NIR news
2017, Vol. 28(4) 16–22

DOI: 10.1177/0960336017706881
journals.sagepub.com/home/nir

Summary of the 2016 IDRC software shoot-out

Benoît Igne[1], Md Anik Alam[2], Dongsheng Bu[3], Pierre Dardenne[4], Hanzhou Feng[2], Ali Gakhani[5], David W Hopkins[6], Shikhar Mohan[2], Charles R Hurburgh[7] and Cathleen Brenner[8]

The Software Shoot-Out has been a staple of the International Diffuse Reflectance Conference (IDRC), a biennial meeting taking place in Chambersburg, Pennsylvania, USA. It is a competition among participants of the conference that acknowledges and rewards the person who develops the best model(s) and obtains the lowest prediction error for a particular diffuse reflectance spectral dataset. For every IDRC, a new challenge is presented. The conference website (http://www.cnirs.org) provides access to the past conference datasets as well as those used for previous challenges. Previous *NIR news* articles have reported results from the past three competitions.[1–3]

Two competitions took place during the 2016 conference: as usual, a dataset was made available for download and completion at home. In addition, an on-site competition was proposed to all conferees. This on-site shoot-out was carried out as an anonymous challenge in which students and professionals used their chemometric skills to come up with the best prediction models for two parameters pertaining to a single dataset. The top three students and the top three professionals were recognized during the conference banquet.

The more traditional shoot-out presentation followed the on-site challenge and was a great occasion at which to learn from and interact with experienced chemometricians presenting their approach to a common multivariate analysis problem. For the first time, a calibration transfer challenge was proposed. The conference thanks Cathleen Brenner of the U.S. Department of Agriculture Grain Inspection, Packers and Stockyards Administration (GIPSA) and Charles R. Hurburgh Jr of Iowa State University for providing the data and facilitating the competition. The conference also thanks Bruins Instruments (http://www.bruins.de), FOSS (http://www.foss.dk), and Perten Instruments (http://www.perten.com) for permitting the use of the data.

Challenge

A common set of whole grain wheat data from three NIR transmission instrument manufacturers was provided. The challenge consisted of matching the spectra from all three instrument manufacturers, then developing a regression model for protein that resulted in equivalent results among the models as measured by the reproducibility. The samples represented wheat grown throughout the United States with spectra collected on five instruments per NIR spectrometer manufacturer. The instruments' make and serial numbers were coded. The spectra were in the spectral range and spacing that the respective instrument manufacturers supported. The reference protein results, provided by GIPSA, were on a 12% moisture basis.

An initial study undertaken by Iowa State University on behalf of the U.S. Department of Agriculture's GIPSA yielded a reproducibility (standard deviation (SD) across NIR models) of 0.14% protein compared to an average SD of 0.07% protein across instrument copies of a given manufacturer. Reproducibility, as defined by the following equation

$$Reproducibility = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)^2}{n-1}}$$

where n is the number of samples was used to judge the effectiveness of the calibration transfer method(s). Note that this definition can apply within copies of a given instrument per manufacturer or across multiple instruments from different manufacturers.

Dataset

- There were 1488 spectra in the calibration dataset for 248 samples analyzed on three instruments for manufacturer A and three instruments for manufacturer B.

[1]GlaxoSmithKline, King of Prussia, USA
[2]School of Pharmaceutical Sciences, Duquesne University, Pittsburgh, USA
[3]Bristol-Myers Squibb, New Brunswick, USA
[4]Walloon Agricultural Research Centre, Gembloux, Belgium
[5]Aunir, Towcester, UK
[6]NIR Consultant, Battle Creek, USA
[7]Agricultural and Biosystems Engineering Department, Iowa State University, Ames, USA
[8]U.S. Department of Agriculture's Grain Inspection, Packers and Stockyards Administration, Kansas City, USA

Corresponding author:
Benoît Igne, GlaxoSmithKline, King of Prussia, PA 19406, USA.
Email: igneben@gmail.com

- There were 744 spectra in the test dataset for the same 248 samples as in calibration analyzed on a fourth instrument for manufacturer A and B in addition to one instrument for manufacturer C.
- There were 450 spectra in the validation dataset for an independent set of 150 samples analyzed once each on a new instrument for each manufacturer A, B, and C.
- Reference protein values were provided for the calibration and test sets only.

The order of the calibration and test samples for all sets and instrument manufacturers was identical as run in the Iowa State University laboratory. However, the order of the validation samples was randomized for each instrument manufacturer.

Figure 1 presents the structure of the dataset and the challenge. Figure 2 presents the mean spectra for the test set from each of the three instruments.

The next section describes approaches as provided by six of the participants.

Participant approaches

Participant 1

The spectral responses from the different instruments were matched in terms of wavelength range and resolution. The common wavelength channels were selected from each instrument to avoid any extrapolation or interpolation of the signals. Spectral baseline differences were observed between instruments after matching the wavelength range and resolution. Savitzky–Golay (SG) second derivative preprocessing was performed using a window size of 9 and a second-order polynomial fit to decrease the baseline difference. A partial least squares (PLS) model was developed using the calibration set and the model

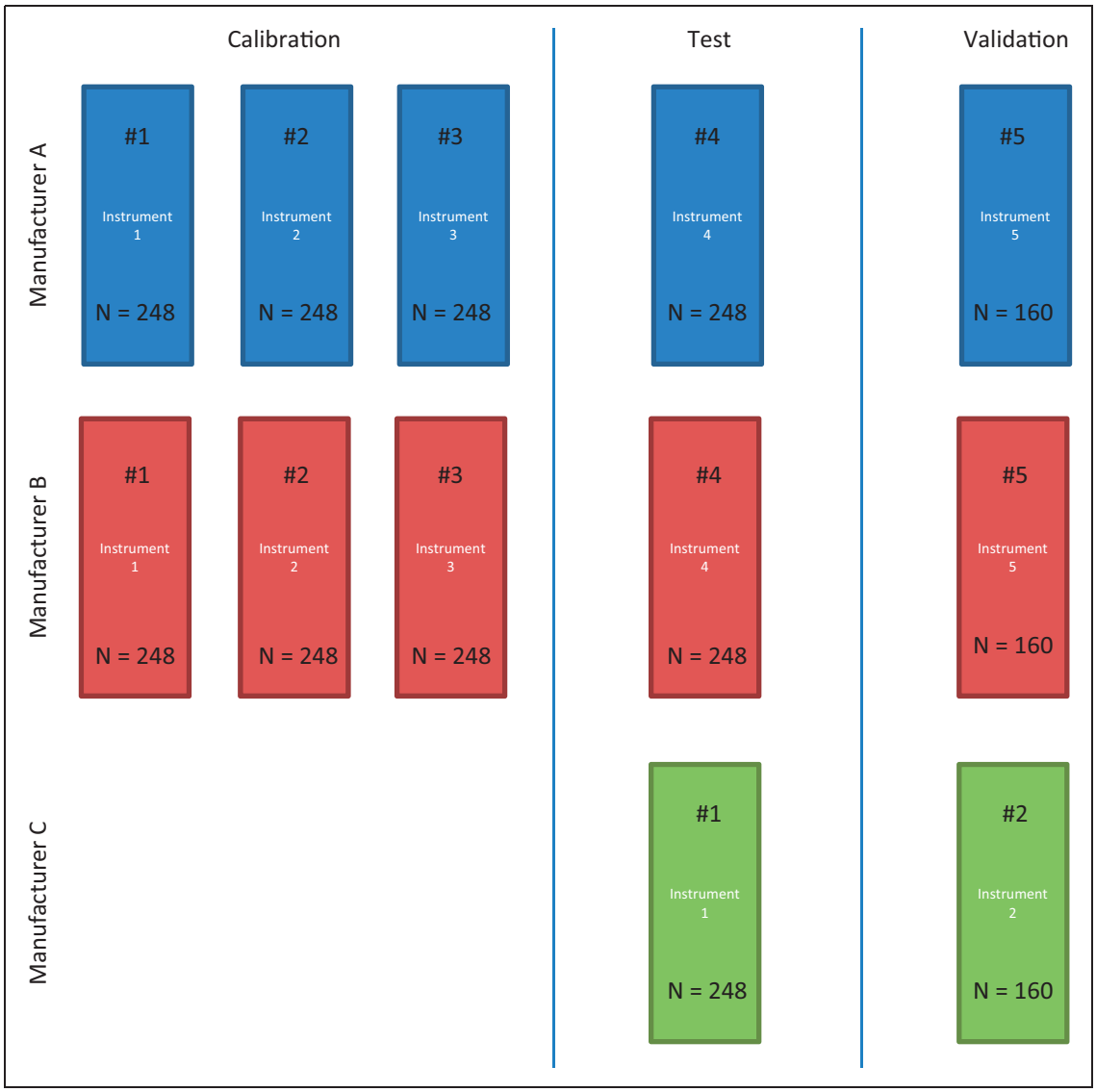

Figure 1. Structure of the dataset in calibration, test, and validation.

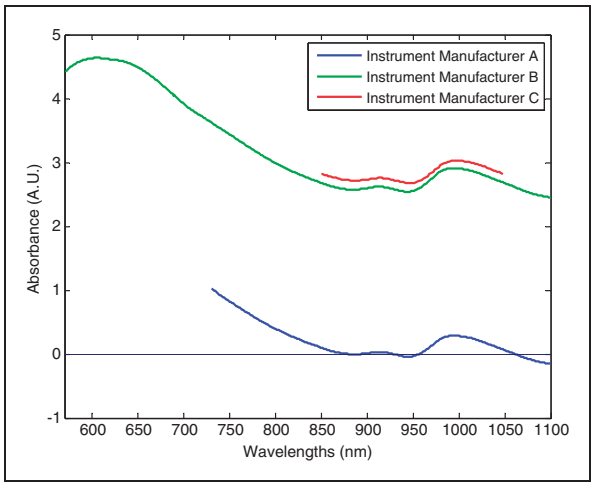

Figure 2. Raw test spectra.

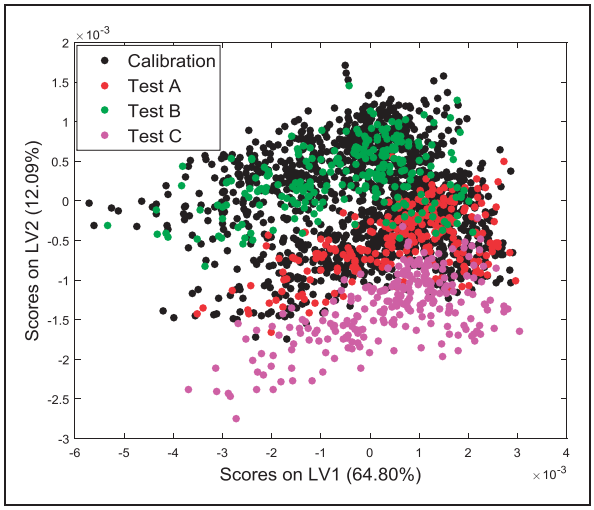

Figure 3. Score plot of the calibration and test set.

performance was assessed on the test set. The RMSEC and RMSEP were 0.36 and 0.37%, respectively. Figure 3 shows the projection of the test spectra of the different instrument vendors in the PLS score space. Grouping based on the instrument sources indicated a significant effect of remaining spectral differences, even after preprocessing.

The spectral difference was further minimized by direct standardization as per the equation below[4]

$$Spectral\ signals_{secondary} * Transfer\ Matrix = Spectral\ signals_{Primary}$$

The primary instrument was selected based on individual instrument's calibration performance. The instrument "A3" (third dataset from instrument manufacturer A) was selected as the primary unit based on its minimum RMSEC (%) among all. Spectra from all other instruments (from same and different manufacturers) were considered as secondary units. Direct standardization was performed to match the net analyte signals between primary and secondary instruments.[5] The idea was to minimize the effect of interference during direct standardization. The net analyte signal was calculated by iteratively removing the spectra that were related to prediction residuals. At each iteration, the information related to wheat concentration increased as the interference information decreased.

The transfer matrix was solved and used to transform the preprocessed spectra from all the instruments to the net analyte signals of the primary instrument (A3). The RMSEC and RMSEP were 0.31 and 0.33% after transformation. Figure 4 shows the projection of the test and validation spectra in the new calibration space. Instrument-based grouping was absent in the score space indicating the minimization of spectral difference upon standardization to the net analyte signals. The increased percent of explained variance by the first loading vector (from 64.8 to 97.44%) indicated that the dataset mostly contained information related to the wheat protein concentration.

Participant 2

It was necessary to lineup and truncate spectra to make all nine instruments have the same wavelength range and number of wavelength variables. Spectra from instrument manufacturers A and B were first pretreated by SG smoothing with an 11-point window and a second-order polynomial filter. Then every fourth wavelength points in the region of 848–1046 nm for manufacturer A, and in the region of 847–1045 nm for manufacturer B were used. Manufacturer C spectra were used as provided.

After spectral alignment and truncation, all spectra had 100 wavelength variables. Further pretreatments were needed to minimize spectral variations from sample physical property and instrument changes. Four steps: standard normal variate (SNV)–SG second derivative (11-point window and second-order polynomial)–SNV–mean centering were applied to all spectral data. SNV was used twice which is unusual for a typical spectral pretreatment scheme.

A PLS regression was used for model development with all samples in the six provided calibration sets and tested with all samples in the provided three test sets. No sample was removed in the final model. It was possible few protein reference values were exchanged on purpose or by mistake. Four pairs of protein reference values were switch for modeling, i.e. pairs of (779,780), (1275,1276), (2003,2004), and (2090,2091).

Wavelength regions used for modeling were 958–982 and 990–1038 nm. Eight (8) latent variables were determined as optimum number of PLS factors. The PLS_Toolbox (Eigenvector Research, Mason, Wa, USA) software was used for searching the optimal conditions on spectral alignment, pretreatment, wavelength

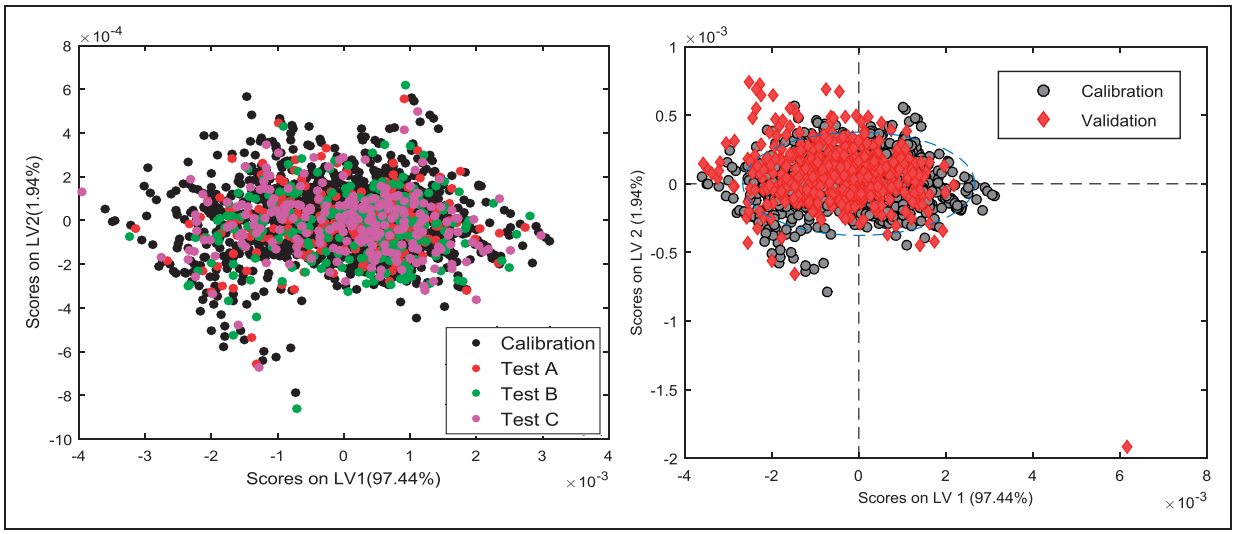

Figure 4. Score plot of the calibration, test, and validation set after standardization of the net analyte signals.

regions, and latent variable selection. The Unscrambler (CAMO, Woodbridge, NJ, USA) software was used for data and model performance visualization check.

PLS predictions were applied to all calibration, test, and validation sets after spectral alignment, truncation, and pretreatment with SNV–SG second derivative–SNV. No sample was removed in the prediction sets. Original order (without switch) of y-reference values in datasets was kept for the standard error (SE), the root mean square error (RMSE), and the bias calculation against predicted proteins. Reproducibility was calculated by averaging the 248 SD values, and each value was the SD of nine predictions on the same sample by nine different instruments in calibration and test sets. Reproducibility of 0.134% was reported. If the SD was calculated by three test sets only, the reproducibility would be 0.106%; further, it would be 0.077% if samples with ID 20140019, 20140020, 20140108, and 20140109 in the test sets were marked as outliers and removed.

Participant 3

The first step was to use the largest common wavelength range among instruments: 850–1048 nm. The second step of the calibration development was to have a look at the X data and to detect potential outliers and/or mistakes. As the data consisted in nine series of the same samples, it was easy to make spectral comparisons and subtractions to detect mistake in the sequences. For y outliers, models were calculated instrument by instrument. After this process eight spectra were removed from the nine series to keep all the series exactly identical. There were sequence errors in the spectra or inversion in the y reference values (samples removed in calibration and test; no. 17, 19, 20, 35, 36, 106, 107, 188).

After cleaning, the average of each of the nine sets has been calculated and merged. Plots of the spectra showed big differences in optical densities and their derivatives showed that spectra from instrument manufacturer C had an X-axis shift compared to the other instrument manufacturers. Based on the nine mean spectra, the external parameter orthogonalization[6] (EPO) method was applied to remove the differences between instruments. The first three PC were removed from all the data. If P is the matrix with the first three loadings of the principal component analysis based on the averaged spectra, the corrected matrix X (X_{cor}) is calculated as follows

$$X_{\mathrm{COR}} = X - XP^{T}P$$

A new calibration dataset was set up with the six calibration series from instrument manufacturers A and B. PLS models showed nonlinear responses. Thus, it was decided to apply a local approach from the package FOSS Winisi.[7] No other math treatments than EPO was used. The test set was predicted with a RMSEP of 0.17% and the reproducibility among the nine sets was 0.11%.

Participant 4

A PLS model was built in which the model optimized reproducibility and accuracy by performing the following spectral pretreatments: spectral truncation and resolution alignment, SG first derivative, and piecewise direct standardization (PDS).[8]

Each instrument manufacturer differed in spectral range and resolution. All spectra were truncated to wavelength range of the instrument with the narrowest range (850–1048 nm). Also the resolution of 2 nm was chosen to match the instrument with the lowest resolution. This first data treatment ensured that the same

spectral information (wavelengths) was being assessed across all the instruments.

The second spectral pretreatment was SG first derivative. Baseline differences were present between instruments of the same manufacturer and were even more pronounced between manufacturers. Derivative is a common method used to remove baseline signal, and SG algorithm is often used to simultaneously smooth the data to reduce the high frequency noise associated with taking derivatives on noisy data. Thus, spectral similarity between instruments and material manufacturers was improved after the second pretreatment.

PDS was the final spectral pretreatment. After baseline removal, peak shifts were still present particularly between manufacturers. This technique is based on selection of a primary spectrum and spectra from secondary units. In this case the primary spectrum was the mean of all spectra in the calibration set. The overall mean spectrum was used in order to reduce the bias associated with arbitrarily defining a standard manufacturer. Multiple regression models were built between the individual wavelengths of the primary spectrum and window of wavelengths of the secondary spectra. The resulting regression coefficients obtained from these models were then used to transform the calibration, test, and prediction sets spectra. The parameters of PDS involved a window size of five variables with two regression components. Residual plots of before and after applying PDS confirmed that the difference in spectra between manufacturers and instruments was reduced.

In the end, a PLS model with six latent variables was built with the pretreated spectra. An independent test set was present to evaluate the model performance. The prediction error (RMSEP) and the reproducibility were 0.564 and 0.161%, respectively. Note that no outlier was removed in this approach.

Participant 5

Band shift was identified as the major source of variability among the instruments under study and hence variable alignment was used as key strategy to achieve harmonization. After trimming the spectra to the common wavelength range (850–1048 nm), a fourth-order derivative (13-point window, fourth-order polynomial) was applied to the spectra. Choice of fourth derivative was important because it helped resolve chemical information (by emphasizing protein content and linearizing the corresponding spectral response) and also sharpen the peak positions that were essential for an optimal variable alignment.

A piece-wise spline alignment was used to adjust all the peaks in all the spectra (A, B, and C) to a target sample (e.g. sample 1 from instrument A).[9] SNV was then applied on the aligned derivatized spectra. This was proved essential most probably by correcting for the path length differences among individual samples.

Finally, a variable selection step was added to further improve transferability. This was incorporated in the PLS algorithm in the form of i-PLS (i.e. interval PLS).[10] Five latent variables were used for the final regression. The test set prediction error (RMSEP) and the reproducibility were 0.195 and 0.09%, respectively.

Participant 6

In 1984, Karl Norris and Phil Williams[11] published that derivative quotient math (DQM; as distinguished from individual absorbance math) would remove multiplicative scattering differences between samples due to particle size variation. Differences between instruments in the adjustment of absorbance scales and optical bandpass would also appear in the spectra as multiplicative effects, and therefore such instrumental differences would be removed. This technique was employed against this challenging shoot-out dataset.

To make the challenge greater, the calibration set was modified to only contain the data from the first instrument of manufacturer A, which would allow the evaluation of the calibration transfer to three units of the same type, four units of the second manufacturer (B), and one unit of the third manufacturer (C). There was no instrument standardization procedure used in this approach. The best calibration was a two-term equation using ratios of fourth derivatives with the wavelengths situated to monitor protein, water, and starch bands in the spectra.

There was close agreement between the results from manufacturer A and B, but no calibrations were found that would transfer well to manufacturer C. Inspection of the wavelength-scale calibrations of all instruments was possible using the peaks at about 915 and 984 nm in the wheat spectra themselves (Figure 5). Precise locations determined from the zero crossings of third derivative of the average spectra from each instrument indicated that the peak positions at the lower wavelength in the instruments from all three manufacturers agreed within about 2 nm, but at the higher wavelength the two units of manufacturer C were high by about 3 nm from the others. The DQM method was somewhat sensitive to the wavelength scale, and making a coarse correction to the wavelength range in the units from manufacturer C simply by removing a point at 938 nm and adding a point at the end to make all the spectra of equal length was sufficient to bring the RMSEP on the Test Set of manufacturer from 1.67 to 0.29% protein, while the other units displayed RMSEP values from 0.25 to 0.34% protein. The DQM method was relatively forgiving to variations in wavelength calibration but could not tolerate differences of 3 nm. In conclusion, it is suggested that after a more rigorous method is used to bring the instruments into wavelength agreement is applied, DQM calibrations would be excellent for assuring that instruments from multiple manufacturers perform equivalently while employing the same, simple calibrations.

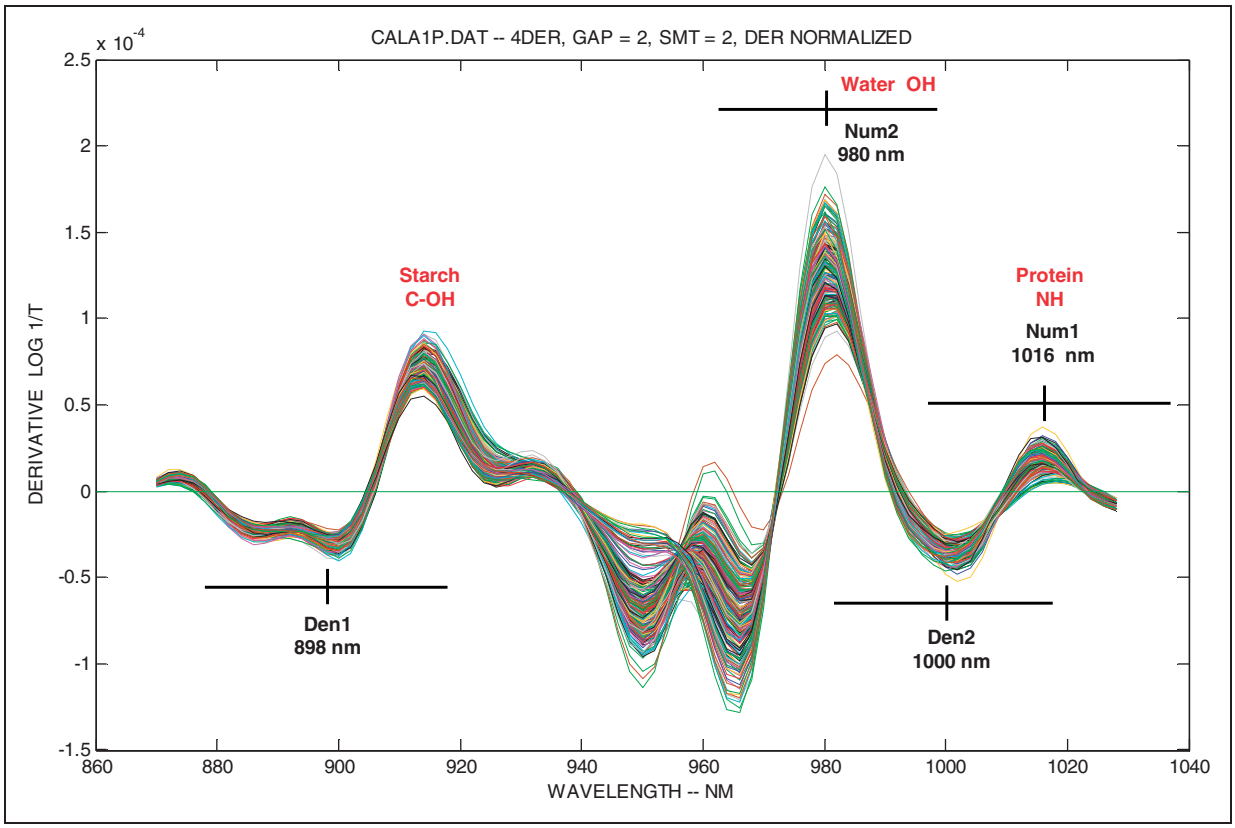

Figure 5. Fourth derivative spectra of scans from Unit #1 of manufacturer A, with crosses indicating the center and width of the convolution functions used in the calibration used for all the instruments.

Table 1. Blind prediction set results.

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Participant 6
Reproducibility	0.156	0.127	0.128	0.346	0.103	0.175
RMSEP	0.307	0.249	0.297	0.445	0.225	0.296
Bias	−0.078	−0.126	−0.113	−0.464	−0.055	0.032
r^2	0.979	0.986	0.980	0.958	0.990	0.982

Results

Table 1 presents the validation results for each participant. RMSEP and bias are presented along with coefficients of determination and the reproducibility values.

The participants chose quite different approaches to get prediction results that also varied significantly. With the overall best statistics, participant 5 won the 2016 IDRC Shoot-Out, followed by participant 2 and 3.

The data are available on the IDRC website (http://www.cnirs.org). The authors would like to thank the 2016 IDRC chair Dr David Funk and the Council for Near-Infrared Spectroscopy for providing funding for the shoot-out and support for the conference. The next conference will take place from 29 July to 3 August 2018.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Igne B, Dardenne P, Honigs D, et al. The 2010 IDRC software shoot-out at a glance. *NIRNews* 2010; 21: 14–16.
2. Igne B, Berzaghi P, Bu D, et al. Summary of the 2012 IDRC software shoot-out. *NIRNews* 2012; 23: 13–15.
3. Igne B, Bogomolov A, Bu D, et al. Summary of the 2014 IDRC software shoot-out. *NIRNews* 2014; 26: 8–11.

4. Wang Y, Veltkamp DJ and Kowalski BR. Multivariate instrument standardization. *Anal Chem* 1991; 63: 2750–2756.
5. Lorber A, Faber K and Kowalski BR. Net analyte signal calculation in multivariate calibration. *Anal Chem* 1997; 69: 1620–1626.
6. Roger JM, Chauchard F and Bellon Maurel V. External parameter orthogonalisation of PLS – application to temperature-independent measurement of sugar content of intact fruits. *Chemometr Intell Lab Syst* 2003; 66: 191–204.
7. Sinnaeve G, Dardenne P and Agneessens R. Global or local? A choice for NIR calibrations in analyses of forage quality. *J Near Infrared Spectrosc* 1994; 2: 163–173.
8. Griffiths ML, Svozil D, Worsford P, et al. The application of piecewise direct standardization with variable selection to the correction of drift in inductively coupled atomic emission spectrometry. *J Anal Atomic Spectrometry* 2006; 21: 1045–1052.
9. PLS_Toolbox Registerspec function help, http://wiki.eigenvector.com/index.php?title=Registerspec (accessed 12 November 2016).
10. Nørgaard L, Saudland A, Wagner J, et al. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectrosc* 2000; 54: 413–419.
11. Norris KH and Williams PC. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size. *Cereal Chem* 1984; 61: 158–165.