

LOCAL: a unifying theory and concept for near infrared analysis

John S. Shenk,^a Paolo Berzaghi^b and Mark O. Westerhaus^a

^aDepartment of Agronomy, The Pennsylvania State University, University Park, PA 16802, USA.

^bDepartment of Animal Science, University of Padova, Agripolis, 35020, Legnaro (PD), Italy.

Introduction

The search for a unifying theory of analysis began in 1972. We asked a simple question. Could the chemical and physical properties of a substance be measured from the light reflected from that substance? Today, it is embodied in an approach we call LOCAL. The theory is simple. It is that the visible/near infrared (vis/NIR) spectrum of a substance contains the information to define the basic chemical and physical properties of a material. LOCAL is a procedure that uses the spectrum of the unknown material, compares the spectrum to a library of information and predicts the chemical and physical composition of the unknown material. This paper will present the LOCAL theory and concept in detail. The application of this method resulted in an improvement of NIR predictions compared with traditional global calibration when applied to single product calibration. LOCAL was also tested with large databases that included a wide variety of agricultural materials. We obtained standard error of prediction (SEP) values similar, or lower than global calibration developed for each product. LOCAL not only improves accuracy, but it greatly simplifies NIR routine analysis.

The accuracy of NIR predictions depends on the structure and distribution of spectra in the calibration file.¹ In particular, the calibration database should include enough samples to cover most of the possible spectral variability encountered during routine analysis. Large calibration data sets are required to predict samples coming from very diverse geographical and climatic areas increasing the complexity of the spectral-reference model.

In theory, the best prediction method should compute a specific calibration equation for each sample analysed, using a small calibration data set tailored to the unknown sample from a large library of samples. This method, often called local calibration, would combine the advantage of global calibration in using one database to cover a large product domain, to the accuracy obtainable with specific calibrations.

The objective of this paper is to present a new procedure called LOCAL that performs rapid predictions applying the local calibration principle using large databases.

Materials and methods

LOCAL procedure

The LOCAL procedure (US patent no. 5,798,526) is designed to search and select samples in large databases that have spectra similar to the sample being analysed. The selected samples are then used to compute a specific calibration equation for predicting the constituents of an unknown sample.

The selection of calibration samples is controlled by the value of the correlation coefficient between the spectrum of the unknown sample and those of the database. The procedure computes all the

correlations between the spectrum of the unknown sample and each spectrum in the product database. The maximum number of samples used for calibration can be set manually by the operator or by a correlation cut-off value. If some of the selected samples have constituents with missing values, LOCAL will continue searching until it finds, for each reference value, the number of requested samples. This ensures that all the constituents will be predicted using the same number of samples, even if the library database has samples with incomplete reference values. If the sample being analysed is included in the calibration dataset, LOCAL will automatically discard it from the selected samples.

After defining the calibration dataset, the procedure uses partial least square (PLS) regression to compute the calibration equation. There are two methods of computation. The first method computes the prediction using a model with a number of PLS factors set by the operator (from 1 to 50). The second method calculates the predicted values using a weighed average of the predicted values generated with various numbers of factors. Weights are computed for each prediction based on how well the model fits the spectra of the unknown sample (X -residuals) and the root mean square (RMS) of the regression coefficients for each factor.

The formula to calculate weights is as follows:

$$\text{Weight}_{1..k} = 1/(\text{rms-residual}_{1..k} \times \text{rms-beta}_{..k})$$

where:

k = number of PLS terms set in the algorithm;

$\text{rms-residual}_{1..k}$ = root mean square of spectral data residual;

$\text{rms-beta}_{1..k}$ = root mean square of regression coefficients of spectral data.

An additional option to this second method is to exclude from the computation of the final prediction, the values obtained with the first PLS factors that usually generate poor predictions. Global and neighborhood H values for the unknown sample are calculated with the PLS scores using the same weights computed for the different models used in the weighed average described previously. To increase the speed of prediction, spectra of the calibration file were reduced to 57.

Databases

The study used 16 feedstuffs, blocked in four databases called Forage, Grain, Meat and Feed, for a total number of samples of 6599. All the silage and TMR samples were dried either in a microwave or in a forced-air oven at 60°C. All the samples were ground through a 1 mm screen and analysed for dry matter (DM) and crude protein (CP).² Forage and grain samples were also tested for ADF,³ while samples included in the Meat and Feed database were analysed for fat content.²

Samples were scanned between 1100 and 2500 nm every 2 nm, using a Foss NIRSystem 5000 monochromator instrument (Silver Spring, MD, USA). The data set of each feedstuff was split into a calibration and a validation file. Using the Centre algorithm,¹ samples were ranked according to their Mahalanobis distance (H) to the average sample. Samples with an H value greater than three were considered outliers and discarded. The validation files were generated selecting every tenth sample of the centred file, with the remaining samples entering the calibration data set.

Math treatments, calibrations and validations

The weighed average method was used to calculate LOCAL predictions, excluding from the computations the values generated with the first few PLS factors.

In order to perform LOCAL predictions, it is necessary to define three parameters ($\text{LOCAL}_{(k, l, m)}$): the number of samples selected from the calibration database (k), the maximum number of PLS factors (l), and the number of predicted values of the first PLS factors to be excluded from the computation of the final prediction (m).

A preliminary test was performed to determine the best LOCAL settings for each database. The test consisted of the prediction of each sample in the databases by LOCAL, changing the number of

Table 1. Standard error of prediction using LOCAL or Global calibrations

	LOCAL within database	LOCAL one database	Global equation	
			Generic ^a	Specific ^b
FORAGE _(200, 35, 4)				
Protein	0.680	0.680	0.758	0.720
ADF	1.470	1.493	1.672	1.539
Dry Matter	0.775	0.769	0.896	0.790
GRAIN _(150, 20, 6)				
Protein	0.343	0.374	0.424	0.337
ADF	0.729	0.756	0.886	0.700
Dry Matter	0.493	0.484	0.591	0.443
FEED _(100, 25, 6)				
Protein	0.805	0.808	1.215	0.866
Fat	0.387	0.406	0.443	
Dry Matter	0.470	0.503	0.598	0.516
MEAT _(100, 25, 6)				
Protein	1.080	1.081	1.387	1.301
Fat	0.612	0.601	0.621	0.583
Dry Matter	0.266	0.266	0.329	0.286

^aGLOBAL Generic: Global calibration using the entire Forage database^bGLOBAL Specific: Global calibration using each single product database^cLOCAL settings: Number of selected samples, number of PLS factors, number of PLS factors discarded

samples selected (from 50 to 350 with steps of 50), the number of PLS factors (from 15 to 40 with steps of five) and the number of PLS factors excluded (from four to ten with steps of two) in the computation of the weighed average, for a total of 168 combinations. The best setting for each database was then used in the prediction of the validation file. Default math pre-treatments were 1,4,4 (derivative, gap and smoothing in nm) with *SNV* and detrend scatter correction.⁴ Then all the spectra used in LOCAL were reduced to 57 datapoints by combining the spectra every 24 nm in the interval between 1124–2468 nm.

Global calibrations were performed (WinISI, Infrasoft International LLC, Port Matilda, PA, USA) for each feedstuff and for each multi-product database, using the same math pre-treatments used in LOCAL, the only difference being that spectra were reduced to 169 data points. Validation files of each feedstuff were then predicted using their specific equation and with the equation of the database the feedstuff was classified. *SEP*, using global equations and LOCAL procedure, were compared.

Results and discussion

On average, using LOCAL procedure improved NIR predictions for Forage, Feed and Meat databases. LOCAL had *SEP* values slightly greater than the traditional global calibrations for grain only. Further details on this evaluation of LOCAL performances can be found in the literature.^{5,6} As could be expected, *SEP* values generated using one single database were similar to those obtained using the four databases and with the product-specific global calibrations (Table 1). The capability of LOCAL to select samples with valuable information is not affected by the number and variety of products included in the database. Actually, because very diverse products were included in the database, LOCAL could probably easily select calibration samples with spectra more similar to the sample being predicted.

Even if LOCAL had to select samples through a large database, computation of predictions was very rapid. On a Pentium II™ 300 MHz personal computer, the speed of LOCAL_(150, 20, 6) prediction for the three parameters (Protein, ADF and DM) of grain samples was 0.6 s per sample.

Conclusions

LOCAL clearly showed the capability of predicting the constituent composition of a large variety of feeds as well as, or better than, traditional global calibration methods. Compared with some of the previously⁷⁻⁹ proposed local calibration techniques, LOCAL has a greater speed of prediction that allows this procedure to be implemented in routine analysis. One of the most important features of LOCAL is the rapid and efficient selection of samples used in the calibrations. This result shows the great selecting capacity of LOCAL and it opens new perspectives on NIR analysis. Laboratories that use NIR on a daily basis can reduce the number of calibration data sets to be maintained into a few, or even one single, database. Because LOCAL automatically selects the samples and creates a calibration equation, it is not necessary to identify the type of sample being analysed simplifying the analytical process. All these LOCAL features can radically change the operational system of NIR networks. In fact, it would be possible to have remote NIR instruments which simply collect NIR spectra and send them to a central NIR laboratory which stores and develops databases. These NIR laboratories can compute predictions even without knowing the type of sample being analysed. Remote instruments can be used by an operator with very little training and the NIR expertise are required only at the central laboratory.

References

1. J.S. Shenk and M.O. Westerhaus, *Crop. Sci.* **31**, 469 (1991).
2. AOAC, Association of Official Analytical Chemists, 15th Edition (1984).
3. H.K. Goering and P.J. Van Soest, *Forage fiber analysis (apparatus reagents, procedures, and some application)*. Agric. Handbook No. 379, ARS-USDA, Washington DC, USA (1970).
4. R.J. Barnes, M.S. Dhanoa and S.J. Lister. *Appl. Spectrosc.* **43**, 772 (1989).
5. J.S. Shenk, P. Berzaghi and M.O. Westerhaus, *J. Near Infrared Spectrosc* **5**, 223 (1998).
6. P. Berzaghi, J.S. Shenk and M.O. Westerhaus, *J. Near Infrared Spectrosc* **8**, (2000).
7. A.H. Aastveit and P. Marum, *Appl. Spectrosc.* **47**(4), 463 (1993).
8. G. Sinnaeve, P. Dardenne and R. Agneessens, *J. Near Infrared Spectrosc.* **2**, 163 (1994).
9. A.M.C. Davies, H.V. Britcher, J.G. Franklin, S.M. Ring, A. Grant and W.F. McClure, *Mikrochim. Acta [Wien]*. **1**, 61 (1988).