

# Examination of Spectral Pretreatments for Partial Least-Squares Calibrations for Chemical and Physical Properties of Wheat\*

STEPHEN R. DELWICHE† and ROBERT A. GRAYBOSCH

USDA/ARS, Beltsville Agricultural Research Center, Instrumentation and Sensing Laboratory, Building 303, BARC-East, Beltsville, Maryland 20705-2350 (S.R.D.); and USDA/ARS at Department of Agronomy and Horticulture, University of Nebraska, Lincoln, Nebraska (R.A.G.)

Use of near-infrared (NIR) diffuse reflectance on ground wheat meal for prediction of protein content is a well-accepted practice. Although protein content has a strong bearing on the suitability of wheat (*Triticum aestivum* L.) for processed foods, wheat quality, as largely influenced by the configuration and conformation of the monomeric and polymeric endosperm storage proteins, is also of great importance to the food industry. The measurement of quality by NIR, however, has been much less successful. The present study examines the effects and trends of applying mathematical transformations (pretreatments) to NIR spectral data before partial least-squares (PLS) regression. Running mean smooths, Savitzky–Golay second derivatives, multiplicative scatter correction, and standard normal variate transformation, with and without detrending, were systematically applied to an extensive set of hard red winter wheat and hard white wheat grown over two seasons. The studied properties were protein content, sodium dodecyl sulfate (SDS) sedimentation volume, number of hours during grain fill at temperature <24 °C, and number of hours during grain fill at temperature >32 °C. The size of the convolution window used to perform a smooth or second derivative was also examined. The results indicate that for easily modeled properties such as protein content, the importance of pretreatment was lessened, whereas for the more difficult-to-model properties, such as SDS sedimentation volume, wide-window (>20 points) smooth or derivative convolutions were important in maximizing calibration performance. By averaging 30 PLS cross-validation trial statistics (standard error) for each property, we were able to ascertain the inherent modeling ability of each wheat property.

Index Headings: Near-infrared preprocessing; NIR; Partial least squares; PLS; Wheat quality; Statistical analysis software; SAS.

## INTRODUCTION

Wheat (*Triticum aestivum* L.), in ground or bulk form, has long been a favorite of NIR spectroscopists and chemometricians in the testing of the performance of instruments and calibration equations. The present study follows in this tradition with the use of this commodity to explore the effect of spectral pretreatments during partial least-squares (PLS) calibration equation development. Wheat protein content and protein quality (specifically, the glutenin–gliadin complex) are fundamental properties that have a bearing on the suitability of flour in specific food products.<sup>1</sup> The quality of wheat is a product of genetics and the growth environment. Environmental ef-

fects often play a larger role in defining quality than genetic effects.<sup>2</sup> The difficulty with quality from the standpoint of the cereal scientist is that of its measurement, and specifically, how the easily identifiable aspects of quality that appear in the finished product (color, loaf volume, crumb grain structure in pan breads) are determined by the biochemical properties of the gluten proteins and their interaction with starch, lipids, and other classes of compounds that make up the wheat kernel. In addition to the classical procedures such as Kjeldahl digestion and Dumas (combustion) for protein content, numerous other devices and procedures have been developed to measure physical properties such as consistency, strength, elasticity, and tolerance of wheat flour during dough development. Instruments such as the Mixograph, Farinograph, Extensograph, and Alveograph were developed to characterize such properties.<sup>3</sup> Although very good at measuring specific rheological features, these instruments suffer either from the length of time needed per analysis (several minutes) or require a large mass of flour per sample, thus making them ill-suited for early generations in plant breeding programs. A biochemical technique, such as size-exclusion high-performance liquid chromatography (SE-HPLC),<sup>4–6</sup> is based on the quantitative measurement of the polymeric ( $M_r > 100$  k, primarily, glutenin) and large monomeric ( $M_r = 30–70$  k, primarily gliadin) protein molecules that largely influence dough rheological behavior and the texture and appearance of the finished product. Size exclusion HPLC procedures are very time-consuming, require extensive use of solvents, and require highly trained laboratory personnel. A simpler biochemical technique for measurement of protein quality is the sodium dodecyl sulfate (SDS) sedimentation volume method, which collectively measures the gluten protein complex and is reported to be a good indicator of heat stress in the developing grain.<sup>7,8</sup> Still, laboratories are limited to 50–100 SDS sedimentation volume analyses per day, owing to the time needed for sediment formation and cleanup. Such reasons have fostered the use of NIR spectroscopy for protein content analysis during the past 30 years and why continued effort is involved in developing this technique for wheat quality analysis.<sup>9,10</sup>

Recent efforts on the use of NIR spectroscopy in wheat quality analysis have dealt with the feasibility of using this procedure on harvested wheat for gauging environmental stress (particularly temperature) to the plant during development.<sup>11</sup> Although warm temperatures (>30 °C) generally favor wheat quality for breadmaking,<sup>12</sup> ex-

Received 2 January 2003; accepted 14 July 2003.

\* Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply endorsement of recommendation by the USDA.

† Author to whom correspondence should be sent. E-mail: delwiche@ba.ars.usda.gov.

cessive temperatures during grain fill, for still-uncertain reasons, can have the opposite effect. The exact mechanism for the manner in which temperature during the postanthesis development period affects the endosperm protein structure of wheat is a topic of current research. Explanations such as an alteration in the ratio of glutenin to gliadin, changes in the size of the glutenin polymers due to changes in the formation of disulfide bonds of the peptides, conformational changes affecting the folding and polymerization of the peptides, or changes caused by the direct influence of heat shock proteins on wheat dough quality are under consideration.<sup>13</sup> Therefore, a rapid, NIR technique for plant stress indicators would be useful in wheat breeding programs outside of the temperate regions of the world, where certain genotypes may be particularly susceptible to temperature extremes.

The success of NIR reflectance calibrations such as those of partial least-squares (PLS) design is often dependent on the structuring of the spectral data before the actual decomposition and regression procedure begins. This structuring, or pretreatment, serves several purposes: (1) removal of random noise, (2) reduction of the physical effect of sample-to-sample variation in scatter, caused by differences in particle size distributions, and (3) enhancement of a weak absorption band that is either intrinsically, or by instrument limitations, convoluted with neighboring spectral data. Unfortunately, the best pretreatment is seldom known beforehand, which is especially true when developing a calibration for a primary procedure or instrument (e.g., viscosity, dough resistance) rather than the concentration of a known chemical species. Spectral analysts have traditionally combined their prior knowledge of calibration development with educated guesswork in an *ad hoc* manner in an effort to find the optimal pretreatment procedure. The present research has taken a new approach, this being the systematic analysis of the effect of spectral pretreatments on wheat protein content and quality calibrations. Our study objective has been to determine the sensitivity of NIR spectral pretreatments, such as smooths, derivatives, and scatter correction, on four wheat properties. These properties range from that which is chemically well characterized (protein content), to that which is more difficult to chemically define (SDS sedimentation volume), to those that are proxies (hours above or below critical temperatures during plant growth development) for biochemically ill-defined constituents that affect wheat quality. The ease with which such trials are performed in a full (one-sample-out) PLS1 cross-validation is afforded by a custom-designed batch computer program that possesses the capability of analyzing hundreds of pretreatment combinations in one unattended run.

## EXPERIMENTAL

**Wheat.** Two consecutive years of growth trial samples from the Nebraska Winter Wheat Variety Tests program were obtained. For each season, ten commercial cultivars or advanced breeding lines of hard red winter wheat and ten cultivars/lines of hard white wheat were grown in field-replicated plots at each of ten non-irrigated sites throughout Nebraska. With a replication factor of two, approximately 400 laboratory samples per year were

available for reference and NIR analyses. Growing location from year 1 to year 2 tended to be the same at the county level but not at the field level. The cultivars/lines were consistent between the years for eight of the ten red wheats and seven of the ten white wheats. Fertilizer (N, P, K) was applied according to standard practices at levels commensurate with soil fertilization needs. Though the Tests program includes sites that possess irrigation, all sites chosen for the current study were dryland sites. Weather information (hourly temperature, humidity, and precipitation) was obtained from the Nebraska Automated Weather Data Network, High Plains Regional Climate Center, University of Nebraska–Lincoln (<http://hprcc.unl.edu/data.htm>). For this study, only temperature information translated into the number of hours above or below critical values (<24 °C or >32 °C) during the growth stage of grain fill was used in the NIR analysis of weather data.

**Protein Analysis.** Measurement of SDS sedimentation volume was accomplished via a modification of AACC Approved Method 56–70.<sup>14</sup> Briefly, a constant mass (typically 2 g, 14% moisture, wet basis) of ground (passing a 0.5 mm screen in a cyclone grinder) wheat is added to a prescribed volume (25 mL) of distilled water in a graduated cylinder, whereupon the cylinder is agitated for several minutes. An equal volume of solution containing 3.0% (w/w) sodium dodecyl sulfate and 2.0% (w/w) 1.2 N lactic acid stock solution is added to the cylinder, and then the mixture is agitated continuously for several minutes before being allowed to rest in the vertical position for 20 min. At the end of this period, the sediment volume is measured. The error of the SDS sedimentation procedure, calculated as the standard deviation of 30 measurements of a control that was measured during a one-month period, was 2.2 mL.

Protein content ( $N \times 5.7$ ) was determined by combustion (Model FP-428, Leco Corp., St. Joseph, MI) on duplicate 150-mg portions of each laboratory sample. Duplicate values were averaged. The error of the combustion procedure, also calculated as the standard deviation of single determinations of a control run in quadruplicate at the beginning and end of each of 11 analysis days (=88 measurements) throughout a one-month period, was 0.109% protein.

**Near-Infrared Acquisition and Calibration Equation Development.** To reduce spectral variation attributed to moisture, test samples were conditioned to a constant relative humidity of 33% by placement<sup>15</sup> in batches of 20 in a desiccator containing saturated  $MgCl_2$  solution. NIR reflectance (1100–2498 nm) values were recorded at 2-nm increments using an analytical scanning monochromator (Foss-NIRSystems Model 6500, Silver Spring, MD) equipped with a rotating sample cup. Each test sample was scanned twice, with repacking between scans. A scan was defined as the average of 32 repetitive passes referenced to the average of an equal number of passes of ceramic. Scan averages [ $\log(1/R)$ ] were the basis of input in calibration equation development.

Partial least-squares (PLS) regression on mean-centered data without variance scaling was performed independently on each of four properties: protein content, SDS sedimentation volume,  $t(T < 24\text{ }^\circ\text{C})$ , and  $t(T > 32\text{ }^\circ\text{C})$ . One-sample-out cross-validation was performed on

one-half the samples (i.e., the first field rep) from each year ( $n = 198$  and  $196$  for years 1 and 2, respectively), treating each year separately and then in combination. The remaining samples ( $n = 200$  and  $195$  for years 1 and 2, respectively) became the test sets. PLS calibration development was performed using two packages: SAS (specifically Proc PLS, Cary, NC) and Grams PLSplus (Galactic Industries, Salem, NH). The first package was used to evaluate the effect of spectral pretreatments (smoothing, derivatives, and scatter or pathlength correction). A sequence of SAS data steps and procedures were written and placed within nested loop structures controlled by the SAS Macro language. Pretreatment combinations could be run in one unattended session in batch mode, as described in detail in a recent article.<sup>16</sup> The second package was used for detailed examination of individual cross-validations and in calibration equation testing. Up to 15 factors were evaluated for each PLS regression, with the optimal number of factors decided by an F-test at a probability level of 0.75.<sup>17</sup>

Spectral pretreatments for minimizing particle distribution variation caused by grinding or packing density consisted of multiplicative scatter (or signal) correction (MSC)<sup>18,19</sup> or standard normal variate (SNV) transformation.<sup>20</sup> Each of these transformations attempts to eliminate additive and multiplicative differences between spectra that are caused by variation in the physical size of particles (hence, causing variation in scattering and absorption) rather than an actual variation in chemical concentration. Multiplicative scatter correction accomplishes this by regressing each spectrum to a common spectrum, typically the mean spectrum of a calibration set, then applying a correction to each wavelength according to the coefficients of the regression equation. It is assumed that the beneficial aspect of minimizing the effect of physical variation outweighs the deleterious effect of inadvertently removing any spectral dependencies on the concentration of the analyte during the MSC procedure. In contrast to MSC, SNV operates on each spectrum independently by normalizing the spectrum to have a standard deviation of unity and a mean of zero across its wavelength region.

Other spectral pretreatments included a running mean smooth and a Savitzky–Golay second derivative. These transformations were applied with or without the MSC and SNV transformations, as shown by the flowchart in Fig. 1. Eleven convolution window sizes, ranging from 5 points to 25 points, were employed in either the smooth or the derivative operation. Coefficients for the second-derivative convolution of a quadratic polynomial were extracted from the original paper of Savitzky and Golay.<sup>21</sup> When an MSC or SNV operation was used in conjunction with the smooth or derivative operation, the order of the two operations depended on the particle size correction operation. With MSC, differentiation and smoothing occurred first; with SNV, the order was reversed. Although the ordering was chosen to mimic that of Grams PLSplus (with the intention of using the graphical features of the commercial software once the best pretreatment had been identified), recent research on the application of MSC and second derivatives in single wheat kernel protein content calibrations has reported the superiority of the derivative-before-MSC approach over

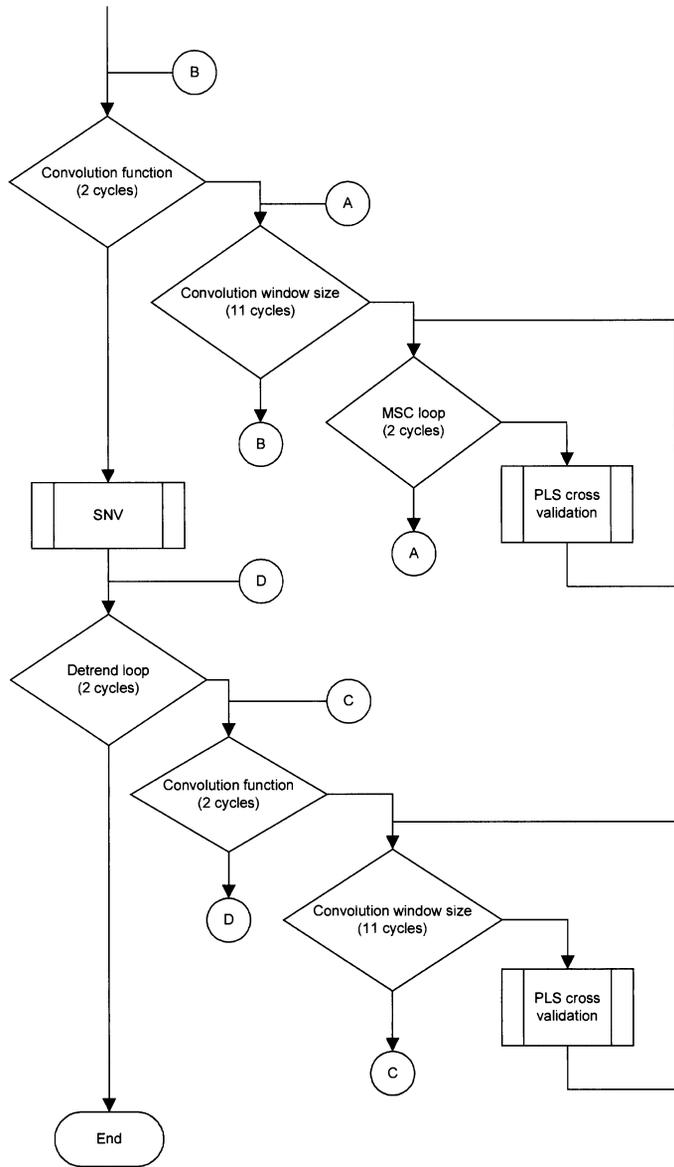


Fig. 1. Flowchart of nested loop structure of spectral pretreatments. (Flowchart symbols: diamond = SAS Macro loop; rectangle with two vertical bars = SAS procedure; circle = connector.)

that of the opposite order.<sup>22</sup> For trials that employed SNV, detrending was subsequently applied to one half the number of trials. As initially reported by Barnes et al.,<sup>20</sup> detrending consists of the least-squares fitting of a quadratic polynomial to each SNV-corrected spectrum, whereupon a new spectrum is formed as the difference between the SNV-corrected spectrum and the polynomial.

The total number ( $M$ ) of cross-validation trials examined for each calibration set was as follows:  $M = \text{none} + \text{only MSC} + [2 \text{ convolution function types (i.e., smooth or second derivative)} \times 11 \text{ convolution window sizes (i.e., 5, 7, \dots, 25)} \times 2 \text{ MSC options (i.e., MSC or no MSC)}] + \text{SNV alone} + \text{SNV with detrend} + [2 \text{ detrend-after-SNV options (i.e., detrend or no detrend)} \times 2 \text{ convolution function types} \times 11 \text{ convolution window sizes}] = 92 \text{ trials.}$

For each trial, cross-validation performance was deter-

mined as the square root of the mean square error of cross-validation (RMSECV), defined as:

$$\text{RMSECV} = \left[ \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N} \right]^{1/2} \quad (1)$$

where  $\hat{y}_i$  is the predicted value of the property of each sample  $i$  as it is first rotated out for development of a regression equation and subsequently predicted during a cross-validation cycle;  $y_i$  is the reference value; and  $N$  is the number of samples in the calibration set.

The optimal number of PLS factors for each trial was determined by an F-test.<sup>17</sup> Specifically, a variance ratio was constructed for each number of PLS factors by dividing the predicted residual error sum of squares (PRESS) by the minimum PRESS value, shown as follows:

$$F_j = \frac{\text{PRESS}_j}{\text{PRESS}_{\min}} = \frac{\sum_{i=1}^N (\hat{y}_{i,j} - y_i)^2}{\sum_{i=1}^N (\hat{y}_{i,\min} - y_i)^2} \quad (2)$$

where  $\hat{y}_{i,j}$  and  $\hat{y}_{i,\min}$  are the cross-validation predictions (see Eq. 1) corresponding to the  $j$ th ( $1 \leq j \leq \min \leq 15$ ) factor and the factor (min) that produces the smallest PRESS value, respectively. With the numerator and denominator degrees of freedom set to  $(N - 1)$ , the optimal number of factors is the smallest  $j$  for which the probability,  $P$ , that an observation from an F distribution being less than or equal to  $F_j$  is smaller than the prescribed value of 0.75. Depending on the calibration conditions, the optimal number is sometimes the number corresponding to the smallest PRESS value, and in rarer circumstances, it can be the largest number of factors examined (15 in the present case). Once the optimal number of factors was determined for each cross-validation trial, the coefficient of determination ( $R^2$ ) was determined by correlating the predictions from the corresponding PLS calibration equation based on all calibration samples to reference values for these samples.

In addition to reporting the individual statistics of each PLS cross-validation trial, these trials were ordered by RMSECV value, whereupon a grouping of  $x$  contiguous trials from the better end formed the basis of averages for the following figures of merit: RMSECV,  $R^2$ , the optimum number of PLS factors, the number of factors at PRESS minimum, and the number of convolution points used in the smooth or second-derivative operation. The procedure was:

$$\begin{aligned} & \overline{\text{figure of merit}} \\ &= \frac{1}{x} \sum_{i=j}^{j+x-1} \{\text{figure of merit of an ordered trial}\}_i \quad (3) \end{aligned}$$

This procedure (with  $x = 30$  and  $j = 6$  in the present case) is utilized for the purpose of representing the realistic modeling power of the NIR PLS calibrations for each property, thus minimizing the possibility of reporting spurious (high or low) calibration statistics.

## RESULTS AND DISCUSSION

Our previous research on the analysis of the Year 1 data indicated a high correlation between SDS sedimentation volume and protein content ( $r = 0.903$ ).<sup>11</sup> When a similar analysis is performed on Year 2, the degree of correlation is much smaller ( $r = 0.243$ ,  $N = 195$  test samples), which demonstrates that while protein content does indeed have a strong effect on the quality-related properties of wheat, environmental effects such as those brought on by yearly changes in weather are very important in moderating this influence. Two common techniques have been used to isolate the contribution of protein content from a wheat quality property: partial correlation analysis, as explained in Fisher and van Belle<sup>23</sup> and applied by Delwiche et al.,<sup>10</sup> and curve fitting of estimated component spectra.<sup>9,24-25</sup> Both techniques have shown some degree of success, though practitioners of each acknowledge the difficulty of spectrally isolating protein-related quality features from protein content, whose molecular groups have numerous overlapping combination and first and second overtone vibrations of CH, OH, and NH throughout the 1100 to 2500 nm region.<sup>26</sup>

Recent research suggests that elevated temperatures alter the timing of gluten protein gene activation; however, contrary to early research,<sup>27</sup> the relative proportions of glutenin to gliadin are largely unaffected.<sup>28</sup> In the present study, yearly differences in weather were demonstrated by the fact that Year 2 was more moderate, having a range of low-temperature ( $<24^\circ\text{C}$ ) hours that was shorter and having cooler high temperatures ( $>32^\circ\text{C}$ ) as well (Table I). This may also have contributed to the shorter range in SDS sedimentation volume values for Year 2, despite the lack of correlation between  $t(T < 24^\circ\text{C})$  and either SDS sedimentation volume or protein content ( $r = 0.002$  and  $-0.085$ , respectively,  $N = 195$ ). For  $t(T > 32^\circ\text{C})$ , a positive significant correlation ( $r = 0.380$ ,  $P < 0.001$ ) existed with protein content, while a negative significant correlation ( $r = -0.288$ ,  $P < 0.001$ ) existed with SDS sedimentation volume.

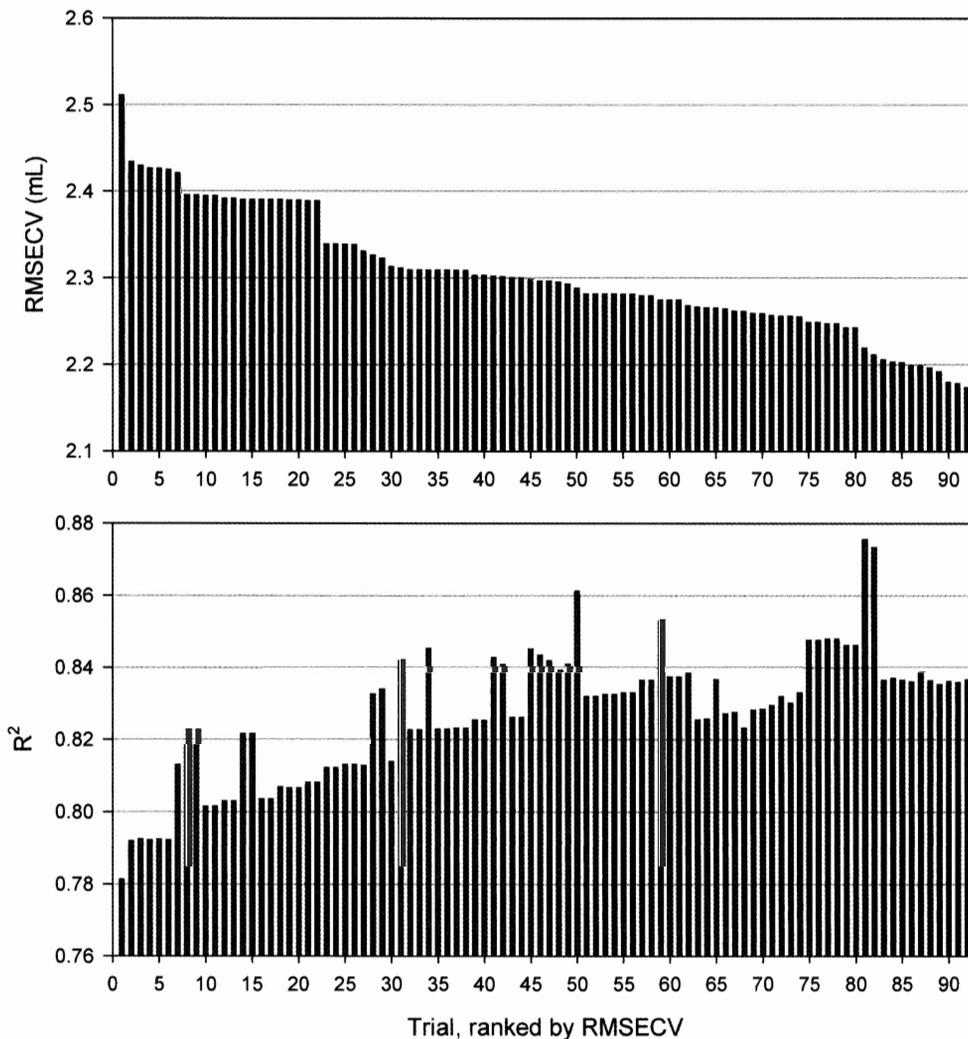
**Effect of Spectral Pretreatment on Partial Least-Squares Calibrations.** By way of example, the RMSECV and corresponding  $R^2$  values are shown for all pretreatment trials of the PLS calibrations for SDS sedimentation volume on Year 1 data (Fig. 2). When ordered by RMSECV, the pretreatment trial results typically demonstrate one to three pretreatments with particularly high error, followed by a very gradual improvement (i.e., lowering) of error throughout the remaining portion of the 92 trials. Unlike the difference between the poorest and the low end of the intermediate trials, the difference between the best and the upper end of the intermediate trials is generally not as large, as demonstrated by the calibrations for SDS sedimentation. Interestingly, the trend toward improved correlation between measured and modeled property, as gauged by  $R^2$ , was not always in parallel with the improvement in RMSECV. For example, the trial with the third-highest  $R^2$  value for SDS sedimentation volume was only intermediate in terms of RMSECV, ranking 43rd best (i.e., trial 50 in Fig. 2). This phenomenon was typical for the other properties regardless of year, which suggests that a strong reliance on  $R^2$  as an

**TABLE I. Properties under study.**

Property	Calibration sets <sup>a</sup>		Test sets <sup>b</sup>	
	Range	Mean $\pm$ std	Range	Mean $\pm$ std
Protein content (%)				
<i>year 1</i>	9.09–18.27	13.35 $\pm$ 2.21	8.13–18.90	12.98 $\pm$ 2.26
<i>year 2</i>	10.61–18.54	13.92 $\pm$ 1.68	10.31–19.22	13.72 $\pm$ 1.77
<i>both years</i>	9.09–18.54	13.64 $\pm$ 1.98	8.13–19.22	13.34 $\pm$ 2.06
SDS sed. vol. (mL)				
<i>year 1</i>	10–35	21.2 $\pm$ 5.2	9–33	20.4 $\pm$ 5.6
<i>year 2</i>	14–29	19.6 $\pm$ 3.2	12–30	19.4 $\pm$ 3.4
<i>both years</i>	10–35	20.4 $\pm$ 4.4	9–33	19.9 $\pm$ 4.6
<i>t</i> ( <i>T</i> < 24 °C) (h)				
<i>year 1</i>	792–1099	965.5 $\pm$ 89.3	792–1099	965.7 $\pm$ 88.9
<i>year 2</i>	419–715	606.8 $\pm$ 100.8	419–715	606.7 $\pm$ 101.1
<i>both years</i>	419–1099	787.0 $\pm$ 203.2	419–1099	788.5 $\pm$ 203.3
<i>t</i> ( <i>T</i> > 32 °C) (h)				
<i>year 1</i>	33–114	72.8 $\pm$ 23.9	33–114	72.8 $\pm$ 24.1
<i>year 2</i>	21–92	65.8 $\pm$ 20.7	21–92	65.6 $\pm$ 20.8
<i>both years</i>	21–114	69.3 $\pm$ 22.6	21–114	69.2 $\pm$ 22.8

<sup>a</sup>*N* = 198, 196, and 394 samples for year 1, year 2, and both years, respectively.

<sup>b</sup>*N* = 200, 195, and 395 samples for year 1, year 2, and both years, respectively.



**FIG. 2.** PLS one-sample-out cross-validation results for SDS sedimentation volume for Year 1 samples (*N* = 198). Trials are placed in order of decreasing RMSECV. The cross-validation (upper graph) for each trial was determined at the optimal number of PLS factors by an F-test. The correlation (lower) graph bars are arranged in the same order as the upper graph.

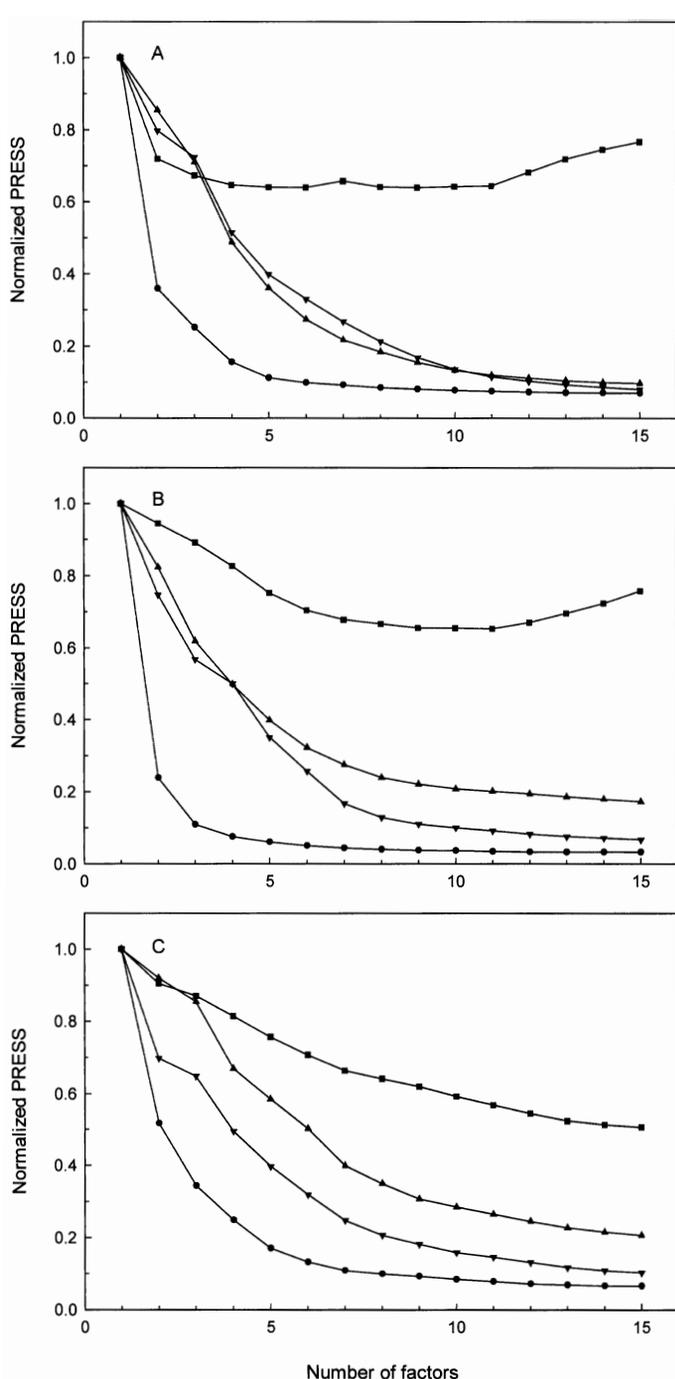


FIG. 3. Average (of 30 pretreatment trials) cross-validation error as a function of the number of PLS factors, normalized to the error at factor 1. (●) protein content; (■) SDS sedimentation volume; (▼)  $t(T < 24\text{ }^{\circ}\text{C})$ ; and (▲)  $t(T > 32\text{ }^{\circ}\text{C})$ . Graphs A, B, and C correspond to Year 1, Year 2, and combined years, respectively.

indicator of calibration performance during cross-validation is not advised.

Aside from the ability to compare the effect of spectral pretreatments, the large number of pretreatment trials has made it possible to more clearly define the performance of the PLS model for a given constituent or property. The number of recommended PLS factors by cross-validation, be it by selecting the number corresponding to the lowest PRESS or by a statistical test that measures the significance of adding factors, can often vary by several units.

TABLE II. Summary of 30-pretreatment-trial averages of leave-one-out PLS1 cross-validations of four wheat quality properties.

Property	RMSECV	$R^2$	Optimal number of factors	Number of factors at PRESS minimum	Number of convolution points
Protein content (%)					
year 1	0.1014	0.999	11.13	14.30	19.40
year 2	0.1111	0.997	10.67	13.67	19.53
both years	0.1029	0.998	12.77	14.87	18.73
SDS sedimentation volume (mL)					
year 1	2.25	0.839	4.56	8.63	14.13
year 2	2.30	0.649	8.00	9.60	16.00
both years	2.40	0.791	12.70	14.87	20.93
Time at $T < 24\text{ }^{\circ}\text{C}$ (h)					
year 1	22.17	0.981	13.60	14.67	11.00
year 2	24.54	0.982	13.30	14.93	11.93
both years	44.78	0.974	14.13	15.00	11.80
Time at $T > 32\text{ }^{\circ}\text{C}$ (h)					
year 1	5.57	0.978	12.43	14.03	12.53
year 2	7.99	0.949	11.97	14.43	12.40
both years	9.06	0.913	13.57	15.00	13.33

The ramification of this is that the often-published graphs that show cross-validation error (e.g., PRESS) as a function of the number of PLS factors are overly specific to the pretreatment procedure and therefore cannot describe the general capability of the PLS model for the constituent or property itself. To alleviate this problem, averaging was performed on the performance statistics of 30 (i.e., ranks 6–35 of RMSECV) pretreatments (Fig. 3, Table II). Apparent from Fig. 3 is a difference among the properties being modeled by PLS regression. With the PRESS values normalized to the PRESS at factor 1 (i.e.,  $\text{PRESS}_{\text{factor } i} / \text{PRESS}_{\text{factor } 1}$ ) for each property, it is seen that calibration equation improvement during the progression of the first several factors was greatest for protein content, irrespective of whether a single year's analysis or the combined years' analysis was used. Beyond five factors, improvement in the protein content calibrations was comparatively slight, though significant, with the average optimal number approximately 11 for single-year calibrations and 13 for combined-year calibrations (Table II). In contrast, the improvement in SDS sedimentation volume calibration equation performance (with respect to the RMSECV at factor 1) was not as pronounced as that for protein content. For the single-year trials, SDS sedimentation volume calibrations worsened after a certain number of factors (averages of 4.6 and 8.0 for Years 1 and 2, respectively), as demonstrated by the rise in PRESS values beyond these minima, suggesting that over-fitting occurred with higher-factor calibrations. The decline in average  $R^2$  value of Year 2 (Table II) compared to that of Year 1 (0.649 vs. 0.839) is thought to be caused by the overall smaller range in SDS sedimentation volume values for Year 2, as well as the weaker relationship between this property and protein content. Therefore, in lieu of a partial correlation analysis,<sup>9</sup> the average statistics for Year 2 may be more realistic than those for Year 1 in assessing NIR spectroscopy's ability to measure wheat protein quality apart from protein quantity.

In contrast to protein content and SDS sedimentation volume, the calibrations for the two proxy weather prop-

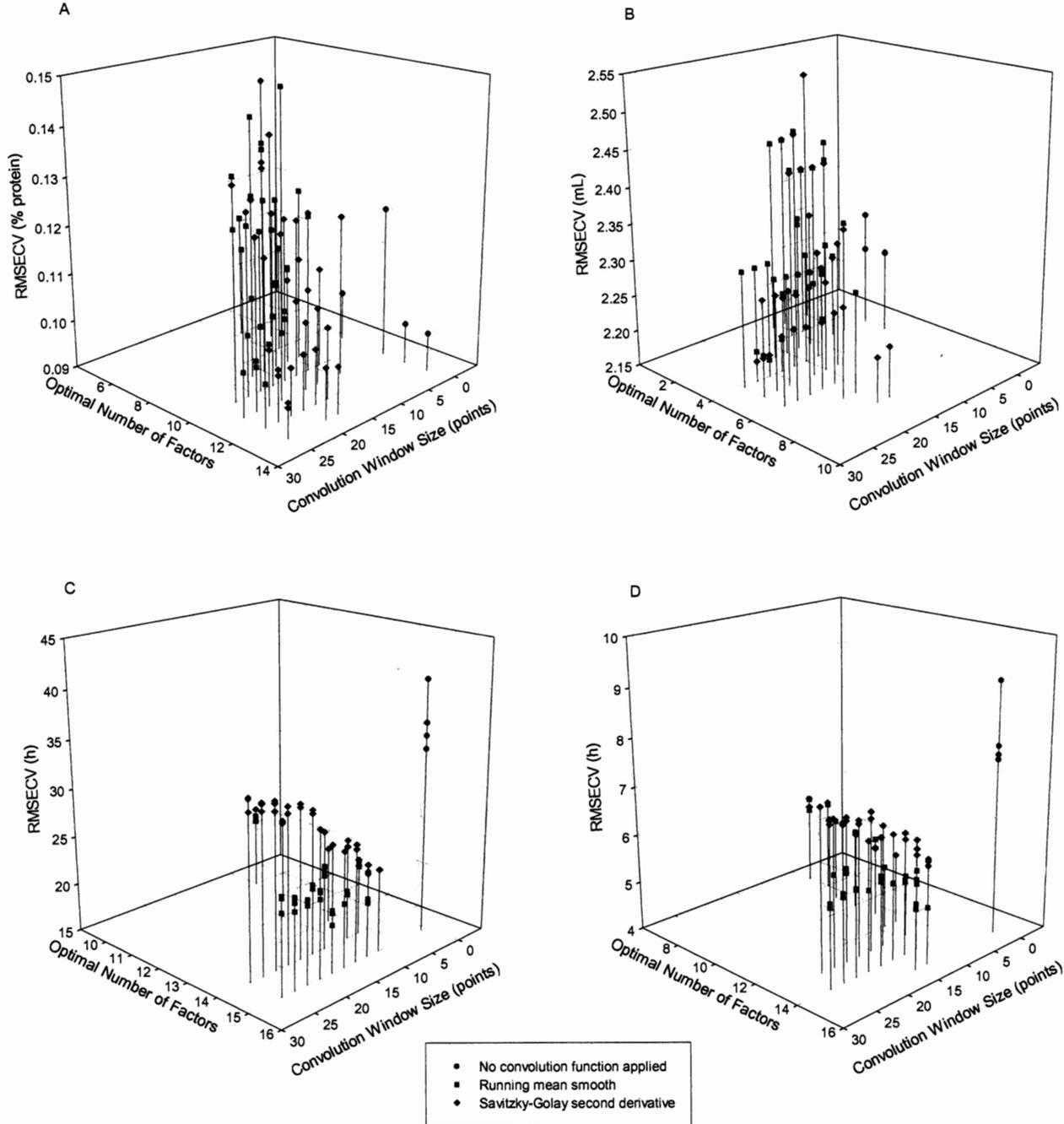


FIG. 4. The effect of spectral pretreatment on cross-validation error; Year 1 RMSECV ( $N = 198$ ) values for (A) protein content, (B) SDS sedimentation volume, (C)  $t(T < 24\text{ }^{\circ}\text{C})$ , and (D)  $t(T > 32\text{ }^{\circ}\text{C})$ .

erties [ $t(T < 24\text{ }^{\circ}\text{C})$  and  $t(T > 32\text{ }^{\circ}\text{C})$ ] continued to improve with increase in the number of PLS factors, even out to 15 factors (Fig. 3). The average optimal number of factors was similar between these two properties as well as similar among the single- and combined-year trials, with an approximate range of 12 to 14 factors (Table II).

**Trends in Spectral Pretreatments.** As demonstrated in the ordered ranking of PLS RMSECVs for SDS sedimentation volume in Fig. 2, the choice of spectral pretreatment had a marked effect on calibration error. The relationships among the convolution function (second derivative, smooth, or none), the size of the convolution

window, the optimal number of PLS factors, and their effect on calibration error are shown for Year 1 samples in Fig. 4 and Years 1 and 2 combined in Fig. 5 (Year 2 plots are not shown due to their similarity to Year 1). Protein content calibrations tended to favor wide ( $>20$  point) convolution window sizes, regardless of single year or multiple year analysis. However, even the use of no convolution function produced acceptable results for protein content, as shown by two of the four non-smoothed and non-derivatized trials in each year group (Figs. 4A and 5A, plotted as convolution window size of one point) having RMSECV values of approximately 0.1%. With convolution, the second-derivative and run-

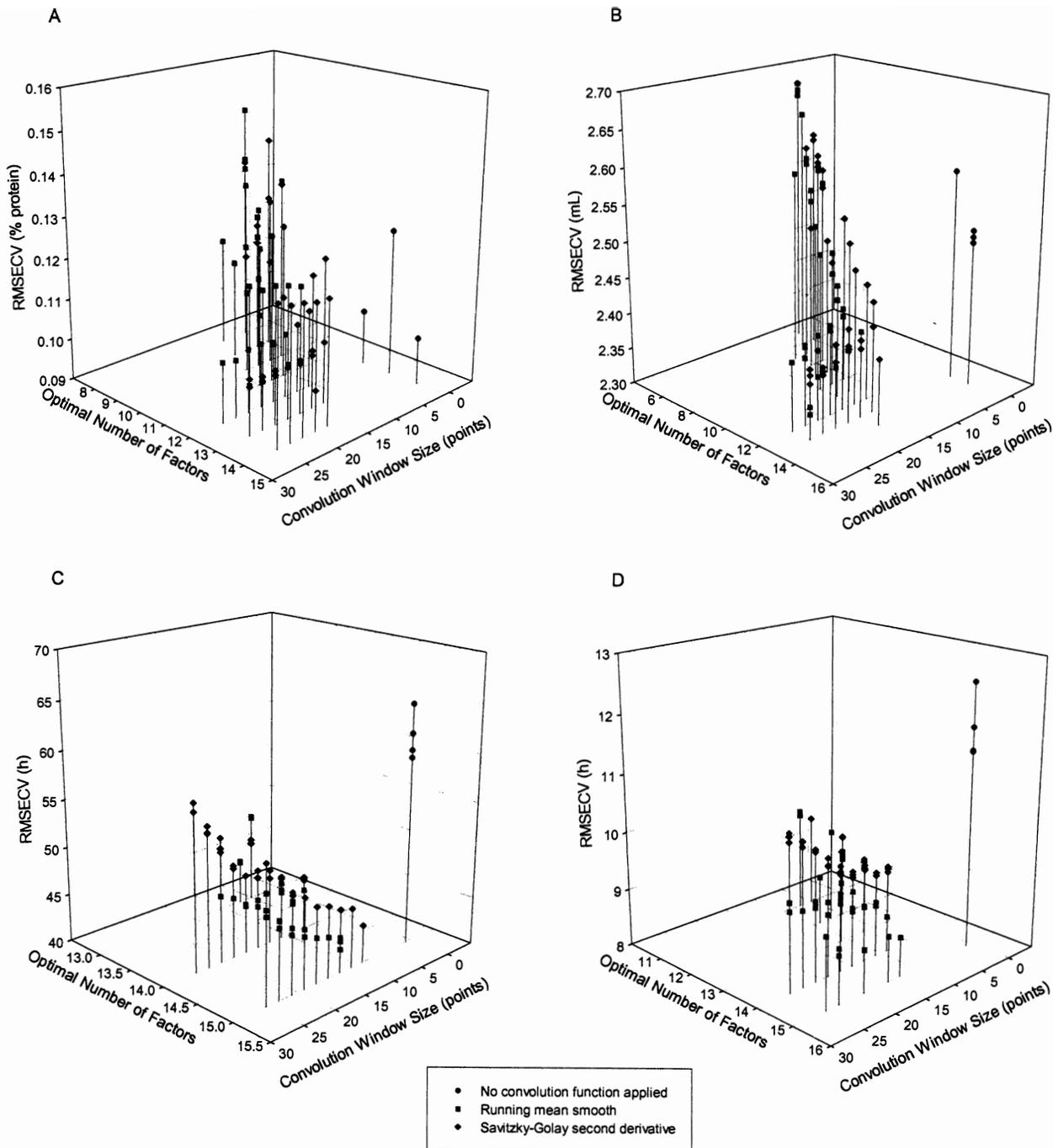


FIG. 5. The effect of spectral pretreatment on cross-validation error; Combined year (1 and 2) RMSECV ( $N = 394$ ) values for (A) protein content; (B) SDS sedimentation volume; (C)  $t(T < 24\text{ }^{\circ}\text{C})$ ; and (D)  $t(T > 32\text{ }^{\circ}\text{C})$ .

ning mean pretreatments demonstrated equivalent overall levels of performance. An increase in convolution window size usually caused an increase in the optimal number of PLS factors.

For SDS sedimentation volume (Fig. 4B for Year 1 and Fig. 5B for the combined years), the benefit of a running mean or second-derivative transformation is more apparent than in the calibrations for protein content. Without such pretreatments, the combined years RMSECV values were in excess of 2.50 mL, compared with <2.35 mL for the best trials that used a convolution function for smoothing or derivatization. The relationship

between the size of the smooth or second-derivative convolution window and the optimal number of factors is also more apparent for SDS sedimentation volume than protein content, especially when both years are combined (Fig. 5B). Generally, as the window width increased, the optimal number of PLS factors also increased, while the cross-validation error decreased. Again, the actual choice of convolution function (smooth or second derivative) was not as important as the size of the convolution window.

For the two time-temperature properties [ $t(T < 24\text{ }^{\circ}\text{C})$  and  $t(T > 32\text{ }^{\circ}\text{C})$ ], use of a smoothing or second-deriv-

ative transformation always produced superior calibrations compared to calibrations that did not utilize such transformations (Figs. 4C and 4D for Year 1; Figs. 5C and 5D for the combined years). The relationship between the convolution window size and the optimal number of factors for these two properties was not as apparent as that for protein content or SDS sedimentation volume, perhaps because the optimal number of factors was usually greater than 10. The unique aspect of these properties is the tendency for the smoothing pretreatment to produce slightly smaller RMSECV values than the second-derivative pretreatments.

**Validation of Selected Partial Least-Squares Calibrations.** When applied to the test set samples, the level of performance of the PLS calibrations was dependent on the property. When the best second-derivative calibration, with or without MSC or SNV, for each property and year was applied to the corresponding test set, the degree of fit was seen to be primarily dependent on the property and, to a lesser extent, on the year(s) under consideration (Figs. 6A–6D). Among the four properties studied, protein content calibrations were most accurate, with little difference among the accuracies of the Year 1, Year 2, and combined year calibrations ( $r^2 = 0.998, 0.996, \text{ and } 0.997$ ; SEP = 0.096, 0.105, and 0.106%; bias = 0.005, -0.008, and -0.006%, respectively). SDS sedimentation volume calibrations had a much greater variation in performance across years, with Year 2 ( $r^2 = 0.551$ , SEP = 2.3 mL, bias = 0.6 mL) being poorer than either Year 1 ( $r^2 = 0.861$ , SEP = 2.1 mL, bias = 0.2 mL) or Years 1 and 2 combined ( $r^2 = 0.796$ , SEP = 2.1 mL, bias = 0.2 mL). This poorer performance for Year 2 is attributed to the weak relationship between SDS sedimentation volume and protein content that was mentioned earlier. The calibrations for  $t(T < 24^\circ\text{C})$  benefited from the combining of the yearly data sets when viewed from the standpoint of the goodness of fit ( $r^2 = 0.953$  for the Years 1 and 2 combined vs. 0.919 and 0.914 for the separate-year trials), in which the overall range in  $t(T < 24^\circ\text{C})$  doubled when the years were combined. However, when viewed from the standpoint of the SEP, error worsened with the combining of years (SEP = 26.6, 29.8, and 44.4 h; bias = 1.7, -3.4, and 1.5 h for Year 1, Year 2, and Years 1 and 2 combined, respectively), which suggests that the choice of broadening a data set should be made based on the future intended use of the ensuing calibration equation. The fact that there was a lack of correlation between protein content and  $t(T < 24^\circ\text{C})$  for each single year or the two years combined indicates that the NIR spectra of mature wheat is truly responsive to the environmental conditions of the developing plant. Lastly, for  $t(T > 32^\circ\text{C})$ , the performance for each year alone was equivalent ( $r^2 = 0.881$  and 0.854; SEP = 8.3 and 8.0 h; bias = -2.2 and -4.4 h for Years 1 and 2, respectively). However, when the years were combined, the performance decreased ( $r^2 = 0.789$ , SEP = 10.5 h, and bias = -3.1 h).

Identification of the near-infrared absorbers responsible for the ability of each property to be modeled is quite difficult. In our previous paper, we demonstrated through PLS2 scores analysis the relationships between historically identified absorption bands (e.g., starch O–H and C–O combination at 2100 nm, amide I and amide III combination band at 2180, and oil  $\text{CH}_2$  stretch–bend

combination band at 2306  $\text{nm}^{29}$ ) and a superset of the properties of the current research.<sup>9</sup> As a way of demonstrating the complexity of each property's PLS calibration equation, Fig. 7 depicts the regression coefficient vectors of a commonly structured PLS calibration (15 point Savitzky–Golay second derivative, followed by MSC) on each of the four properties examined, using the combined year calibration set. This spectral pretreatment regime was selected because of its general level of acceptable performance across all properties and years. Likewise, 13 factors were selected as the fixed number of PLS factors to permit property-to-property comparisons in what otherwise (if the number of factors was based on the optimal number for each property) would have ranged between 12 and 14 factors. Even the calibration equation for protein content (graph B) possesses a high degree of variation in its regression coefficients, which emphasizes the overlapped nature of absorption bands throughout the NIR wavelength region.

#### Implications of Pretreatment Searches and Trends.

At the onset of NIR calibration development for a new analyte or property, the spectroscopist is often unaware of the potential calibration performance. The following observations are drawn based on the NIR spectra of ground wheat: (1) The types of pretreatments necessary for good calibration performance are specific to the property. For example, when developing calibrations for protein content, reasonable equations were developed that possessed no spectral pretreatments, whereas for  $t(T < 24^\circ\text{C})$ , a lack of any spectral pretreatment resulted in poor SECVs. Secondly, the differences in overall shape of the four graphs of Fig. 4, and again in Fig. 5, demonstrate that the effect of spectral pretreatment on calibration error is specific to the property. (2) More important than the choice of the convolution function is the size of the convolution window. All four studied properties tended to prefer a wide (>15 point) convolution window. Future trials that utilize functions other than the running mean smooth and the Savitzky–Golay second derivative will be needed to corroborate this trend. (3) When developing PLS calibrations for wheat quantity or quality properties, a large number of factors (10–15) are preferable to a smaller number (<10). Additional work is needed to ensure that this trend holds true for independent test sets, such that the cross-validation procedure is not recommending high-factor PLS calibration equations that are inherently over-fitted.

## CONCLUSION

This study has demonstrated the application of a multitude of spectral pretreatment trials on the analysis of NIR reflectance of ground wheat for protein content and quality indicators. By systematically applying smooths, derivatives, and particle size correction procedures, then performing a full cross-validation to the PLS regression of each combination, the importance of spectral pretreatment to calibration performance was affirmed. The potential for calibration improvement with application of a pretreatment is dependent on the property. Protein content, a property that is easily and accurately modeled, produced calibrations whose accuracies were not affected as much by changes in spectral pretreatment as were

Year 1

Year 2

Years 1 &amp; 2

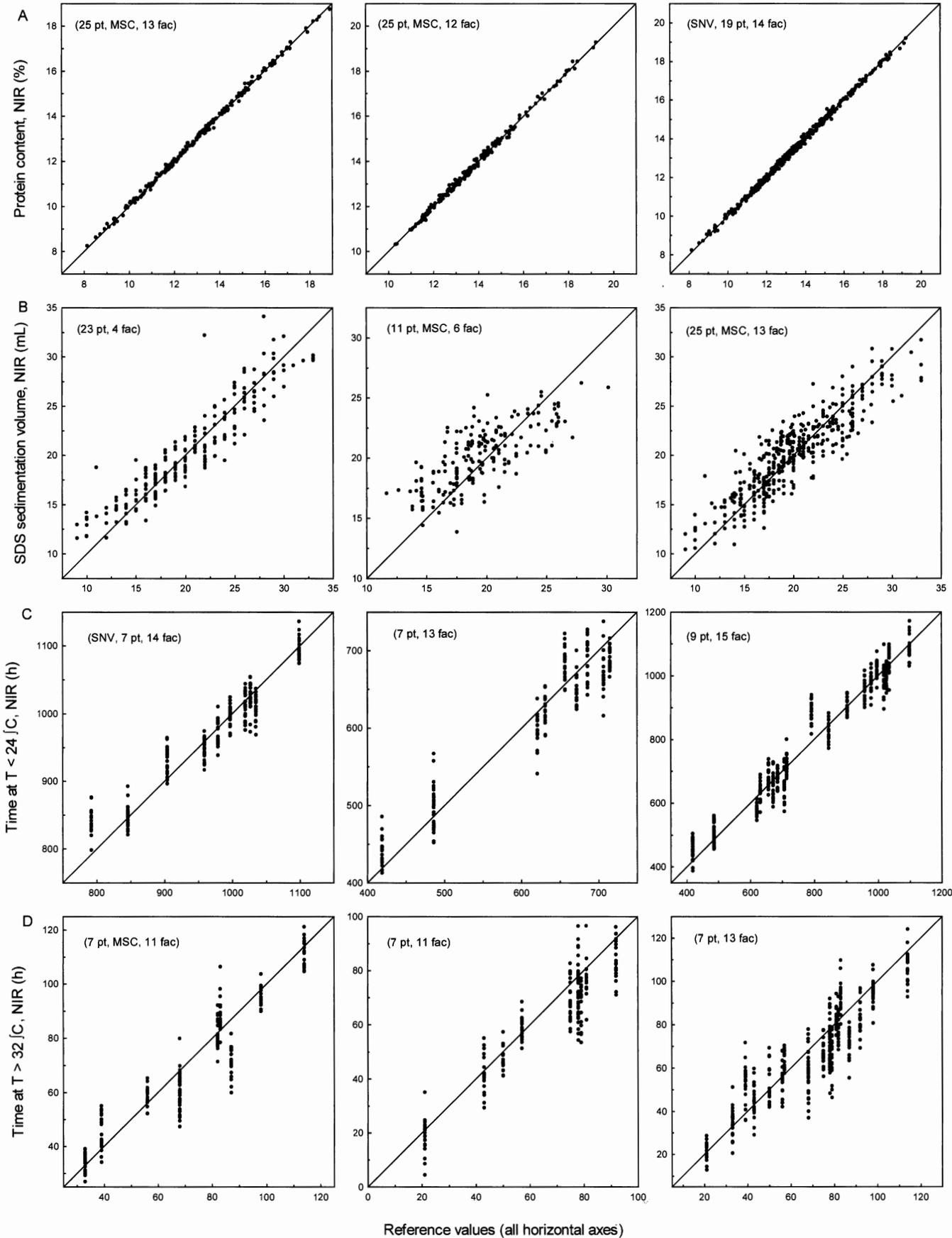


Fig. 6. NIR-predicted vs. reference values for the four studied properties. For each property  $\times$  year condition, the best second-derivative calibration with pretreatment conditions and number of PLS factors in parentheses) as determined by cross-validation is applied to the test samples ( $N = 200, 95,$  and  $395$  for Year 1, Year 2, and both years, respectively). The  $45^{\circ}$  line is included to represent the ideal calibration.

We wish to thank Joyce Shaffer (USDA-ARS, Beltsville) for collection of spectral data and combustion analysis, Richard Samson (USDA-ARS, Lincoln) for SDS sedimentation volume analyses, and Greg Dorn, John Eis, Bob Klein, Dave Baltensperger, Steve Baenziger, Chris Hoagland, and Glen Frickle for their contributions to the Nebraska Wheat Statewide Variety Trial.

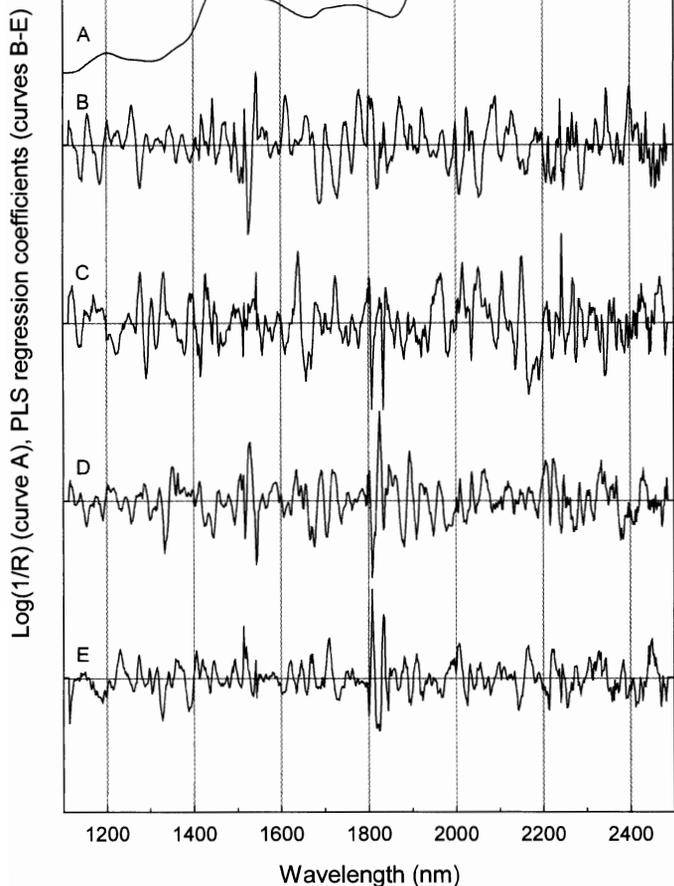


FIG. 7. Average spectrum for the calibration samples of the combined year set plotted above the PLS regression coefficient vectors for the four properties studied. Calibration conditions for all properties are as follows: 15 point Savitzky–Golay second derivative, followed by multiplicative scatter correction, 13 PLS factors [A = average, B = protein content, C = SDS sedimentation volume, D =  $t(T < 24\text{ }^\circ\text{C})$ , and E =  $t(T > 32\text{ }^\circ\text{C})$ ].

those for SDS sedimentation volume, a wheat quality property. The present study demonstrated that by averaging the performance indicators of a PLS regression cross-validation (e.g., RMSECV) across many pretreatment trials (30 in this case), a better indication of the calibration potential of NIR is obtained. Reliance on the  $R^2$  value of a calibration equation should be avoided as this value does not ensure the lowest standard error of the residuals. By use of two growth seasons of wheat, it was determined that the degree of correlation between protein content and quality (SDS sedimentation volume) is strongly influenced by environmental factors, such as variation in weather. Lastly, the proxy weather properties themselves, such as  $t(T < 24\text{ }^\circ\text{C})$ , which may have a very weak correlation with protein content, may still produce reasonable NIR calibration equations.

1. F. MacRitchie, "Physicochemical Properties of Wheat Proteins in Relation to Functionality", in *Advances in Food and Nutrition Research*, J. E. Kinsella, Ed. (Academic Press, San Diego, CA, 1992), Vol. 36, pp. 2–87.
2. C. J. Peterson, R. A. Graybosch, P. S. Baenziger, and A. W. Grombacher, *Crop Sci.* **32**, 98 (1992).
3. A. H. Bloksma and W. Bushuk, "Rheology and Chemistry of Dough", in *Wheat Chemistry and Technology*, Y. Pomeranz, Ed. (Am. Assoc. Cereal Chem., St. Paul, MN, 1988), 3rd ed., Vol. 2, Chap. 4, pp. 131–217.
4. N. K. Singh, G. R. Donovan, I. L. Batey, and R. MacRitchie, *Cereal Chem.* **67**, 150 (1990).
5. N. K. Singh, G. R. Donovan, and R. MacRitchie, *Cereal Chem.* **67**, 161 (1990).
6. R. B. Gupta, Y. Popineau, J. Lefebvre, M. Cornec, G. J. Lawrence, and F. MacRitchie, *J. Cereal Sci.* **21**, 103 (1995).
7. R. A. Graybosch, C. J. Peterson, P. S. Baenziger, and D. R. Shelton, *J. Cereal Sci.* **22**, 45 (1995).
8. P. J. Stone, P. W. Gras, and M. E. Nicolas, *J. Cereal Sci.* **25**, 129 (1997).
9. I. J. Wesley, O. Larroque, B. G. Osborne, N. Azudin, H. Allen, and J. H. Skerritt, *J. Cereal Sci.* **34**, 125 (2001).
10. S. R. Delwiche, R. A. Graybosch, and C. J. Peterson, *Cereal Chem.* **75**, 412 (1998).
11. S. R. Delwiche, R. A. Graybosch, L. A. Nelson, and W. R. Hruschka, *Cereal Chem.* **79**, 885 (2002).
12. P. J. Randall and H. J. Moss, *Austr. J. Agric. Res.* **41**, 603 (1990).
13. C. Blumenthal, P. J. Stone, P. W. Gras, F. Bekes, B. Clarke, E. W. R. Barlow, R. Appels, and C. W. Wrigley, *Cereal Chem.* **75**, 43 (1998).
14. AACC, *Approved Methods of the AACC* (Am. Assoc. Cereal Chem., St. Paul, MN, 2000), 10th ed., Method 56–70.
15. L. Greenspan, *J. Res. Nat. Bur. Stand.* **81A**, 89 (1977).
16. J. B. Reeves III and S. R. Delwiche, "SAS Partial Least Squares Regression for Analysis of Spectroscopic Data", *J. Near Infrared Spectrosc.*, paper in press (2003).
17. D. M. Haaland and E. V. Thomas, *Anal. Chem.* **60**, 1193 (1988).
18. P. Geladi, D. McDougall, and H. Martens, *Appl. Spectrosc.* **39**, 491 (1985).
19. H. Martens and T. Næs, *Multivariate Calibration* (John Wiley and Sons, Chichester, UK, 1989), p. 345.
20. R. J. Barnes, M. S. Dhanoa, and S. J. Lister, *Appl. Spectrosc.* **43**, 772 (1989).
21. A. Savitzky and M. J. E. Golay, *Anal. Chem.* **36**, 1627 (1964).
22. D. K. Pedersen, H. Martens, J. P. Nielsen, and S. B. Engelsen, *Appl. Spectrosc.* **56**, 1206 (2002).
23. L. D. Fisher and G. van Belle, *Biostatistics: A Methodology for the Health Sciences* (Wiley-Interscience, New York, 1993), pp. 510–512.
24. W. R. Hruschka and K. H. Norris, *Appl. Spectrosc.* **36**, 261 (1982).
25. I. J. Wesley, S. Uthayakumaran, R. S. Anderssen, G. B. Cornish, F. Bekes, B. G. Osborne, and J. H. Skerritt, *J. Near Infrared Spectrosc.* **7**, 229 (1999).
26. C. E. Miller, "Chemical Principles of Near-Infrared Technology", in *Near-Infrared Technology in the Agricultural and Food Industries*, P. C. Williams and K. H. Norris, Eds. (Am. Assoc. Cereal Chem., St. Paul, MN, 2001), 2nd ed., pp. 19–37.
27. C. S. Blumenthal, I. L. Batey, F. Bekes, C. W. Wrigley, and E. W. R. Barlow, *J. Cereal Sci.* **11**, 185 (1990).
28. S. B. Altenbach, K. M. Kothari, and D. Lieu, *Cereal Chem.* **79**, 279 (2002).
29. B. G. Osborne and T. Fearn, *Near Infrared Spectroscopy in Food Analysis* (Longman Scientific and Technical, Harlow, U.K., 1986), pp. 117–161.