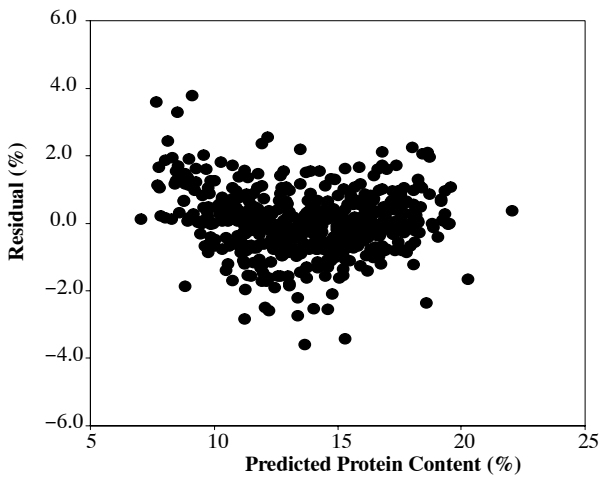

Figure 5. Residual scatter plot of weight-corrected protein content prediction model results.

sample set as a whole. The inclusion of the individual kernel weight correction improved the RPD to 7.69, which is considerably better than reported for prediction models that do not include weight.

The protein prediction model was developed using the entire sample set, but the model performed differently with each of the four cultivars (table 2). The model incorporating kernel weight was most accurate when predicting protein content of individual Jagger kernels with an SECV of 0.32%, r^2 of 0.93, and a RPD of 9.15. Model performance was also strong when predicting protein content of Ike (SECV = 0.36%, r^2 = 0.90, RPD = 6.9). The prediction model performed less consistently when evaluating kernel protein content for 2137 (SECV = 0.43%, r^2 = 0.79, RPD = 7.58) and TAM 107 (SECV = 0.51%, r^2 = 0.82, RPD = 6.77). The reasons for this inconsistency may be due to the kernel characteristics themselves (table 1), as evidenced by the larger standard deviation in kernel protein content and weight. In addition, 2137 was not normally distributed. For 2137, the weight correction factor did not improve the prediction model's coefficient of determination (table 2), which may also be attributable to individual kernel weight skewness.

Conclusion

This research documented the development and application of a single-kernel protein prediction model using instrumentation developed for industry and research applications. An 11-factor PLS protein prediction model was developed using hand-collected single-kernel samples from fields under commercial production in southwest Kansas. Model performance was slightly outside the range of previously published single-kernel NIR results, but the model development process and research application of this model may more closely mirror the type of use and variability found in a commercial wheat production setting.

There was a correlation between kernel weight and protein content for the entire data set that had not been seen in published single-kernel NIR model development work. This correlation affects model performance at the high end of the protein range (>20%) where the model appeared to be under-predicting protein content. The use of a weight correction factor reduced the standard error of calibration and coefficient of determination.

The model performance statistics differed for each of the varieties, with variety Jagger performing best, followed by Ike, 2137, and TAM 107. When applying the model to the research application for which it was developed (Bramble et al., 2002), the model appeared to perform satisfactorily.

References

American Association of Cereal Chemists (AACC). 2000. *Approved Methods of the AACC*, 10th ed. St. Paul, Minn.: AACC.

Bramble, T., T. J. Herrman, T. Loughin, and F. Dowell. 2002. Single kernel protein variance structure in commercial wheat fields in Western Kansas. *Crop Sci.* 42(5): 1488-1492.

Bicsak, R. C. 1993. Comparison of Kjeldahl method for determination of crude protein in cereal grains and oilseeds with generic combustion method: Collaborative study. *J. AOAC* 76(4): 780-786.

Bushuk, W., K. G. Briggs, and L. H. Shebeski. 1969. Protein quantity and quality as factors in the evaluation of bread wheat. *Canadian J. Plant Sci.* 49(2): 113-122.

Delwiche, S. R. 1995. Single wheat kernel analysis by near-infrared transmittance: Protein content. *Cereal Chem.* 72(1): 11-16.

Delwiche, S. R. 1998. Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. *J. Cereal* Sci. 27(3): 241-254.

Table 2. Comparison of single-kernel NIR prediction statistics with kernel weight correction and without kernel weight correction.

	Kernel Weight Correction			Without Kernel Weight Correction		
Cultivar	SECV	RPD	r^2	SECV	RPD	r^2
Jagger	0.32	9.15	0.93	0.93	3.13	0.90
Ike	0.36	6.90	0.90	1.04	2.40	0.84
2137	0.43	7.58	0.79	1.45	2.23	0.80
TAM107	0.51	6.77	0.82	1.52	2.27	0.82

Dowell, F. E., C. R. Martin, J. L. Steele, D. Wang, J. E. Baker, J. E. Throne, and S. R. Delwiche. 1997. Automated single kernel quality measurement using near-infrared reflectance. ASAE Paper No. 973022. St. Joseph, Mich.: ASAE.

Hruschka, W. R. 2001. Data analysis: Wavelength selection methods. In *Near Infrared Technology in the* Agricultural and Food Industries, 39-58. St. Paul, Minn.: American Association of Cereal Chemists, Inc.

Hunt, W. H., D. W. Fulk, B. Elder, and K. Norris. 1977. Collaborative study on infrared reflectance devices for determination of protein in hard red winter wheat, and for protein and oil in soybeans. *Cereal Foods World* 22(10): 534-536.

Wang, D., F. E. Dowell, and R. E. Lacey. 1999. Single wheat kernel size effects on near-infrared reflectance spectra and color classification. *Cereal Chem.* 76(1): 34-37.

Watson, C. A., W. C. Shuey, O. J. Bariasik, and J. W. Dick. 1977. Effect of wheat class on near-infrared reflectance. *Cereal Chem.* 54(6): 1264-1269.

Wilkins, D. E., C. L. Douglas, Jr., and D. B. Churchill. 1993. Soft white winter wheat kernel separation by percent protein. *Transactions of the ASAE* 36(6): 1841-1845.

Williams, P. C. 2001. Implementation of near-infrared technology. In *Near Infrared Technology in the Agricultural* and Food Industries, 145-169. St. Paul, Minn.: American Association of Cereal Chemists, Inc.