University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Publications from USDA-ARS / UNL Faculty

U.S. Department of Agriculture: Agricultural Research Service, Lincoln, Nebraska

2010

A Graphical Method to Evaluate Spectral Preprocessing in Multivariate Regression Calibrations: Example with Savitzky–Golay Filters and Partial Least Squares Regression

Stephen R. Delwiche
USDA-ARS, stephen.delwiche@ars.usda.gov

James B. Reeves III
USDA-ARS

Follow this and additional works at: https://digitalcommons.unl.edu/usdaarsfacpub

Delwiche, Stephen R. and Reeves, James B. III, "A Graphical Method to Evaluate Spectral Preprocessing in Multivariate Regression Calibrations: Example with Savitzky–Golay Filters and Partial Least Squares Regression" (2010). *Publications from USDA-ARS / UNL Faculty*. 1550.
https://digitalcommons.unl.edu/usdaarsfacpub/1550

This Article is brought to you for free and open access by the U.S. Department of Agriculture: Agricultural Research Service, Lincoln, Nebraska at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Publications from USDA-ARS / UNL Faculty by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

A Graphical Method to Evaluate Spectral Preprocessing in Multivariate Regression Calibrations: Example with Savitzky–Golay Filters and Partial Least Squares Regression

STEPHEN R. DELWICHE* and JAMES B. REEVES, III

USDA/ARS, Beltsville Agricultural Research Center, Food Quality Laboratory, Building 303, BARC-East, Beltsville, Maryland 20705-2350 (S.R.D.); and USDA/ARS Environmental Management and Byproduct Utilization Laboratory, Beltsville, Maryland, 20705-2350 (J.B.R.)

In multivariate regression analysis of spectroscopy data, spectral preprocessing is often performed to reduce unwanted background information (offsets, sloped baselines) or accentuate absorption features in intrinsically overlapping bands. These procedures, also known as pretreatments, are commonly smoothing operations or derivatives. While such operations are often useful in reducing the number of latent variables of the actual decomposition and lowering residual error, they also run the risk of misleading the practitioner into accepting calibration equations that are poorly adapted to samples outside of the calibration. The current study developed a graphical method to examine this effect on partial least squares (PLS) regression calibrations of near-infrared (NIR) reflection spectra of ground wheat meal with two analytes, protein content and sodium dodecyl sulfate sedimentation (SDS) volume (an indicator of the quantity of the gluten proteins that contribute to strong doughs). These two properties were chosen because of their differing abilities to be modeled by NIR spectroscopy: excellent for protein content, fair for SDS sedimentation volume. To further demonstrate the potential pitfalls of preprocessing, an artificial component, a randomly generated value, was included in PLS regression trials. Savitzky–Golay (digital filter) smoothing, first-derivative, and second-derivative preprocess functions (5 to 25 centrally symmetric convolution points, derived from quadratic polynomials) were applied to PLS calibrations of 1 to 15 factors. The results demonstrated the danger of an over reliance on preprocessing when (1) the number of samples used in a multivariate calibration is low ($<$50), (2) the spectral response of the analyte is weak, and (3) the goodness of the calibration is based on the coefficient of determination (R^2) rather than a term based on residual error. The graphical method has application to the evaluation of other preprocess functions and various types of spectroscopy data.

Index Headings: **Preprocessing; Savitzky–Golay; Near-infrared spectroscopy; NIR spectroscopy; Partial least squares; PLS; Derivative; Smoothing; Regression.**

INTRODUCTION

Near-infrared (NIR) spectroscopy is widely used for quantitative analysis in the chemical, food, and pharmaceutical industries because of its ability to generate rapid results and, more often than not, the accuracy of its predictions. The success of this technology has come about through the parallel development of the multivariate statistical regression procedures. Foremost among these procedures has been partial least squares (PLS) regression. Initially developed in the field of econometrics[1] and later expanded into chemometrics,[2] this procedure has become the technique of choice for NIR practitioners. Essentially, the PLS procedure, when applied to NIR spectra typically consisting of several hundred points per spectrum, reduces the number of needed points of the **x** block (the spectra) to a number that is representative of the rank of the data with respect to the **y** block (the regression variable). The PLS regression algorithm works to maximize the covariance between the **y** vector and any linear function of **X**, the matrix of the spectra, performing this successively with the residuals from the preceding component.[3]

Often, to enhance the wanted features of the spectra before the actual application of the PLS procedure, certain transformations are performed on the spectral data to reduce unwanted effects of light scatter caused by features of the physical structure (i.e., particle size) of the medium. The most common transformations are the first and second derivatives, which allow for the removal of vertical offsets and linearly sloping baselines. In certain instances, derivative preprocessing can produce calibrations with the lowest prediction error.[4] However, accompanying the derivative transformations is the potential of an increase in noise in the transformed spectra and the possibility of an apparent, but false, improvement in the correlation between spectral and chemical readings. These derivatives when applied to spectra consisting of several hundred discrete values are actually numerical approximations of true derivatives of continuous functions that match the appearance of the spectra.

The common numerical algorithm for the derivative is the Savitzky–Golay (S-G) approach,[5] which is based on a localized linear regression of several neighboring points to determine a best fit polynomial, whereupon this polynomial can be mathematically differentiated and evaluated at the x values coincident with wavelength collection points. In practice, a mathematical equivalent of the regression and differentiation procedure is performed by a convolution with a set of derived coefficients.[6] The NIR practitioner then finds the convolution window width that produces the best PLS calibration. Too large a window results in a distortion of the derivatized curve, while too small a window can introduce unwanted noise. Prior knowledge of the filter settings for S-G derivatives that produce optimal model performance is generally not known and therefore requires time-consuming search schemes.[7]

The goal of this study is to showcase a graphical procedure, based on contour plotting, that can be used in the evaluation of PLS preprocessing operations. An additional goal is to demonstrate through example the dangers of an over-reliance on the derivative, including the zeroth order (smooth), as a spectral pretreatment. This is done by using a set of low-noise NIR spectra of ground wheat, for which two chemical properties are known. The first property, protein content, is known to produce highly accurate NIR calibrations. The other property, sodium dodecyl sulfate (SDS) sedimentation volume, is of intermediate quality in modeling ability. Used as a protein quality indicator, SDS sedimentation is a measurement of the

Received 8 May 2009; accepted 3 November 2009.

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

0003-7028/10/6401-0073$2.00/0

© 2010 Society for Applied Spectroscopy

This document is a U.S. government work and is not subject to copyright in the United States.

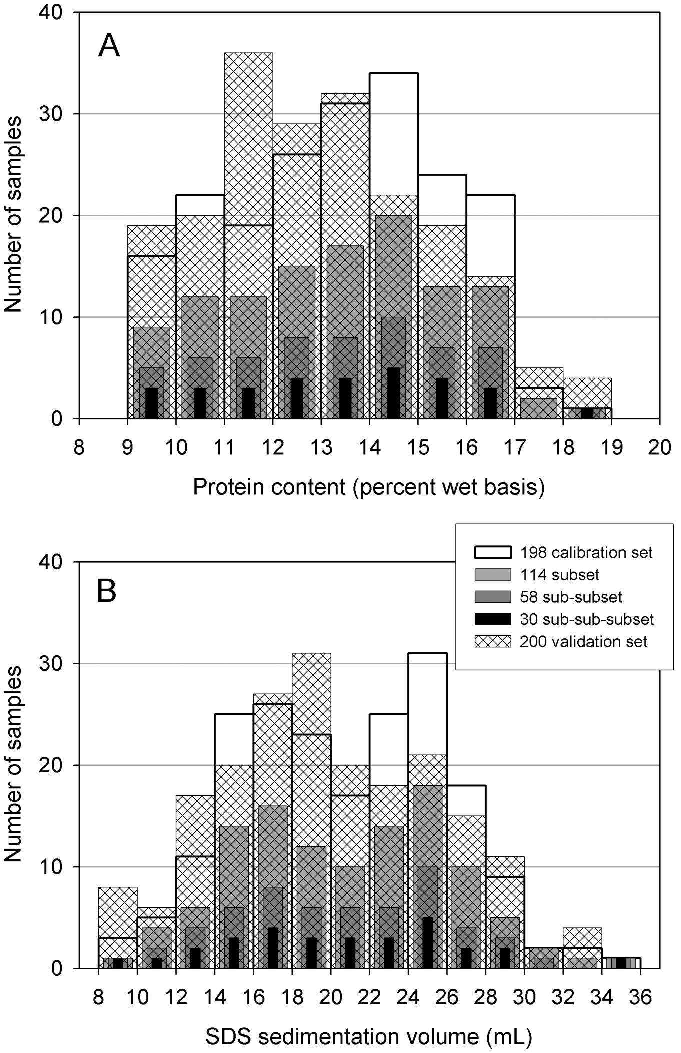

FIG. 1. Distributions of the analytes used in PLS calibration and validation trials: (**A**) protein content and (**B**) SDS sedimentation volume.

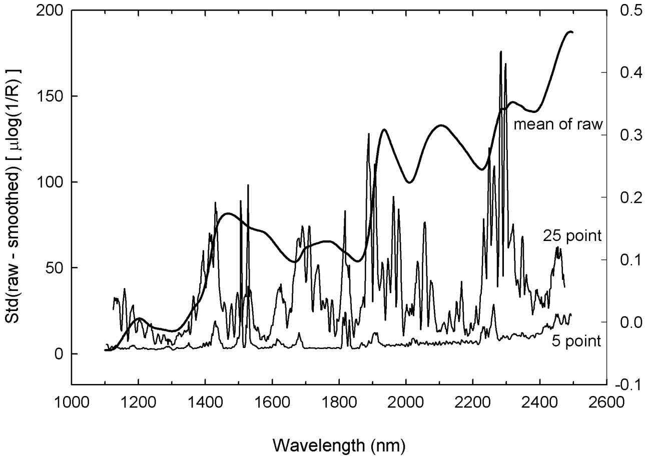

FIG. 2. Characterization of spectral noise, as represented by the standard deviation of the difference between raw and Savitzky–Golay smoothed spectra of the calibration samples ($n = 198$) at the two extremes in convolution window width (5 and 25 points). Also included is the mean raw spectrum (right axis scale).

settling volume of flour or meal hydrated in a solution (of water, SDS, and lactic acid) that produces a differential swelling and flocculation of glutenin and other insoluble constituents.[8] A third, fictitious property, derived by random number assignment, is introduced to demonstrate the pitfalls of a PLS analysis when sample size, method of validation, derivative window width, and number of PLS factors are not carefully considered.

EXPERIMENTAL

Samples. Ground wheat was used in this study, as this is a formulation that is ideally suited to diffuse reflection NIR spectroscopy and has been reported on for more than 30 years.[9] The wheat samples originated from a breeder's field performance trials of hard red winter and hard white wheat in Nebraska. Field replicated plots of approximately ten white and ten red cultivars/lines of wheat were planted at ten geographical locations within the state, which provided diversity in weather and soil conditions. Additional information on these samples can be found in a previous study.[10]

At each location, randomized plots were completely replicated twice, such that the number of samples available for spectral analysis was as follows: 2 field reps × 2 wheat classes × 10 lines × 10 locations ≈ 400. Approximately 15 g of each sample was ground in a cyclone mill (Udy, Fort Collins, CO) equipped with a 0.5 mm screen. The ground samples were held, one location at a time, at 33% relative humidity (at ≈22 °C) in a desiccator containing a saturated solution of $MgCl_2$ until the time of scanning. Grinding, conditioning, and scanning took place as each location's samples were delivered, which occurred over a two-month period.

Analytes. Although the original study contained a number of properties of interest to wheat breeders, such as yield and environmental stress indicators,[10] only two are used herein. The two properties, protein content and SDS sedimentation volume, were purposely chosen because the first is easily and very accurately measureable by NIR, while the second is also measurable by NIR but at a much lower level of performance. A third "analyte" was generated for this study. Random numbers were drawn from a normal distribution with a mean equal to 100 and standard deviation equal to 15.

Protein Content. Protein content ($N \times 5.7$) was measured by combustion (Model FP-428, Leco, St. Joseph, MI). Each sample was measured in duplicate 150 mg portions, then averaged. Precision of this procedure, determined as the standard deviation of 88 single analyses conducted on portions of the same sample over a one-month period, was 0.109% protein by weight.

Sodium Dodecyl Sulfate Sedimentation Volume. Measurement of SDS sedimentation volume was by AACC Approved Method 56-70,[11] with slight modifications. Briefly, ground wheat (≈2 g, <0.5 mm particle size) was added to a fixed volume of distilled water (≈25 mL) in a graduated cylinder and agitated for several minutes. Upon agitation, the volume was doubled by addition of a solution of 3.0% (w/w) sodium dodecyl sulfate and 2.0% (w/w) 1.2 *N* lactic acid stock solution, whereupon the mixture was agitated for several more minutes. After resting in the vertical position for 20 minutes, the sediment volume was measured. Precision of this procedure, as determined by the standard deviation of 30 replications on one control wheat over a one month period, was 2.2 mL.

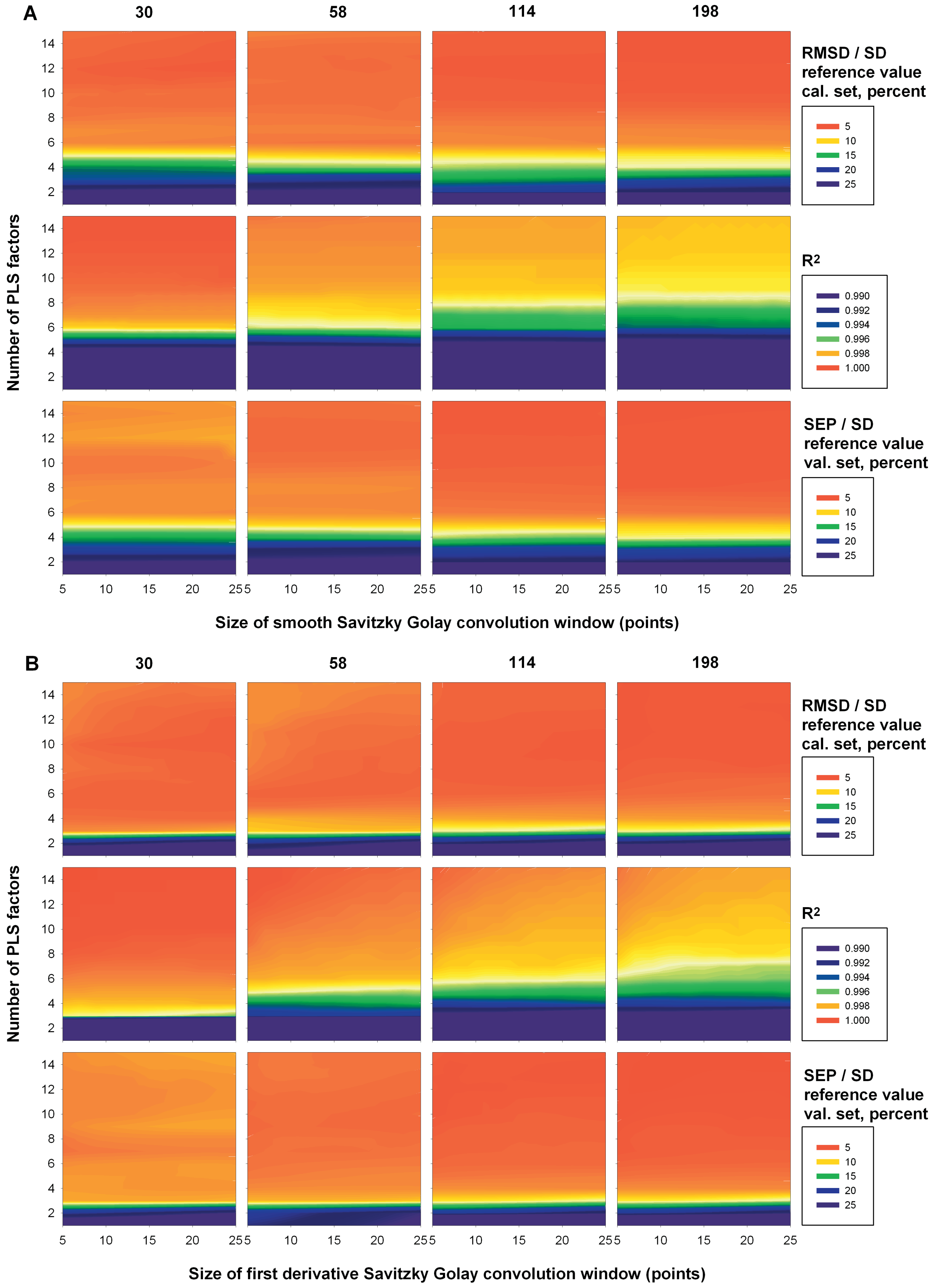

FIG. 3. Contour plots of PLS regression trials of protein content, using three Savitzky–Golay preprocess functions, (**A**) smooth, (**B**) first derivative, and (**C**) second derivative. Within each preprocess function, a column corresponds to the number of samples used in calibration (*first* = 30, *second* = 58, *third* = 114, *fourth* = 198) and a row corresponds to a normalized statistical figure of merit (*top* = cross-validation RMSD/SD$_{198 \text{ calibration samples}}$, *middle* = R^2, *bottom* = validation SEP/SD$_{200 \text{ validation samples}}$).

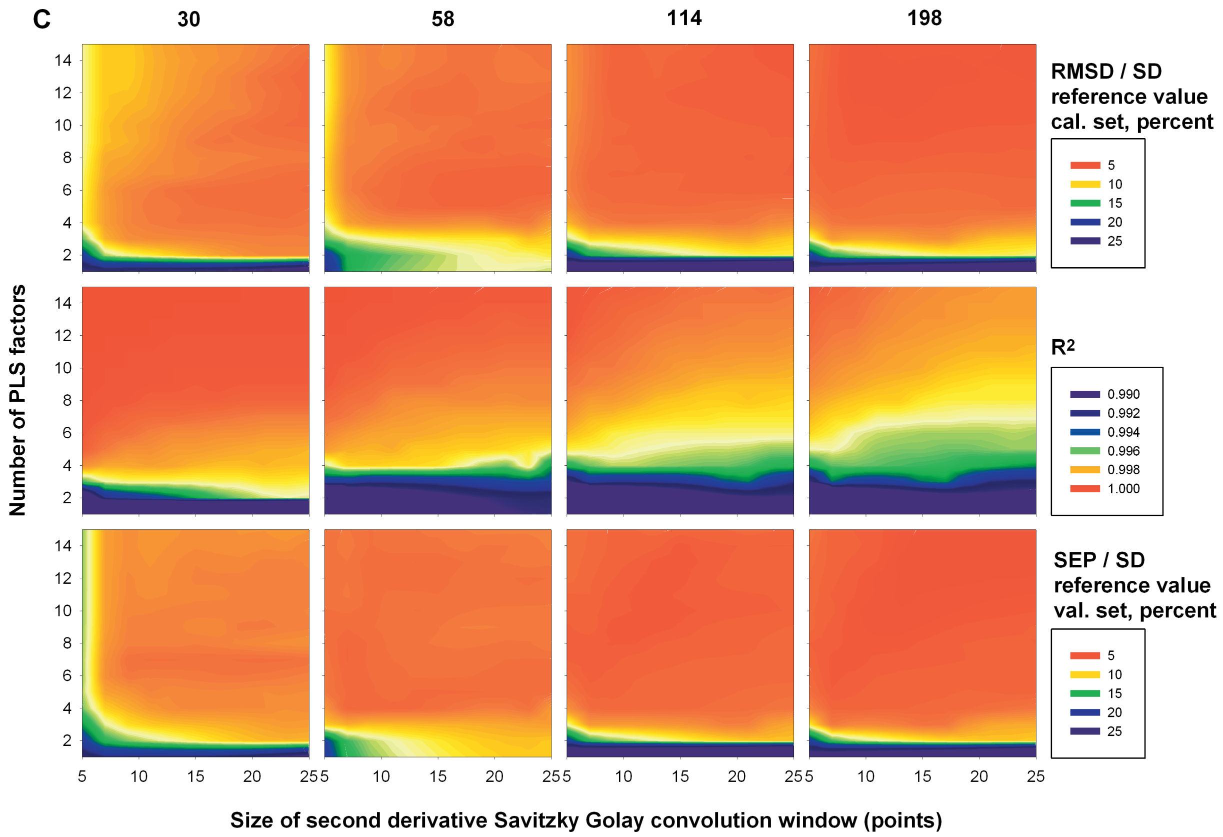

FIG. 3. Continued.

Near-Infrared Acquisition. Diffuse reflection (1100–2498 nm) readings of ground meal in a 30 mm diameter ring cell with quartz window were recorded at 2 nm increments (700 points total) using an analytical scanning monochromator (Foss-NIRSystems Model 6500, Laurel, MD) equipped with a spinning cup module. Reflectance readings (32 scans) were referenced to corresponding readings from a ceramic tile collected before each sample. Log($1/R$) values of duplicate packs were averaged and stored to disk for later analysis.

Partial Least Squares Analysis. For each analyte, the number of calibration samples consisted of four sizes, ranging from the maximum number available in one field replicate (198), to a first subset of 114 samples, and to a remaining subsubset and sub-sub-subset of 58 and 30 samples, respectively. The smallest set was formed by ordering all calibration samples by the analyte, selecting the first and last sample, and selecting a sample from each of 28 equally spaced groups. The intermediate sized sets were formed by selecting one or three additional samples from each group. The ordering procedure and subset formation was separately performed for each of the three analytes. The other field replicate ($n = 200$) was reserved for model validation. Histograms of the sets used in PLS calibrations for protein content and SDS sedimentation volume are shown in Fig. 1. Similarly, histograms of the validation set for these two analytes are contained in this figure.

Partial least squares (PLS-1) regressions on mean-centered data were performed separately on each of the three analytes. A one-sample-out cross-validation scheme was employed. Three common spectral preprocesses involving S-G convolutions (smooth, first derivative, and second derivative) were separately applied. In each case, centrally symmetric convolution windows varying from 5 points (10 nm) to 25 points (48 nm) in increments of 2 points (4 nm), based on quadratic polynomials, were applied and PLS calibrations were developed thereupon. The degree of noise in the spectra is represented by plots of the standard deviation of differences between the raw and smoothed spectrum at convolution windows of 5 and 25 points (Fig. 2). During PLS modeling, the number of factors ranged between 1 and 15, with the following statistical indices recorded at each number: root mean square of the differences (RMSD) of the cross-validation, multivariate coefficient of determination of the calibration equation (R^2), and standard deviation of the residuals (i.e., standard error of performance) of the validation set (SEP). Partial least squares routines were implemented in SAS (v. 9.1.3, SAS Institute, Inc., Cary, NC) using the PLS procedure (PROC PLS) in a macro program statement structure that allowed for looping across the number of factors and the convolution window width.[12] Later, the RMSD and SEP values were divided by the standard deviation (SD) of the analyte's reference values from the calibration and validation sets, respectively. Contour plots of the normalized statistical indices were formed from 165 entries (11 convolution window sizes × 15 PLS factors) per plot.

RESULTS

Protein Content. The modeling results of the various PLS trials with $n = 30$, 58, 114, and 198 calibration samples, with a S-G smoothing preprocess of varying window width (5 to 25 points) and PLS factors (1 to 15) are summarized in the contour plots of Fig. 3A. The graphs in Fig. 3A, as well as all remaining figures, are arranged in three rows and four columns. Each column corresponds to a calibration with a fixed number

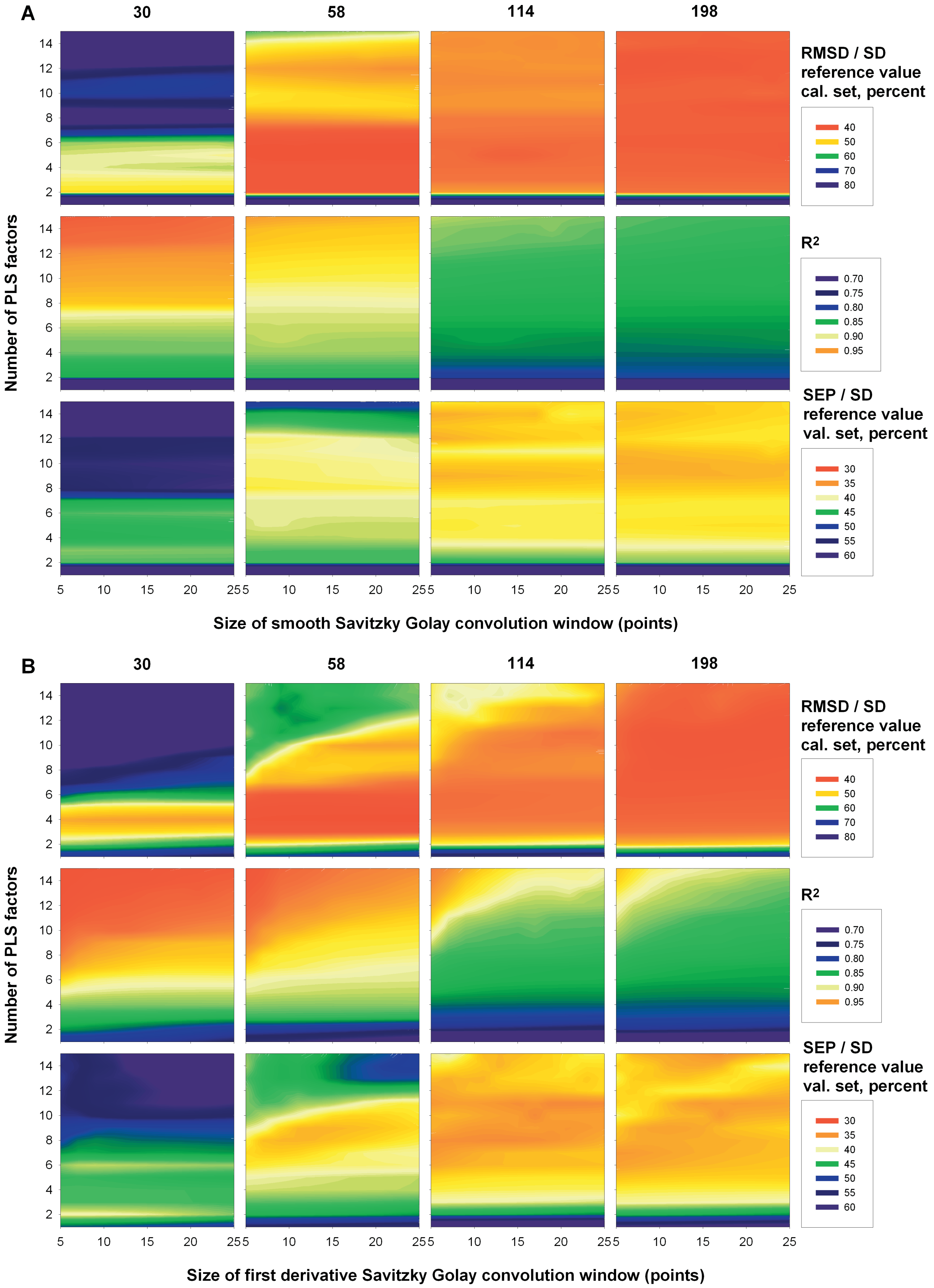

FIG. 4. Contour plots of PLS regression trials of SDS sedimentation volume, using three Savitzky–Golay preprocess functions, (**A**) smooth, (**B**) first derivative, and (**C**) second derivative. Within each preprocess function, a column corresponds to the number of samples used in calibration (*first* = 30, *second* = 58, *third* = 114, *fourth* = 198) and a row corresponds to a normalized statistical figure of merit (*top* = cross-validation RMSD/$SD_{198\ \text{calibration samples}}$, *middle* = R^2, *bottom* = validation SEP/$SD_{200\ \text{validation samples}}$).

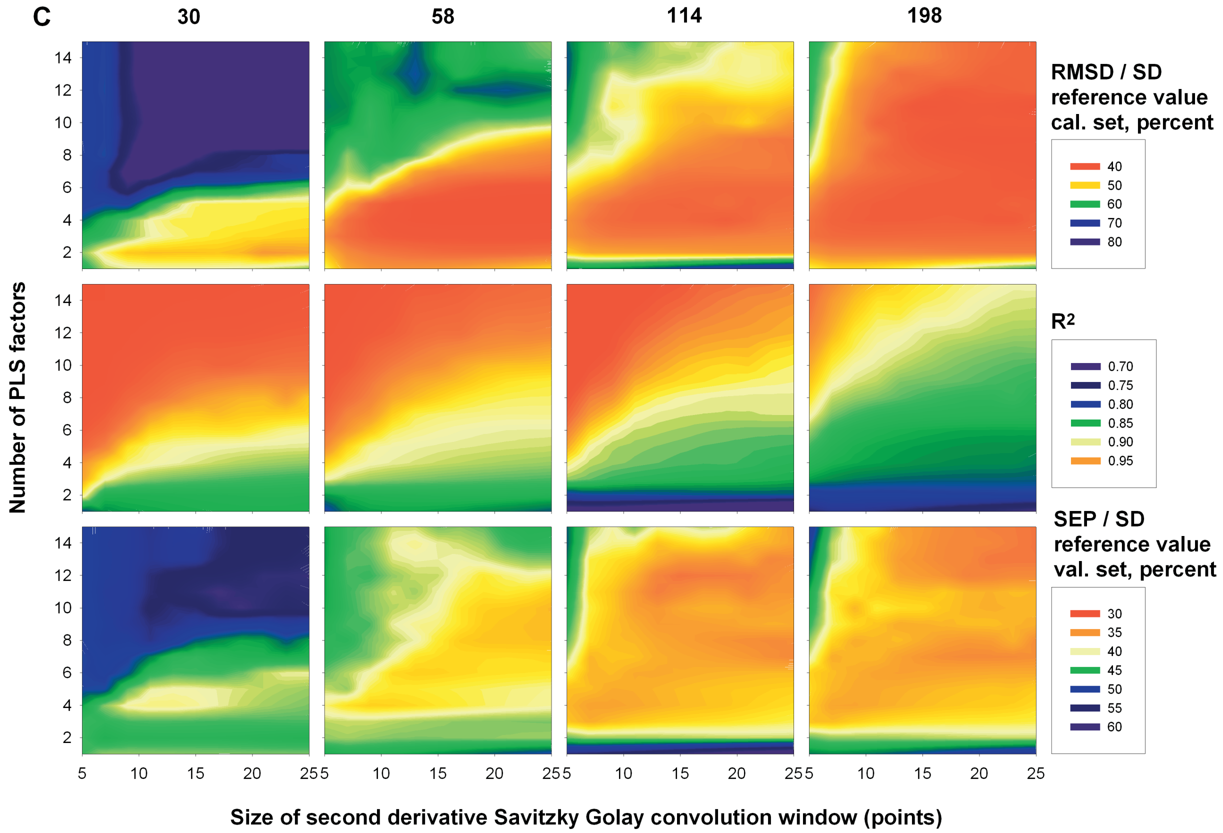

FIG. 4. Continued.

of samples: 30, 58, 114, and 198 for the first, second, third, and fourth columns, respectively. The first row represents one-sample-out cross-validation RMSD normalized by the SD ($n = 198$) of the analyte by the reference method (i.e., RMSD/SD). The second row represents the R^2 value of the calibration, and the third row of graphs is similar to the first row, but with the SEP of the validation set ($n = 200$) normalized by the SD of the reference values for this set (SEP/SD). Each row contains a legend for the contour values that applies to all graphs within the row.

Apparent from normalized RMSD values (first row of graphs) of Fig. 3A is a lack of effect of S-G window width, as shown by the horizontal striped appearance of these graphs. Over the convolution width range of 5 points (8 nm) to 25 points (48 nm), RMSD is basically unchanged. The much larger effect on RMSD occurs as the number of PLS factors is varied; in the case of protein content, the RMSD improves as the number of samples is increased and the number of factors is increased. At the two larger sample number sets (114 and 198), model performance is nearly identical, such that for 10 to 15 PLS factors the RMSD approaches 5% of the natural variation of the analyte. Even the 30-sample RMSD graph indicates model performance consistent with the models developed using the larger numbers of samples. This observation is not true when comparing the R^2 graphs (second row). Based on R^2 alone, the 30-sample calibration appears superior to the higher sample calibration equations, both in terms of higher values and a smaller number of factors needed to achieve these elevated values, with $R^2 > 0.98$ for calibrations of seven factors and higher. However, as the number of samples increases, especially when going from 58 to 114, the R^2 values decline slightly, though they are still greater than 0.95. The appearances of the 114- and 198-sample R^2 graphs are nearly identical, which indicates that the potential change in protein content calibration performance with sample augmentation is minimal. The result that 30- and 58-sample calibrations lead to inflated correlation statistics supports the recommendation of ASTM that the number of spectra in a calibration should be at least six times larger than the number of PLS latent variables.[13]

Although external set validation is considered to be the best indicator of PLS calibration performance, this form of validation is often absent in studies, especially when the number of samples is limited. The importance of external validation is demonstrated in the third row graphs of Fig. 3A. Whereas according to the RMSD and R^2 indicators, in which the small sample number set ($n = 30$) is favored, the calibration resulting therein is not as robust as the calibrations based on more samples. Again, in comparing the SEP graphs of the rightmost two graphs of this row, little difference in performance is observed between the $n = 114$ and 198 calibrations. Similar to the contour plots of RMSD and R^2, the effect of window width of the smooth convolution on SEP is minimal for protein content.

When the first-derivative convolution function is applied to the PLS protein calibrations (Fig. 3B), the effects on RMSD, R^2, and SEP are somewhat similar to the corresponding indicators of the smooth convolution. The most noticeable change is an expansion of the region of high model performance. Defining such performance to be RMSD/SD < 10%, $R^2 > 0.98$, and SEP/SD < 10%, the change from smooth

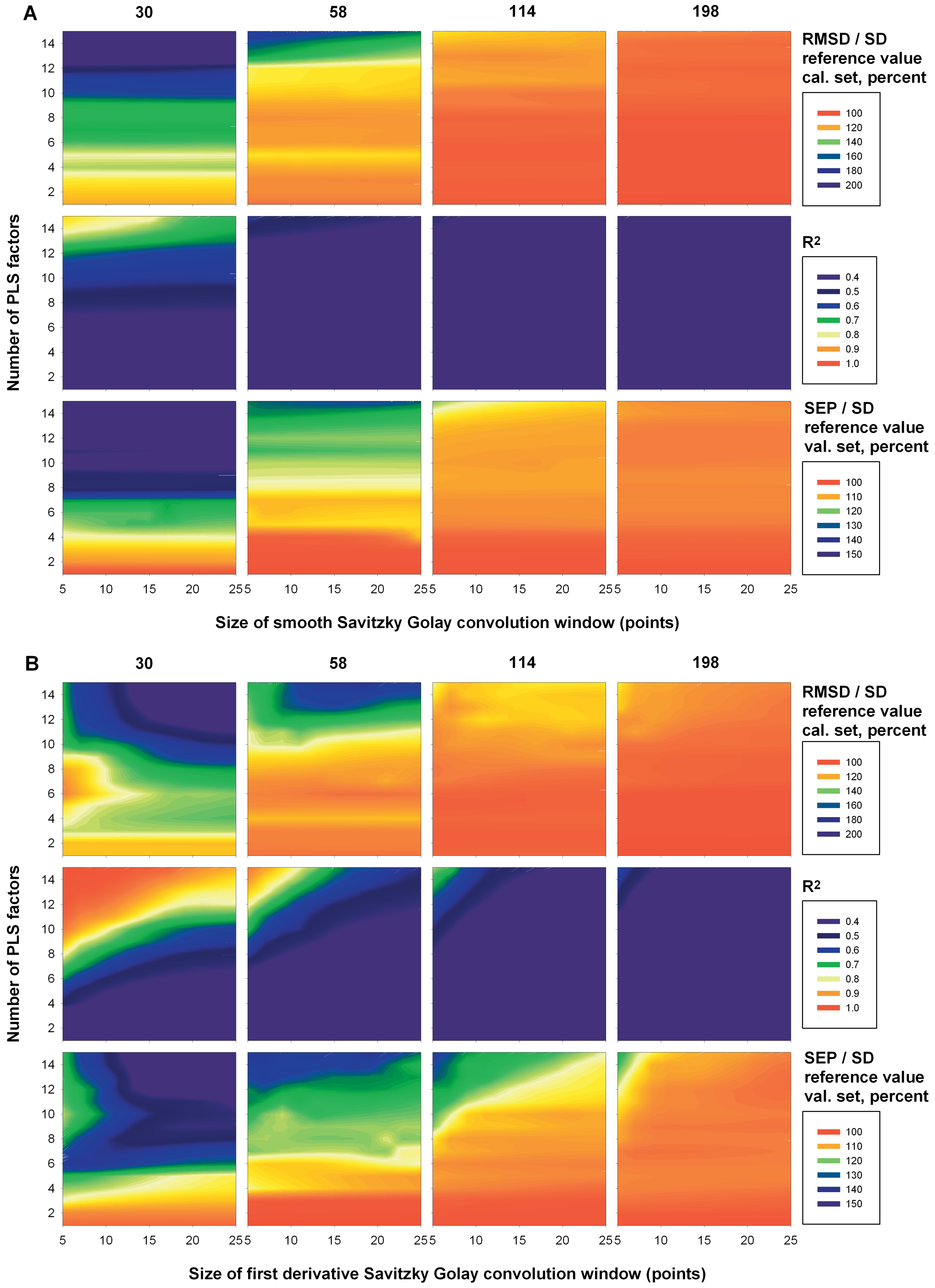

FIG. 5. Contour plots of PLS regression trials of a random number, using three Savitzky–Golay preprocess functions, (**A**) smooth, (**B**) first derivative, and (**C**) second derivative. Within each preprocess function, a column corresponds to the number of samples used in calibrations (*first* = 30, *second* = 58, *third* = 114, *fourth* = 198) and a row corresponds to a normalized statistical figure of merit (*top* = cross-validation RMSD/SD$_{198\ \text{calibration samples}}$, *middle* = R^2, *bottom* = validation SEP/SD$_{200\ \text{validation samples}}$).

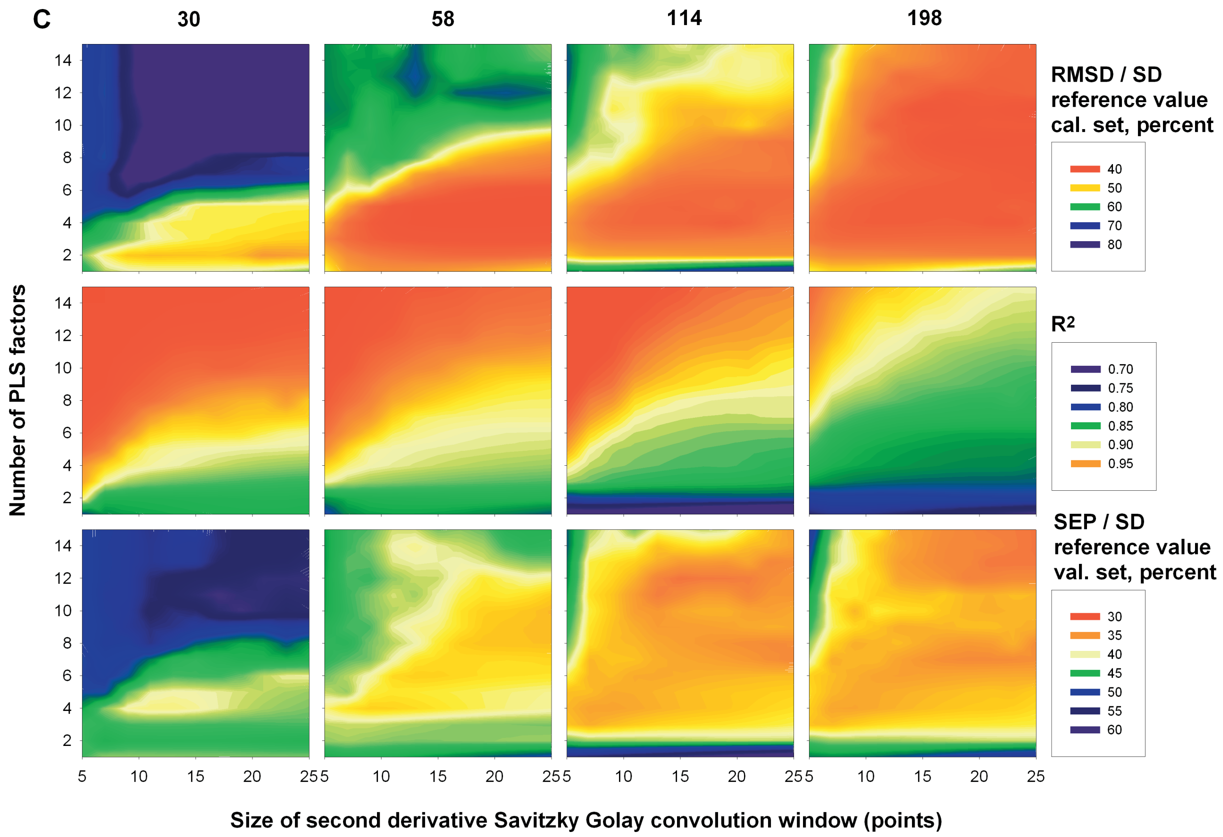

FIG. 5. Continued.

to first derivative typically allows similar performance to occur with one to two fewer PLS factors. For example, in a calibration involving all 198 samples, the high performance region of the normalized SEP begins at 4 and 3 factors for the smooth and first-derivative convolutions, respectively. Also, similar to the smooth convolution, the effect of window width of the first-derivative convolution is insignificant. Recent simulation modeling of light through turbid liquid with four absorbing components produced a similar finding that derivative preprocessing leads to PLS models that require fewer latent variables, but without an improvement in performance.[14]

Second derivatives are often the transfer functions of preference because of their dual ability to reduce the effect of baseline shifts and sloping baselines and to accentuate the appearance of previously overlapping absorbers, with peaks coincident with the absorption frequencies. For protein content (Fig. 3C), the main result of the second derivative has been to expand the region of high performance (defined above). However, in this case, the slight curved nature of the contours (most evident by following the regions of yellow color in the color versions of the graphs) suggests that convolution window width now has a slight effect on model performance. In particular, wider windows are favored when the number of factors is small (<4); however, this effect is diminished as the number of calibration samples is increased. Another feature that is common to all three convolution functions is the misleading nature of R^2. Whereas the 30-sample calibrations consistently appear superior to the higher sample calibrations, regardless of convolution function, the opposite effect is borne out in the SEP.

Sodium Dodecyl Sulfate Sedimentation Volume. Partially due to an inherent positive correlation to the overall concentration of protein (i.e., protein content), SDS sedimentation volume, which is an indicator of gluten protein quality, is an analyte that achieves moderate success in NIR equation development.[10] This is demonstrated in the contour plots of Figs. 4A–4C, which are arranged in the same manner as the plots for protein content in Figs. 3A–3C. However, the scales for the contour regions of RMSD/SD, R^2, and SEP/SD are different with respect to those for protein content to reflect the overall reduction in model performance for this analyte. In examining the effect of the smooth convolution function, it is apparent that the best RMSD/SD is approximately 40%, as opposed to between 5% and 10% for protein content. Similarly, the best normalized SEP values rise by the same amount. Whereas protein content produces the best NIR models of any naturally occurring analyte, SDS sedimentation volume is probably more typical for NIR calibration development trials.

Poorer model performance is attributed to two main reasons. First, the property, instead of being a concentration, is an indicator of the quality of the wheat gluten protein that comes from a measurement procedure that is less precise than either Kjeldahl or Dumas (combustion). Secondly, the homologous nature of the endosperm proteins makes it very difficult to spectrally discern quality-enhancing glutenin from quality-neutral proteins. When RMSD/SD graphs are examined, it is found that unlike protein content the SDS sedimentation volume calibrations are noticeably influenced by the number of samples. Calibrations involving a small number of samples (30 and 58) favor a small number of PLS factors (2 to 6). As the number of samples increases, however, the behavior begins to resemble that of protein content such that the recommended number of PLS factors expands to form a wide range, between 2 and 15. This phenomenon is mirrored in the validation graphs, where it is noted that the validation set performance is improved when the regression equation is based on the second to largest (114) or largest number of samples (198). In fact, the 30-sample and 58-sample calibrations produce misleading R^2 results because they seemingly favor a large number of factors, which is contrary to the behavior of either the calibration set RMSD or the validation set SEP. Consistent with protein content modeling there is no effective influence of the width of the smooth convolution window.

Application of a first-derivative preprocessing to the SDS sedimentation volume PLS calibrations results in performance trends that differ from smooth preprocessing by showing a small effect of convolution window width, as seen by the departure from a strictly horizontal nature of the contours in all of the R^2 graphs and in the intermediate (58- and 114-sample) graphs of RMSD/SD and SEP/SD (Fig. 4B). Again, the R^2 graphs of the first-derivative preprocessing are misleading. In fact, even the 114-sample and 198-sample calibrations seemingly favor a large number of factors (12–15) and small convolution window (<15 points), which is not substantiated by either the RMSD/SD or SEP/SD graph series. When these latter two series are compared with the corresponding graphs of the smooth preprocessing, it is apparent that the first-derivative preprocess has not produced better performance. This leads to the question of whether advantages exist for higher order derivative preprocessing of SDS sedimentation calibrations. The graphs in Fig. 4C do not reveal an advantage, as this higher order convolution has introduced a greater dependency on window size, but with no improvement in best model performance.

Random Numbers. The results of the PLS calibrations of a randomly assigned number are shown in the graphs in Figs. 5A–5C. To be expected, the smallest cross-validation RMSD or validation SEP approaches but never becomes less than the standard deviation of this analyte due to its artificial fabrication. Therefore, the scales of the RMSD/SD and SEP/SD have 100% as the minimum value in the range. Although the R^2 graphs do not show promise for the smooth preprocessing step, as they shouldn't, the pitfall of too small a sample set is shown in the $n = 30$ calibration, in which values approaching 0.9 occur when the number of factors is high and the convolution window width is low. A better analysis of the response is revealed through the RMSD/SD and SEP/SD graphs. Contrary to well-behaved calibrations such as protein content, the fact that the RMSD and SEP values are lowest at the smallest number of factors is an indication of the weakness of the calibration.

As the derivative order increases from zeroth (smooth) to first and second, the cross-validation RMSD and validation SEP indicators remain essentially unchanged with the exception that the contour patterns now indicate a departure from horizontal striping, which seemingly indicates an effect of convolution window width. The most misleading graphs are the ones of R^2. Although this has been the case for the two real analytes, it is striking to see that R^2 values can exceed 0.90 when the number of PLS factors is sufficiently high (>10), even for calibrations involving more than 100 samples.

DISCUSSION

Early research on the applications of derivatives in preprocessing was done for the purpose of enhancing individual bands that were otherwise overlapped with other bands.[15–17] Later research has routinely targeted derivatives as a preprocessing step that removes or reduces baseline offsets (first derivative) or baselines that are sloped with respect to the wavelength axis (second derivative). Often unknown in the application of derivatives is their effect on the quality of the spectra in terms of signal-to-noise ratio and on the performance of the regression models that follow. Because of this, NIR practitioners run the risk of overly relying on chemometric software to ascertain the best preprocess conditions for a particular analyte. Such practices, if not properly interpreted, can lead to exaggerated model performance, as demonstrated by the relatively large values for R^2 on PLS regression calibrations of an analyte (random number) that has no inherent correlation to its corresponding spectra (Figs. 5A–5C). By example, the general invariance of protein content or SDS sedimentation volume model performance as the convolution window width varied in the smooth preprocessing operation corroborates the findings of Brown and Wentzell,[18] who determined by simulation and analysis of simple mixtures (using conditions of spectral noise being independent and identically distributed across wavelengths and calibration errors being negligible) that smoothing results in spectral distortion. The spectral noise of adjoining wavelengths becomes correlated and the ensuing multivariate calibration error is not reduced.

Brown and co-workers also theoretically examined derivative preprocessing.[7] In that work, the researchers addressed a commonly held idea that noise is disproportionately increased with respect to the signal when derivatives are applied. In fact, when derivatives are accomplished through application of polynomial least squares convolution filters, such as S-G filters, signal-to-noise ratio can actually improve. This is especially true in circumstances when the "noise" or drift of neighboring wavelength points is correlated, which is the usual case for NIR spectroscopy data. The order of the derivative affects the degree of the attenuation of the broad-based trends such as drift, with higher order derivatives providing greater attenuation. The width on the polynomial filter has an effect on the "high frequency" components, or sharpness, of the spectra such that larger windows produce a greater suppression of these features. With respect to the spectra of this study, the fact that the size of the convolution window had little effect on the cross-validation RMSD or validation SEP for protein content suggests that the NIR PLS models of well-modeled naturally occurring food analytes do not owe their efficacy to high frequency absorption bands.

The slight effect of the convolution window size for SDS sedimentation volume, especially for the first and second derivatives, alludes to the more difficult nature of modeling this analyte. In fact, the improvement in performance with derivative order, which was not seen for protein content, points to the more subtle nature of the SDS sedimentation spectral response. However, the more apparent effect on SDS sedimentation modeling is sample size. When the analyte is easily modeled, the number of samples used in calibration does not matter nearly as much as when the analyte is more difficult to model, as seen by the similarity of the protein contour plots within a statistical figure of merit (RMSD, R^2, or SEP) and within a preprocessing treatment (Figs. 3A–3C). This is contrasted with the SDS sedimentation plots, especially those for SEP/SD, which indicate that the 30- and 58-sample calibrations were noticeably inferior to the 114- and 198-sample calibrations.

The clearest indication of the danger of PLS regression calibration development with a small sample set are the second-derivative contour plots for the random number models (Fig. 5C). Without proper attention paid to the residual errors during either cross-validation or external validation, a reliance on the multivariate coefficient of determination can lead the practitioner into accepting calibration equations of little or no predictive power. The literature suggests that the performance of multivariate calibrations are enhanced by increased numbers of calibration samples for two reasons: better reliability of models through a more accurate representation of the correlations between the analyte concentrations and the spectral responses, and better assurance that the larger number of samples will mathematically span the multitude of components within a complex mixture, such as a natural biological compound.[19] It is interesting to note that from the analysis of PLS and PCR models of simulated mixtures,[19] model quality categorization was described that is similar to the current study, these being that multivariate calibrations could be divided into three "regions" (as defined by the analyte being modeled): a region of no predicative ability, a region of good predictive ability, and a transition, or intermediate region. Respectively, these regions correspond to the random number, protein content, and SDS sedimentation volume in the current study. Although several studies have been performed with synthetic datasets on the number of calibration samples, the preprocessing method, or the multivariate regression method itself from which recommendations on modeling have been made, the use of experimental data consisting of many complex compounds contributes to the complexity of the regression analysis and its predictability.[20]

CONCLUSION

The intention of this study was to provide the NIR chemometric analyst with a visual representation of the effect of S-G convolution filters (zeroth, first, and second differentiation order) that are commonly used in spectral preprocessing. While such filters are used to remove unwanted effects of noise (zeroth order), or offset (first order) and sloped baselines (second order) brought on by a scattering medium, careful interpretation after their use is needed in order to guard against the development of misleading PLS calibrations. By applying the visual technique to a spectral set of ground wheat meal, we have found that smoothing alone does not appear to improve calibration performance. Secondly, the higher the order of differentiation, the greater is the risk that PLS calibrations will produce exaggerated performance, particularly when (1) the analyte itself is inherently weak in its ability to be modeled, (2) small sample sets (i.e., $n < 50$) are used in calibration, and (3) the multivariate coefficient of determination (R^2) is used instead of residual error-based terms.

ACKNOWLEDGMENTS

We wish to thank Robert Graybosch (USDA-ARS, Lincoln) for supplying the wheat samples and performing the SDS sedimentation volume analyses. Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply endorsement or recommendation by the USDA.

1. H. Wold, "Estimation of Principal Components and Related Models by Iterative Least Squares", in *Multivariate Analysis,* P. R. Krishnaiaah, Ed. (Academic Press, New York, 1966), pp. 391–420.
2. S. Wold, H. Martens, and H. Wold, "The multivariate calibration problem in chemistry solved by the PLS method", in *Proceedings of the Conference on Matrix Pencils,* A. Ruhe and B. Kagström, Eds. (Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 1983), p. 286.
3. T. Næs, T. Isaksson, T. Fearn, and T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification* (NIR Publications, Chichester, 2002), p. 27.
4. L. Stordrange, F. O. Libnau, D. M. Sorenssen, and O. M. Kvalheim, J. Chemom. **16,** 529 (2002).
5. A. Savitzky and M. J. E. Golay, Anal. Chem. **36,** 1627 (1964).
6. J. Steinier, Y. Termonia, and J. Deltour, Anal. Chem. **44,** 1906 (1972).
7. C. D. Brown, L. Vega-Montoto, and P. D. Wentzell, Appl. Spectrosc. **54,** 1055 (2000).
8. C. F. Morris, B. Paszczynska, A. D. Bettge, and G. E. King, J. Sci. Food Agric. **87,** 607 (2007).
9. P. C. Williams, Cereal Chem. **52,** 561 (1975).
10. S. R. Delwiche and R. A. Graybosch, Appl. Spectrosc. **57,** 1517 (2003).
11. AACC, Approved Methods of the AACC (American Association of Cereal Chemists, St. Paul, MN, 2000), 10th ed., Method 56–70.
12. J. B. Reeves, III and S. R. Delwiche, J. Near Infrared Spectrosc. **11,** 415 (2003).
13. ASTM, E1655-08, Standard Practices for Infrared Multivariate Quantitative Analysis (ASTM International, West Conshohocken, PA, 2008).
14. S. N. Thennadil and E. B. Martin, J. Chemom. **19,** 77 (2005).
15. V. J. Hammond and W. C. Price, J. Opt. Soc. Am. **43,** 924 (1953).
16. J. D. Morrison, J. Chem. Phys. **21,** 1767 (1953).
17. F. R. Stauffer and H. Sakai, Appl. Opt. **7,** 61 (1968).
18. C. D. Brown and P. D. Wentzell, J. Chemom. **13,** 133 (1999).
19. P. D. Wentzell and L. V. Montoto, Chemom. Intell. Lab. Syst. **65,** 257 (2003).
20. F. Navarro-Villoslada, L. V. Pérez-Arribas, M. E. León-González, and L. M. Polo-Díez, Anal. Chim. Acta **313,** 93 (1995).