How do we do it: a brief summary of the methods we use in developing near infrared calibrations

Phil Williams and Debbie Sobering

Canadian Grain Commission, Grain Research Laboratory, 1404–303, Main Street, Winnipeg, Manitoba, Canada, R3C 3G8.

Introduction

Most of the near infrared (NIR) analysis carried out by the Canadian Grain Commission (CGC) concerns whole-grain analyzers, the Tecator Models 1225 and 1221, and the Foss/Multispec Grainspec. For these instruments, sample preparation is limited to removing foreign material (cleaning) and re-blending. Factory software is used to generate the calibrations, the "Calibration-maker" option of "Unscrambler" for the Tecator and "Tracker" for the Grainspec. These packages include systems for storing spectral and composition data, developing the calibration equations, using a combination of principal component analysis and partial least squares (PCA/PLS) regression, systems for validation and the computing of validation statistics (r^2 and *SEP*) and systems for detecting spectral outliers.

For these instruments at least 100 samples are used in calibration development, and preferably more, with further sample sets being used for validation. As well, about 5% of samples analyzed subsequent to the installation of a calibration are submitted to reference analysis and used for continuous monitoring. Of these, some may be selected to extend the calibration.

Outliers in analytical results are probably caused by differences in the interaction between sample and instrument, relative to the calibration constants. These differences mean that the optimum wavelength regimes for analysis of the individual (outlier) samples do not conform to the optimum wavelength regime for determination of protein, oil etc. represented by the calibration. In the use of whole-grain NIR transmittance instruments in our protein testing and segregation operation, outliers are rarely eliminated. The chief reasons are (i) their identification is based on spectral characteristics and in our experience the predicted results of these "outliers" are often within acceptable limits (±1.5 X the *SEP*), whereas samples with larger analytical residuals are often not detected as having any spectral aberrations and (ii) many of these analyses represent farmers' deliveries to primary elevators in the country, or railcars, which have been delivered to grain terminal elevators. The elimination of, say, 3% of the samples as "outliers" would mean eliminating 6000–7000 railcars from testing at delivery time, and could involve $15–20 million in protein premiums to the elevator companies. Were the same criteria for elimination to be applied to farmers' deliveries from trucks, the loss in potential revenue would also be about $20,000,000. So—we test everything we get!

The rest of this summary will refer to calibrations developed for the NIRSystems Model 6500 visible/NIR scanning spectrophotometer. All of our calibrations are based on analytical, rather than spectral data.

From Near Infrared Spectroscopy: The Future Waves

Calibrations

Calibrations for the scanning spectrophotometer are developed as follows:

We identify the sources of variance, as far as we can—this includes growing location and season, grade and class of grain (e.g. wheat, barley, flax etc.) and range in composition—we already have analytical data and select samples with as uniform a distribution of reference data as possible. Sometimes this involves very large sample sets, in order to accommodate the most comprehensive array of variance, while still ensuring a fairly uniform range of composition.

For conventional calibrations we use a combination of NIRSystems ISI and NSAS software. Spectra are sorted from lowest to highest reference values and divided into calibration and validation sample sets using the superior sample-managing facility of ISI. Sample sets may be assembled in ratios of calibration : validation spectra of 3 : 1, 2 : 1 or 1 : 1. Following this step, NSAS is used to optimize the treatment of the optical signal and wavelength range, using the AutoCal, stepwise multiple linear regression (MLR) option. There are four areas which require optimization: the mathematical treatment of the original log 1/*R* optical signal, the number of wavelength points required and the wavelength range and the method of correction for light-energy scatter due to sample characteristics.

The statistical methods we record in the evaluation of equations are the standard error of prediction (performance) or *SEP*, which is the standard deviation of differences between NIR and reference data, the coefficient of determination (r^2) the *RPD*, which is the ratio of the standard deviation of the reference data for the validation samples to the *SEP*, the *RER*,[1] which is the ratio of the range in reference data for the validation samples to the *SEP*, the coefficient of regression and intercept, the *RMSD* and the bias. The *RPD* and *RER* relate the *SEP* to the variance and range in the original reference data. The *RPD* should ideally be at least three and the *RER* at least 10. When the range, and therefore the variance in reference data is low, for example in the development of calibrations for the prediction of flour ash, the values for r^2 and the *RPD* cannot be very high. But if the *RER* indicates that the equation is capable of predicting the required values with an accuracy of at least one tenth of the range, this is regarded as acceptable for use in certain applications.

The mathematical treatments most frequently used by users of computerized scanning spectrophotometers are the log 1/*R* "raw" or smoothed and the first or second derivatives thereof. The "segment", or degree of smoothing, and the "gap", or dimension of the derivative, vary among commodities and parameters and the best results are obtained by determination of the combination of segment and gap which gives the most accurate and reproducible estimates. NSAS AutoCal allows the selection of up to nine wavelength points, depending upon the number of samples used in the calibration set. We begin by developing as many equations (wavelength points) as the software will allow, but at least four, for log 1/*R* and its first and second derivatives, using four wavelength points per segment and gap. This provides a very rapid (about 10 minutes) indication of the mathematical treatment likely to be the most successful. Further optimization is then carried out by changing the segment and/or gap for this algorithm. In the event of the preliminary search showing no preference between the three options, optimization of segment, gap and number of wavelength points is undertaken for as many as necessary. We prefer the number of wavelength points to be as low as possible, ideally one for derivatized signals, but this is rarely achieved. With whole-grain applications we have found that in general more wavelength points are required than for ground grain or any other form of powdered material applications, such as flour.

We usually employ the full wavelength range for the first optimization. This usually identifies the wavelength range between either 1100–2498 nm, or 400–1100 nm to be the most practical. The wavelengths selected for a commodity/constituent combination have involved the visible/NIR and NIR ranges sufficiently often to indicate that the ideal range for agricultural applications is

likely to be 500–1800 nm, which would enable analysis for coloured constituents such as chlorophyll and yellow pigment, as well as protein, moisture, oil and carbohydrate constituents.

Having arrived at the optimum mathematical algorithm and number of wavelength points, the equation is fine-tuned by re-visiting the wavelength range over which the equation has been developed. This may reveal that in the wavelength range of 1100–2498 nm, the lowest wavelength used was (for example) 1420 nm, while the highest was 2210 nm. The exercise is then repeated over a limited wavelength range, which accommodates the extremes of wavelength selected. This may or may not improve the efficiency of the equation.

Following optimization of the conditions using MLR, the principal component analysis/partial least squares (PCA/PLS) option of NSAS is employed to compute equations using the calibration sample set with untreated log 1/*R* data and with the optimized algorithm. The wavelength range is the same as the optimum identified during the MLR search. The NSAS version of PCA/PLS allows the computing of up to 15 factors. We prefer the number of factors to be as low as possible, and in our experience the most successful PCA/PLS calibrations incorporate from 3 to 10 factors. We have found PCA/PLS treatments to improve the accuracy and reproducibility of prediction for about half of the applications (over MLR).

NSAS provides no means for the evaluation of the application of any form of spectral manipulation, such as scatter correction. On the other hand, ISI software includes several options for determining the effect of scatter correction, but is not so expeditious as NSAS for rapid optimization of the mathematical algorithm and wavelength range. Once the preliminary optimization for these aspects has been completed using NSAS, ISI is employed to determine whether the efficiency of the equation can be improved by correcting for scatter. ISI offers five options for scatter correction. All of these are used in sequence using the calibration set with the optimized mathematical treatment to determine whether an improvement can be achieved. In our experience, improvements induced by the ISI scatter correction options are usually small. If one or more of the scatter-correction methods affords a significant improvement in *SEP*, the mathematical treatment is re-optimized, using ISI.

The final step in development of a reliable prediction equation is to determine the precision, or reproducibility of analysis with the new equation. This is accomplished by using the equation to predict the results of a precision file of samples of the same material. This should be developed during the calibration step by re-loading a check sample from time to time and by re-reading the re-loaded samples, without re-loading. If the sample cell permits, the cell should also be rotated, to provide three aspects of reproducibility of results—re-loading, sample cell rotation and re-reading of the sample without re-loading. When two or more equations differ from each other only subtly in *SEP*, the equation which has provided the best *SEP* may not provide the best precision and it is the equation which gives the best overall performance (accuracy plus precision) which should be used.

The above methods apply only to development of calibration equations for large sample sets. Sometimes only a small sample set is available for a preliminary search to determine the feasibility of NIR for an analysis. In these situations, ISI is employed and cross-validation is the statistical tool used to evaluate the application.

In the early days of applied NIR technology, if a sample set of, say, 40 samples was available, the accepted practice was to select about eight of the samples, representing the full range of composition to serve as validation samples and to develop the calibration equation using the rest of the samples. More recent research has shown that this approach can be misleading. The validation samples have been selected to maximize the range of composition, which tends to improve the correlation and *SEP* statistics "artificially" between NIR and reference data.

Table 1. Evaluation of NIR predictions by calibration/prediction and cross-validation in a small sample set of soft wheat flours.

Method	*N*	Constituent	r^2	*SEP*	*RPD*	*b*	*a*
Calibn/Predn	26/8	Strch dam.	0.85	1.13	3.4	1.117	–1.43
Cross-validation	34	Strch dam.	0.45	1.83	2.1	0.669	2.76
Calibn/Predn	26/8	Protein	0.96	0.07	8.4	0.714	2.71
Cross-validation	34	Protein	0.95	0.10	6.2	0.996	0.04

Cross-validation avoids this type of bias. The statistics may not be as flattering but give a more realistic estimate of the applicability of NIR to the analysis. An example of the difference between normal MLR (calibration/prediction) and cross-validation for a small sample set is given in Table 1, for flour starch damage (by the Farrand method) and protein in soft wheat flours. In this table the calibration/prediction data were achieved from a calibration set of 28 and a validation set of eight samples. For cross-validation all 34 samples were used. The calibration and prediction sets were selected after sorting the samples from low to high in composition. For cross-validation there is no need to sort the samples. The statistics favour the calibration/prediction procedure but the cross-validation statistics are a better representation of the true efficiency of the equation, using this small number of samples. This is particularly apparent in the case of starch damage. The data for protein indicate that the prediction of protein would likely be satisfactory by either method, although the cross-validation method is preferred. The chief implication from this table is that more samples should be included in the starch damage calibration exercise.

And that's how we do it!

Reference

1. C. Starr, A.G. Morgan and D.B. Smith, *J. Agric. Sci.* **97,** 107 (1981).