Multivariate calibration and chemometrics for near infrared spectroscopy: which method?

Pierre Dardenne, George Sinnaeve and Vincent Baeten
Département Qualité des Productions Agricole, Centre de Recherches Agronomiques de Gembloux–CRAGx, 24 Chaussée de Namur, B-5030 Gembloux, Belgium

The four most important regression methods are evaluated on very large data sets: Multiple Linear Regression (MLR), Partial Least Squares (PLS), Artificial Neural Network (ANN) and a new concept called "LOCAL" (PLS with selection of a calibration sample subset of the closest neighbours for each sample to predict). The Standard Errors of Prediction (*SEP*s) are statistically tested and the results show that the regression methods are almost equal and that the data matrices are more important than the fitting methods themselves. The types of pre-treatments (Multiplicative Scatter Correction, Detrend, Standard Normal Variate, derivative etc.) of the spectra are too numerous to be able to test all the combinations. For each test, the pre-treatment found as the best with the PLS method is fixed for the other ones. The second part of the paper emphasises the importance of the number of samples. If any agricultural commodity, and probably any kind of product measured by an NIR instrument, can be considered as a mixture of several constituents, the databases built by collecting actual samples bringing new information can reach hundreds, if not thousands, of samples.

Keywords: regression methods, MLR, PLS, ANN, LOCAL, mixture model, Global *H*, Neighbourhood *H*

Introduction

Near infrared (NIR) spectroscopy is widely used as a quantitative method and the main multivariate techniques consist of regression methods used to build prediction models. Discriminant analysis and pattern recognition are related to qualitative analysis and are used less intensively than the regression methods. There are many techniques to achieve the same goal and the user often finds it difficult to choose the most appropriate one.

The statistical packages available to manage the spectral data usually offer several procedures based on different mathematical algorithms and chemometric tools. Amongst the most currently used calibration (or regression) procedures are Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares regression (PLS), Local Weighted Regression (LWR), Ridge Regression and regression methods based on Artificial Neural Network methodology (ANN).[1–5] Whereas the linear methods assume that the relationship between the independent and dependent variables are linear in nature, they are able to cope with non-linear relationships.[6] Other regression methods such as Genetic Algorithm,[7] Uninformative Variable Selection[8] and Interactive Variable Selection[9] are also used in multivariate calibration. These methods are more oriented to an elimination of part of the wavelength range.

The aim of this paper is to compare four regression techniques (MLR, PLS, ANN and "LOCAL"). NIR spectral databases obtained from three agricultural products (fresh grass silage, wheat and whole plant maize) and from two food products (meat and apple) were used. The databases consist of between 380 and 2400 samples from different years of production and were obtained with several NIR instruments. In addition, the number of samples needed for robust calibrations and the complexity of the spectral multidimensional space are illustrated through mixture design.

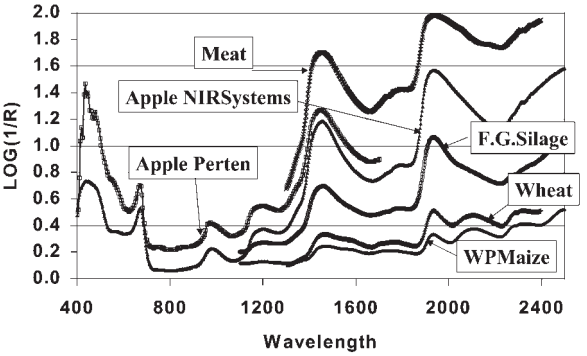

Figure1. Average spectra of the six data sets as defined in Table 1.

Data set description

The data sets used for the comparisons of the regression methods consist of six sets of samples. The products have been chosen to cover the whole range of optical densities that we can expect in reflection mode. Table 1 reports the products and the parameters to be predicted with the statistics of the reference value distributions. Figure 1 represents the average spectrum for each set. The first set is 1000 samples of grass silage measured fresh on a NIRSystems 5000 (NIRSystems-Perstorp, Silver Spring, MD, USA) equipped with a transport module and within the range 1100–2498 nm. The second is 2400 wheat samples measured by the Belgian network between 1990 and 1998 on six NIRSystems 4500 instruments within the range 1300–2398 nm. The wheat samples are ground on a Cyclotec mill (1 mm), filled in mini-ring cups and placed on a rotating drawer. The third set consists of 2250 whole plant maize samples collected from Europe and the USA. The samples are dried at 70°C and ground. The spectra were collected from several NIRSystems 5000 and 6500 instruments in Europe and USA. The fourth database concerns minced meat (pork, beef and chicken) filled in mini-ring cups and scanned on a NIRSystems 4500. Each spectrum is the average of three sub-samples. The last two sets contain the spectra of apples measured whole, either on a NIRSystems 6500 equipped with the DCA (Direct Contact Analyser) module, or on a Perten DA7000 (Perten Instrument, IL, USA), which is a diode array instrument. The wavelength range of the

Table1. Data set description and statistics of the reference analyses.

Data sets	Constituent	*N*	Min	Mean	Max	*SD*
1. Fresh Grass Silage	Protein	1000	6	14.3	23	3.1
2. Wheat	Protein	2400	8	12.1	19	1.3
3. Whole Plant maize	Crude fibre	2250	4	7.9	14	1.2
4. Meat	Fat	650	0.5	6.0	35	7.2
5. Apple – NIRSystems	Brix	775	8	12.3	17	1.5
6. Apple – Perten DA7000	Brix	380	9	13.6	19	1.4

N = number of samples; Min = minimum value; Mean = average value; Max = maximum value; SD = standard deviation of the reference values

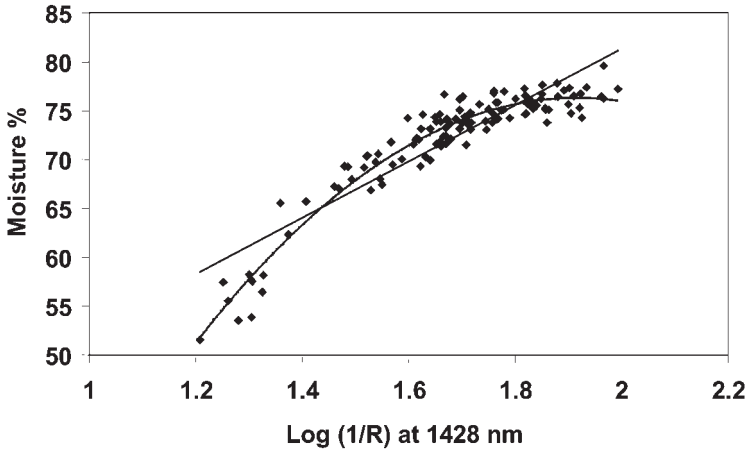

Figure 2. Response of the meat samples at 1428 nm as a function of the moisture content.

NIRSystems 6500 is 400–2498 nm, while the range of the Perten DA7000 is limited to 400–1700 nm with a 5 nm step. In both cases, each spectrum is the average of four scans acquired around the equatorial faces of the fruits. The optical densities cover the whole range between 0.1 and 2.0 units.

Products with high moisture levels display non-linearity and then the regression methods have difficulty in solving the problem. As an example, Figure 2 shows the response at 1428 nm of the meat samples as a function of moisture content.

The number of combinations to test becomes very numerous when pre-treatments are associated with the regression methods. The current pre-treatments found as best for each set using PLS are kept unchanged for the other methods. We know that this gives a certain advantage to the PLS results, but it was a formidable, and even unrealistic, task to test all the possibilities. The math treatments were as follows. Fresh Grass silage: no pre-treatment followed by a 2–5–5 derivative (2^{nd} derivative with 5 point gap and 5 point segment). Wheat: SNV-Detrend followed by a 1–5–5 derivative. Whole plant maize: SNV-Detrend followed by a 1–5–5 derivative. Meat: no pre-treatment followed by a 2–5–5 derivative. Apple NIRSystems: no pre-treatment followed by a 2–10–5 derivative. Apple Perten: no pre-treatment and no derivative. It was observed that the high moisture products do not like scatter correction pre-treatments such as MSC, SNV and Detrend. The influence of the water peak is too important to make a good correction in the parts of the spectrum which are less influenced by water and where the most information is. On the other hand, scatter correction is appropriate for powders with variable particle size.

Regression methods

All the computing work was done using the routines included in the ISI software (Infrasoft International, Port Matilda, PA, USA). The data sets were randomly split into two parts: 4/5 for the calibration and 1/5 for the validation. MLR models are developed following a stepwise procedure which tests the permutation at each stage before entering the next variable. In all cases, the minimum *F* test on the regression coefficients has been fixed at 12 and the whole wavelength ranges have been presented for the variable searches. With PLS, four cross-validation groups were used to fix the number of terms and, among the *X* variables, every 4^{th} data point has been used (8 nm intervals) except for the Perten instrument for which all the data points have been kept since its resolution is already 5 nm. The number of parameters in an artificial neural network is very large and in this case, as we did for the pre-treatment, we fixed most of them, even if we know that we can miss the optimal solution. The input layer consists of the PLS scores and the number of factors have been fixed at a level found in the normal PLS procedure. The number of hidden elements was three, with a sigmoid function between the input layer and the hidden layer and a linear function between the hid-

Table 2. Data set description and comparisons of the *SEP*s.

Data sets	Constituent	Np	MLR	PLS	ANN	LOC	Nmin
1. F.G.Silage[a]	Protein	200	1.08[a]	0.90[b]	0.95[b]	0.91[b]	60
2. Wheat	Protein	500	0.27[a]	0.22[b]	0.21[b,c]	0.20[c]	30
3. W.P. Maize[b]	Crude fibre	450	1.02[a]	0.97[b]	0.96[b]	0.86[c]	60
4. Meat	Fat	130	0.36[a]	0.35[a]	0.32[a,b]	0.29[b]	40
5. Apple NIRSyst.	Brix	155	0.64[a]	0.60[a,b]	0.55[b]	0.60[a,b]	90
6. Apple DA7000	Brix	75	0.57[a]	0.50[a]	0.40[b]	0.60[a]	15

[a]Fresh Grass Silage; [b]Whole Plant Maize
Np = number of independent samples of the test sets. Nmin = number of samples for which all the four methods can not be defined as statistically different

den layer and the output layer. The training sessions were restarted five times with different random seeds. Since ANN needs an internal stop set to freeze the best solution, the calibration sets have also been randomly divided in the proportion 3/4 for a training set and 1/4 for a test set to keep the original validation set unchanged. In the "LOCAL" procedure, each "unknown sample" is spectrally compared with the spectra of a library and the *N* nearest neighbours are selected to build a temporary file. The similarity index is the correlation between the spectra of the library and the "unknown sample". A PLS model is developed on this temporary file. No cross-validation is used, to speed up the system since the algorithm must work in real time with an instrument connected to the PC. The main feature of the procedure is the fact that the final predicted value is weighted according to the standard deviation of the regression model B coefficients and to the residuals on the *X* variables, which can express how well a spectrum can be reconstructed from the PLS scores and loadings. A main advantage of the "LOCAL" method is the use of the spectrum of the "unknown sample" as a guide to reach a better prediction. The two main parameters to set up when applying "LOCAL" are the number of samples to select and the number of PLS terms. For each data set, an optimisation design was set up by varying the number of samples from 50 to 250 in steps of 25, and the number of terms from 12 to 36 in steps of three. This gives a factorial design of 9×9 or 81 runs. The results are entered and a response surface is fitted to find the exact minimum of the *SEP*. The response surface has been computed using Unscrambler® 7.5 (Camo ASA, Norway) and the figure plotted by Minitab® (Minitab Inc., PA, USA). Figure 3 gives the response surface of this design for the meat data set for which the observed minimum is reached with 125 samples and 24 PLS terms.

To be able to carry out the *SEP* comparisons we did not remove any outliers (neither *T* nor *H* outliers).

SEP comparisons

Table 2 shows the *SEP* values for the six data sets. Statistical tests have been performed on a pairwise comparison of *SEP*s for the data sets, following the procedure published by Fearn.[10] Based on the standard deviations [or the *SEP*(C), *SEP* corrected for bias] of the residual vectors coming from two models and on the R^2 between these two vectors, we can define the lower and upper limits of a 95% confidence interval for the ratio of the standard deviations following the formula 1, 2, 3 and 4. *S* is the standard deviation of the *n* errors for each method.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(e_i - m)^2}{n-1}} \qquad (1)$$

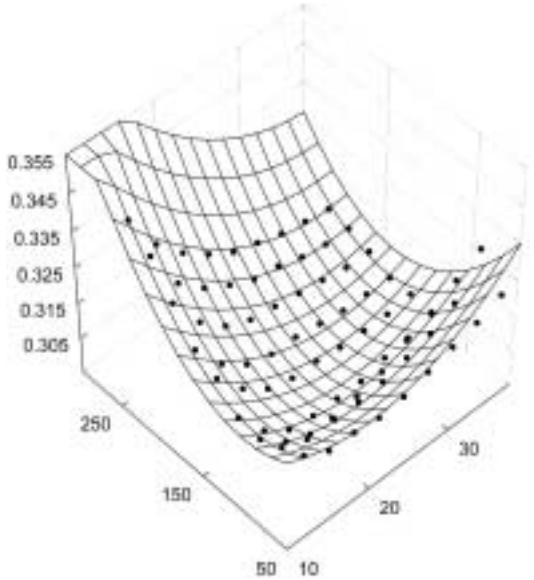

Figure 3. Response surface of the meat data with the combinations of the number of samples (NS) and the PLS terms (NF).

where e_i is the error on sample i (= 1, …, n) and m is the mean error or bias.
Two intermediary variables K and L are calculated:

$$K = 1 + \frac{2(1-R^2)t^2_{n-2,0.025}}{n-2} \qquad (2)$$

where $t^2_{n-2,0.025}$ is the upper 2.5% point of a t distribution for $n-2$ degrees of freedom and R^2 is the square of the correlation between the error vectors by the two methods,

$$L = \sqrt{K + \sqrt{(K^2-1)}} \qquad (3)$$

Then, the two next functions define the lower and upper confidence limits,

$$\frac{S_1}{S_2} \times \frac{1}{L} \quad \text{and} \quad \frac{S_1}{S_2} \times L \qquad (4)$$

Based on these formulae, the comparisons of the *SEP*s lead to the determination of the number of *SEP* pairs with non-significant differences (marked as superscript a, b and c in Table 2). This means that two *SEP*s marked with different superscript letters are significantly different or, conversely, two *SEP*s marked with the same letters are not significantly different. Globally, MLR is the worst method, followed by PLS whereas ANN and LOC give the best and similar results. On the first sample set, *SEP* from MLR is different to the three others and these three are not statistically different. Notice that for the last set, the differences between PLS-LOC and PLS-ANN are 0.1 in each case. PLS and LOC are "equal", but PLS and ANN are different, since the R^2 between the corresponding residual vectors are not the same.

We can see that the PLS and even MLR produce good accuracy and the analysis of the residuals does not reveal any non-linearity. The linear methods can handle some non-linear responses providing the combination of the predictors define a plane in the multivariate space.[11] The tests made with large data sets allow us to conclude that ANN and the "local" algorithm give better results. The local algorithm needs enough information and the number of samples in the data base is important to be able to define subsets with a sufficient number of closest neighbours.

The last column of Table 2 contains the number of samples in the data set which is the minimum (keeping the *SEP*s and the R^2s constant) to reach the threshold of significant differences between the two extreme *SEP*s in each data set. In many publications dealing with regression method comparisons, these statistical tests are not often used and conclusions are taken on too small data sets.

Calibration updates and regression coefficient stability

Using cross-validations [Standard Error of Cross-Validation (*SECV*)] on a data set generally gives an over-optimistic idea of the actual performance of the models. When new and totally independent samples are predicted with a "young" calibration, it is very rare to get an *SEP* at the same level as the *SECV*. In most of the databases, the samples, even if they are "different" (of course not the replicates), are processed in batches. For example, from experimental plots, a breeder can submit 50 samples from the same field. The 50 samples (different varieties) are harvested on the same day, dried at the

Table 3. Interim models for starch content in whole plant maize samples.

Mod	*N*	Min	Max	*SEC*	R^2C	*SECV*	R^2V	Tpls
90	198	2	48	1.26	.99	1.69	.98	8
91	245	2	48	1.37	.98	1.70	.97	9
92	522	2	53	1.37	.98	1.62	.98	13
93	709	2	53	1.59	.98	1.74	.97	10
94	1125	2	53	1.61	.98	1.74	.97	11

N = number of samples included in the calibration sets; Min = minimum value; Max = maximum value; *SEC* = standard error of calibration; R^2C = determination coefficient of calibration; *SECV* = standard error of cross-validation; R^2V = determination coefficient obtained by cross-validation; Tpls = number of PLS term in the model

Table 4. Validation of the interim models on the succeeding years for starch content.

Model	Val	*N*	*SEP*	%*H* > 3
90	91–95	1779	2.80	22
91	92–95	1733	2.47	21
92	93–95	1448	2.41	13
93	94–95	1259	2.31	11
94	95	831	1.89	10

Val = validation sample set; *N* = number of samples; *SEP* = standard error of prediction; %*H* > 3: percentage of the validation samples with a Mahalanobis distance higher than 3

same temperature in the same oven, ground on the same grinder and measured on one instrument by the same operator. The information the samples carry is not completely independent from each other because they have been processed together. Once included in the calibration set with previous other samples, they will be split automatically in the calibration segment and in the cross-validation segment and then the *SECV* is estimated with an over optimistic value.

In the experiment presented, we use totally independent test sets, which is the only way to have a determination of the actual performance of the models. Having whole plant maize samples collected from 1986 to 1995, allowed interim models to be evaluated with the samples of the succeeding years. More details of this experiment can be found in the article by Dardenne.[12] Table 3 gives the results of the calibration for starch and Table 4 the results of the validation for each interim model. Figure 4 shows the typical behaviour of the *SEC*, *SECV* and *SEP* with the interim models for four constituents. In the calibration process, the *SEC*s increase with the number of samples, whereas the *SECV*s increase as well, but at a lower rate. The differences between *SEC* and *SECV* become smaller when the numbers of samples increase. The number of terms used shows a tendency to increase according to the number of samples. The R^2 (R^2C and R^2V) remain almost constant. Calibration models can be viewed as stable when the

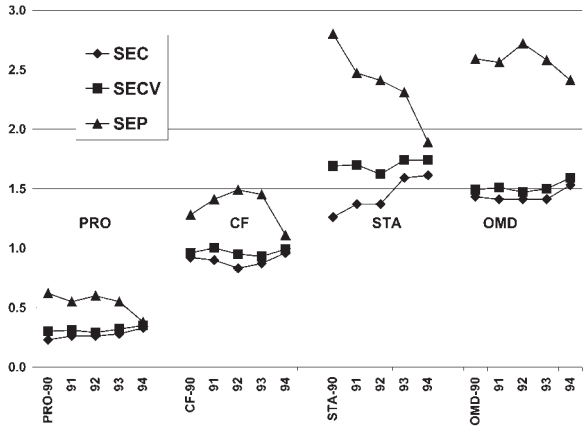

Figure 4. Trends of *SEC*, *SECV* and *SEP* with the interim models for four constituents (PRO = protein, CF = crude fibre, STA = starch, OMD = organic matter digestibility).

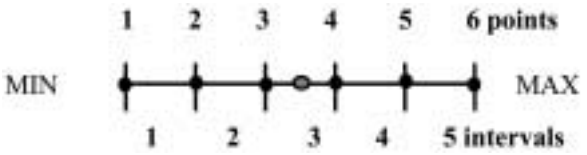

Figure 5. Binary mixture model: 6 points, 5 intervals and standardised neighbourhood distances of 0.66.

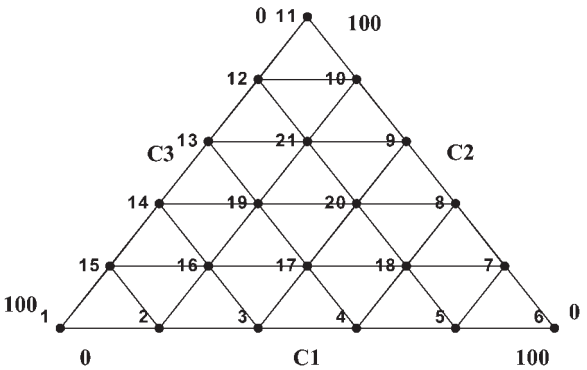

Figure 6. Ternary mixture model: 21 points, 5 intervals on each edge and a standardised neighbourhood distance between the closest point equal to 0.60.

SEP values are close to the corresponding *SECV* values.

Statistical tests, as explained previously, can be used to compare *SEC* and *SECV*, and when they are significantly different we can expect to have to add new information coming from future routine samples. Between *SEP* and *SECV*, the statistical tests exist, but they are different since the objects (samples) are not paired. A full explanation of the statistical tests can be found in the article by Shenk *et al.*[13]

Mixture model

For agricultural products, it becomes obvious as a result of the previous experiment and from the literature[12,14] that many samples are needed to cover all the sources of variation: varieties, climate, fertilisation, water availability, harvest stages, sample preparation etc. Based on a theoretical view, a binary mixture can be represented by only one dimension (Figure 5). When the range of the two constituents is divided by five, six points are defined and the neighbourhood standardised distance between the closest points is 0.66. Standardised distance means that the average distance of the distances between all the points and the centre is one. A mixture of three components (such as meat, with protein, fat and water) can be defined with two dimensions. As the sum of the three components is always 100, the design is the well-known triangle (Figure 6). If each edge is divided into five intervals, 21 points are needed to cover the space and the standardised distance between the closest points is 0.60. A design with four components can be drawn with a tetrahedron (three dimensions) (Figure 7). When the edges are divided into two intervals, 15 points cover the space: four corners, six edge centres, four face centres and the overall centroid. When each edge is divided into five intervals, 56 points are necessary to evenly fill the volume and the standardised distance between the closest points is 0.57.

The sequence can be continued with five components: this four-dimensional design is more difficult to draw and the number of points is 126. The number of points can be calculated with the combination formula below, where p is the number of components ($p - 1$ dimensions) and r the number of intervals for each constituent:

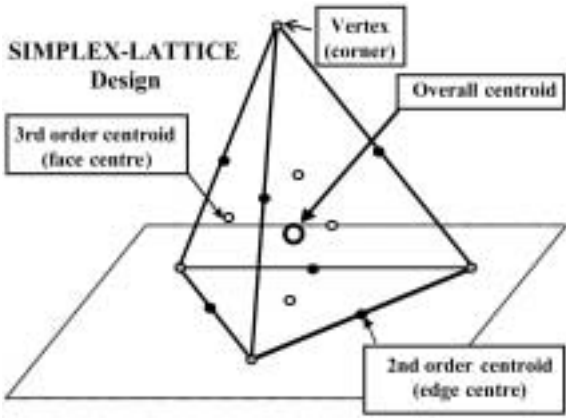

Figure 7. Quaternary mixture model: 15 points with two intervals on each edge; 56 points with five intervals and a standardised neighbourhood distance between the closest point equal to 0.57.

$$C_{r+p-1}^{p-1} = \frac{(r+p-1)!}{r!(p-1)!}$$

Table 5 gives the number of samples (or points) to cover the space as a function of the number of components into a theoretical mixture. We can notice that the number of points increases very quickly with the complexity of the mixture. Of course, a full factorial design restricted to a mixture model is certainly not necessary in all the cases. Fortunately, otherwise NIR spectroscopy would never have produced results. In the case of total linearities, the corner points would be enough to calculate the models. But as soon as the shapes of the relationships are not well known, and especially when the "LOCAL" concept is applied, the spectral space must be evenly filled.

On the basis of only the main constituents, a forage sample can be considered as a mixture of water, ash, proteins, fibre, starch and sugar. But these constituents consist of hundreds of different molecules. In practise, a forage database can be reduced to 15 to 30 principal components including the variation coming from the sample preparation. Considering the complexity of such mixtures, it is not surprising to see how it is difficult to select samples to cover all the variations, and many years, as well as thousands of screened samples, are needed to obtain robust calibrations.

Table 5. Relationship between the number of constituents in a theoretical mixture and the number of points necessary to cover the space with five intervals on each constituent.

Const.	Dim.	Nspl	Const.	Dim.	Nspl	Const.	Dim.	Nspl
3	2	21	13	12	6188	23	22	80730
4	3	56	14	13	8568	24	23	98280
5	4	126	15	14	11628	25	24	118755
6	5	252	16	15	15504	26	25	145506
7	6	462	17	16	20349	27	26	169911
8	7	792	18	17	26334	28	27	201376
9	8	1287	19	18	33649	29	28	237336
10	9	2002	20	19	42504	30	29	278256
11	10	3003	21	20	53130	31	30	324632
12	11	4368	22	21	65780	32	31	376992

Const. = number of constituents; Dim. = dimension; Nspl = number of samples

Conclusion

A first conclusion is that multivariate linear regression methods can deal with non-linearity and produce accurate models for the analysis of agricultural products. The differences between the methods (MLR, PLS, ANN and "LOCAL") are quite small and more work and time must be dedicated to collecting the right spectra and good reference values than to testing several methods. Also, when methods are compared, the set of samples must be wide enough to be able to compare the results. Adequate statistical tests on the *SEP*s coming from sets of paired samples must be done before reaching any conclusion.

A second comment is the difficulty of obtaining robust and stable equations over time. During the calibration process, the gap between *SECV*s and *SEC*s informs the user that he will have to update the calibration when new samples arrive. New samples must be added to the calibration data base while the *SECV* remains higher than the *SEC* and of course while the *SEP* is higher than the *SECV*. In the latter case, other statistical tests can be done to compare *SECV* and *SEP*. With big databases, it has been shown that *SEC*, *SECV* and *SEP* can be very close providing enormous effort has been taken in selecting samples and determining the reference values.

Based on theoretical mixture models, the complexity of the agricultural products has been shown and thus the need for wide databases. If it is possible to define the population boundaries, it is much more difficult to fill all the intervals in the multidimensional space. We believe that a calibration data base can always be improved and should be open to new samples, especially with the concept of the "LOCAL" calibration.

References

1. M. Blanco, J. Coello, H. Iturriaga, S. Maspoch and C. de la Pezuela, *J. Near Infrared Spectrosc.* **5,** 67 (1997).
2. G. Sinnaeve, P. Dardenne and R. Agneessens, *J. Near Infrared Spectrosc.* **2,** 163 (1994).
3. I.E. Frank and J.H. Friedman, *Technometrics* **35,** 109 (1993).
4. T. Næs and T. Isaksson, *Appl. Spectrosc.* **46(1),** 34 (1992).
5. H. Martens and S.A. Jensen, in *7th Proc. World Cereal Congr.*, Ed by J. Holas and J. Kratochvil. Elsevier, Amsterdam, The Netherlands, pp. 205–235 (1983).
6. S. Sekulic, M.B. Seasholtz, Z. Wang, B.R. Kowalski, S.E. Lee and B.R. Holt, *Anal. Chem.* **65(19),** 835A (1993).
7. D. Jouan-Rimbaud, D.L. Massart, R. Leardi and E. de Noord, *Anal. Chem.* **67(23),** 4295 (1995).
8. V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste and C. Sterna, *Anal. Chem.* **68,** 3851 (1996).
9. F. Lindgren, P. Geladi, S. Rannar and S. Wold, *J. Chemometrics* **8,** 349 (1994)
10. T. Fearn, *NIR news* **7(5),** 5 (1996).
11. H. Martens and T. Næs, *Multivariate Calibration*. John Wiley & Sons, Chichester, UK (1989).
12. P. Dardenne, *NIR news* **7(5),** 8 (1996).
13. J.S. Shenk, M.O. Westerhaus and S.M. Abrams, in *Near infrared reflectance spectroscopy (NIRS): analysis of forage quality*, Ed by G.C. Marten, J.S. Shenk and F.E. Barton, II. United States Department of Agriculture, Agriculture Handbook No. 643, p. 104 (1989).
14. T. Næs and T. Isaksson, *NIR news* **5,** 16–17 (1994).

Received: 10 September 1999
Revised: 10 August 2000
Accepted: 17 October 2000
Web Publication: 7 December 2000