New approach for calibration transfer from a local database to a global database

Pierre Dardenne
CRAGx, 100 rue de Serpont, 6800 Libramont, Belgium.

Roland Welle
Pioneer Hi-Bred, Apensener Str., 198 D-21614 Buxtehude, Germany.

The goal of the experiment is to transfer a local model to a global model for the prediction of a new parameter. Station de Haute Belgique (SHB) has developed an extensive spectral database to determine 10 parameters of whole plant maize silage samples. Pioneer has determined the Tilley and Terry digestibility (T&T) on a smaller set of samples. The analysis of the Mahalanobis distances between the two sets shows that Pioneer's set has a much smaller variation than SHB's set. Pioneer's set can be predicted by SHB's calibrations, but SHB's database cannot be predicted by Pioneer's calibration. The 10 parameters predicted on Pioneer's set are used to estimate the Tilley and Terry digestibility coefficient through a PLS model. The same model is then used to predict T&T on the large SHB set. The later predicted values are reported as references on SHB's set and a global T&T equation is developed after an adequate sample selection. The procedure shows that the new equation, applied on independent sets, is more accurate and more robust than the local one.

Keywords: Maize silage, Tilley and Terry digestibility, Mahalanobis distance, sample selection, transfer, NIR.

Introduction

The robustness of the NIR calibration models depends mainly of the chemical and spectral spread of the selected samples used to determine the models. The database must include all sources of variation, and especially for agricultural products (i.e. grass, grass silage, maize, maize silage etc) covering all these sources of variation (species, years of crop, climatic conditions, locations, sample preparations etc.) can take several years of sample screening and selection. When it is of interest to calibrate a new parameter, it is quite difficult and, very often, even impossible to determine the reference values on all the previous calibration samples. The samples do not exist anymore or they are too much for the lab capacity. Using whole plant maize samples as an example, it is shown how it is possible to create a "global" model to predict the Tilley and Terry (T&T) coefficient of digestibility of the organic matter using a "local" limited set of samples.

Material

The databases which were available have three origins. SHB has developed for 10 years a global database for whole plant maize[1] (fresh maize dried and ground) with 10 parameters: ash, proteins, crude fibre, NDF, ADF, ADL, starch, soluble sugar, organic matter digestibility and NDF digestibility. The total number of different spectra is 3196 and will be called set A (matrices Y_{A1} and X_{A1}). The number of reference values vary from 814 to 2784, according to the available wet chemistry values. In 1994 and 1995, Pioneer Hi-Bred GmbH performed T&T determinations on 244 samples (set B—vector y_{B1} and matrix X_{B1}). The previous company had carried out 106 other T&T determinations in 1996 and this set was used as an independent prediction set (set C). A fourth set of maize samples from a third source was available from RVP (Ir. J. Van Waes, Rijksstation Voor Plantenveredeling) who carried out the T&T determinations as well (set D—crop 1992). The spectra of all the samples concerned were acquired on 6500 or 5000 NIRSystems spectrometers which were standardised to the master instrument at SHB. The global database (set A) contains spectra from 12 standardised instruments whereas sets B and C include only spectra from one standardised instrument. The samples of set D were measured at SHB. The data treatments were carried out by means of the ISI (Infrasoft International) software.[4] Table 1 reports the statistical results of the local equation developed from set B.

The analysis of the PCA scores shows that the global set includes Pioneer's spectra and that the global equations can be applied on this data set to predict the different parameters. The average Mahalanobis standardised distance (*H*) of the 244 Pioneer samples versus the centre of the global data set is 0.78, whereas the average distance of the 3196 samples from the centre of the 244 Pioneer samples is 12.7. Though the *SECV* reported in Table 1 is very good for the prediction of an *in vitro* digestibility coefficient, the spectral space defined by Pioneer's set is too small to rely on the predicted value obtained by this model and then the T&T can not be predicted by the set B equation on a global wide data set such as A.

Transfer of the T&T calibration

The various steps to transfer a T&T equation from set B to set A are the following.

1. With all the reference values available in set A, a suite of 60 MPLS equations are developed for each parameter. The number of factors in each model is fixed by a cross-validation with four groups. There are six pre-treatments: none, SNVD (Standard Normal Variate and Detrend), SNV, Detrend, MSC (Multiplicative Scatter Correction) and Weighted MSC. Associated with these corrections, there are 10 math treatments: log(1/*R*), smoothing of five data points, four first derivatives 1-5-5, 1-10-5, 1-15-5 and 1-20-5 and four second derivatives 2-5-5, 2-10-5, 2-15-5 and 2-20-5. (The first digit gives the degree of the derivation, the second figure is the gap for the subtraction and the third one is the smoothing segment according the calibration program of the ISI software.) The 60 equations are built automatically using the "auto teach sequence" of the ISI package. Table 2 reports the results of the retained equations according to the smallest *SECV* for each parameter.

2. These 10 equations are used to get the 10 predicted parameters on set B (matrix 10 × 244 elements = X_{B2}).

3. The T&T actual values (Y vector) are set with these 10 predicted parameters and a PLS equation (E_{B2}) is then calculated to fit the T&T actual values. The results of the model are presented in Table 3 and the regression coefficients in Figure 1. The main contribution to fit the T&T values is coming from the OMD cellulase method with the highest positive co-

Table 1. Statistical results of Pioneer's Tilley and Terry organic matter digestibility calibration on set B (MPLS) (E_{B1}).

Const.	*N*	Mean	Min	Max	*SD*	R^2C	*SEC*	*SECV*	*T.PLS*
T&T	244	75.2	66	83	3.11	0.85	1.18	1.35	10

N: number of samples; Min: minimum value; Max: maximum value; *SD*: standard deviation; R^2C: coefficient of determination of calibration; *SEC*: standard error of calibration; *SECV*: standard error of cross-validation; *T.PLS*: number of PLS factors.

Table 2. Statistical results of the global calibrations (E_{A1}) for whole plant maize. (MPLS – ISI software) (set A).

Const.	*N*	Mean	Min	Max	*SD*	R^2P	*SECV*	Math treatment
1.ASH	2784	4.6	2.5	11.5	1.09	0.83	0.83	WMSC—2-5-5
2.PRO	1902	8.02	4	13.0	1.20	0.93	0.33	WMSC—1-5-5
3.CF	1846	21.5	9.0	37.5	4.20	0.96	0.81	NSVD—2-10-5
4.NDF	1520	44.4	0.6	59.2	6.20	0.94	1.65	NSV—1-5-5
5.ADF	843	24.8	11.5	40.7	5.20	0.94	1.28	NSV—1-5-5
6.ADL	814	3.1	1.0	7.4	1.02	0.84	0.37	NSVD—2-10-5
7.STA	2294	26.9	0.7	63.8	10.60	0.97	1.72	NSVD—2-5-5
8.SS	1679	6.8	0	20.8	4.90	0.96	1.01	WMSC—2-5-5
9.OMD	2642	72.8	49	87.3	6.10	0.91	1.78	WMSC—2-5-5
10.DNDF	990	82.0	63.5	98.7	6.40	0.79	3.01	NSVD—1-15-5

PRO: protein, CF: crude fibre, STA: starch, SS: soluble sugar, OMD: enzymatic organic matter digestibility, DNDF: enzymatic NDF digestibility.[2]

Table 3. Statistical results of the equation to fit the T&T values from 10 NIR predictors (PLS).

Const.	*N*	Mean	Min	Max	*SD*	R^2C	*SEC*	*SECV*	*T.PLS*
T&T	244	75.2	66	83	3.11	0.76	1.51	1.56	4

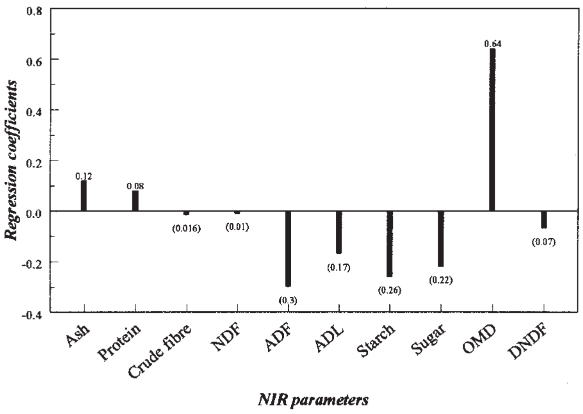

Figure 1. Regression coefficients of the T&T PLS equation using 10 NIR predictors.

efficient (0.64). The single R^2 between OMD and T&T is 0.72. A PCA on the correlation matrix of the 10 predictors gives the next explainable variances for the five first factors: 56.4, 17.1, 11.2, 7.5 and 4.8%. The cumulative variance is 97.0% with five factors. Figure 2 shows the loading plot of the variables in the space of the two first components. The first axis is related to fibre (CF, NDF, ADF, ADL) and in the opposite direction the OMD. The second axis concerns starch and sugar and the third perpendicularly represents the proteins. The fourth is mainly characterised by DNDF and the fifth by the ash content.

We can see that the 10 predictors express less T&T variability than the spectra themselves. R^2 is 0.76 instead of 0.85 in the spectral model (Table 1).

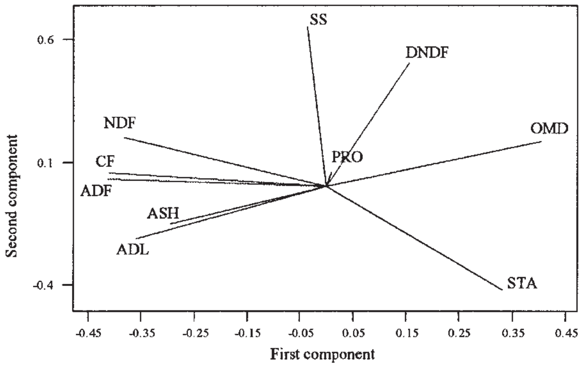

Figure 2. Loading plot of the 10 NIR predictors for the two first principal components (73.5% of the total variance) (set X_{B2}).

The full spectra contain more information than the 10 NIR predictors to estimate T&T.

4. The equations (E_{A1}) of the 10 parameters (Table 2) are applied on the calibration set (X_{A1}) to obtain a full matrix 10 × 3196 (X_{A2}). This was necessary because the database did not contain all reference chemistry values for the 10 parameters.

5. The equation established at point 3 (E_{B2}) is applied on the previous X matrix (X_{A2}) coming from step 4. In this way, a vector of 3196 T&T values was obtained (y_{A2}).

6. The 3196 T&T values of the vector (y_{A2}) are merged with the corresponding spectra.

7. The distribution of the 3196 predicted T&T values are displayed in Figure 3. The distribution tends to be a normal distribution with most of the samples close to the average. This kind of distribution is not the optimum one for calibration and therefore a double selection is performed. The first one is to select on the Y values by keeping simply the extreme points. 513 samples with a T&T value less than 69 and 369 with values higher than 79 are selected. From the remainder, 200 samples are selected on their spectral variation (select routine of ISI) which is not yet included with the 882 first ones. The neighbourhood minimum distance was 0.47 for the latter selection.

8. New equations are developed from the 1082 selected samples to predict T&T directly from the spectra (X_{A1}). After testing 60 models as described above, the selected data pre-treatment was a SNVD and no derivative, but with a smoothing of five data points. The equations with 1 to 16 terms are saved to

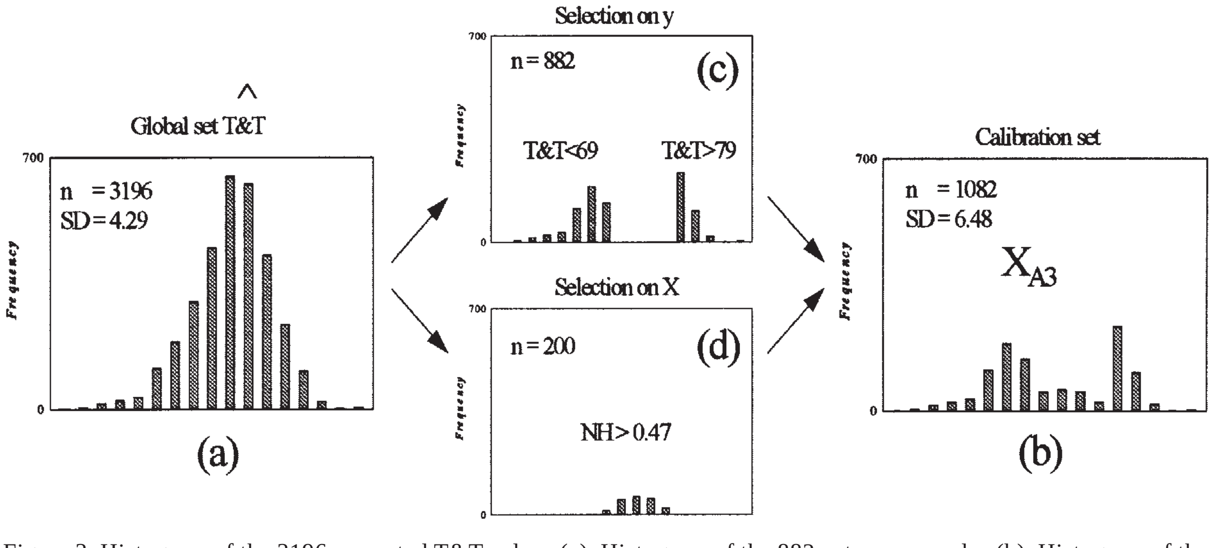

Figure 3. Histogram of the 3196 corrected T&T values (a). Histogram of the 882 extreme samples (b). Histogram of the 200 samples selected on spectral variation (c). Histogram of the final distribution used to calibrate (d).

Table 4. Statistical results of the global equation to predict T&T. NSVD 0-0-5 (no derivative, five point smoothing).

T.PLS	*N*	*SEC*	R^2C	F test	*SECV*	R^2V
1	1082	4.63	0.36	611	4.63	0.36
2	1082	3.59	0.62	710	3.58	0.62
3	1082	3.07	0.72	391	3.07	0.72
4	1082	2.48	0.82	561	2.48	0.82
5	1082	2.38	0.83	89	2.38	0.83
6	1082	2.16	0.86	228	2.16	0.86
7	1082	2.07	0.87	94	2.07	0.87
8	1082	1.71	0.91	498	1.74	0.91
9	1082	1.55	0.93	213	1.58	0.93
10	1082	1.37	0.94	303	1.40	0.94
11	1082	1.23	0.95	224	1.28	0.95
12	1082	1.15	0.96	179	1.16	0.96
13	1082	1.02	0.97	265	1.06	0.97
14	1082	0.94	0.97	235	0.98	0.97
15	1082	0.89	0.98	121	0.90	0.98
16	1082	0.78	0.98	271	0.81	0.98

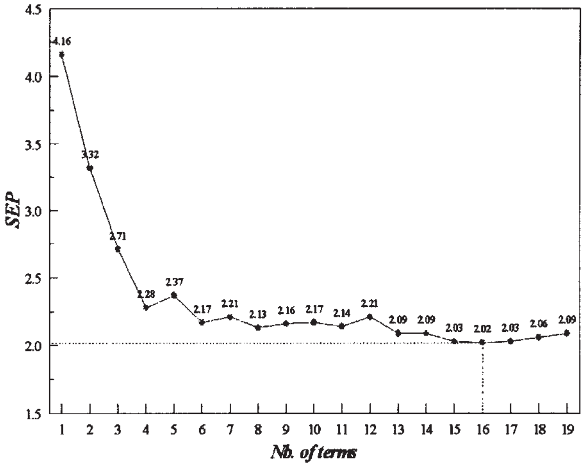

Figure 4. *SEP* on sets C and D in function of the number of PLS terms.

evaluate the performances on the two independent sets. The calibration statistics (Table 4) seem very good, but this is quite normal where a combination of NIR predicted values is used as reference value. In this case, the goal is not to minimise the *SEC* or *SECV*, but to minimise the real *SEP* on the independent sets.

9. These equations are evaluated with totally independent sets and compared with the equation obtained from set B (E_{B1}) with 244 samples. Figure 4 gives the *SEP* for the merged test set (106 + 155) as a function of the number of terms in the T&T global equations (Table 4).The results with 16 terms was the best, but the difference with the results at eight terms was small (2.13 and 2.02). Anyway, the model with 16 terms is kept and Table 5 reports the statistical prediction results for each independent sets.

These results show obviously that the global calibration can predict T&T on independent sets much more accurately than the local calibration. For set C the improvement is smaller than for set D, because set C contains samples prepared and measured in the same laboratory as for set B. For set D, the improvement is very important, because the laboratories used for calibration and validation are different and the instruments are different as well. The internal *SECV*s have been computed by calibrating sets C and D internally. The internal *SECV*s represent limits which can not be exceeded by using an external data set to predict (*SECV* < *SEP*).

The calibration equations for the 10 parameters (Table 2) are made to take account of all the sources of variation coming from the samples (years, varieties, crop conditions and locations, sample preparation etc.) and also from instruments. The MPLS models with scatter correction and derivatives tend to eliminate the unwanted variation to extract the chemical information. Then, the small equation (E_{B1}) based on these NIR predicted values is insensitive to sample preparation and instruments and considers only the chemical variation inside the sample set. The danger of extrapolation and overfitting is reduced by using a simpler model to estimate the T&T digestibility from 10 predictors. The T&T predicted

Table 5. Comparison of the performances between global and local calibration.

Test sets	Set C, n = 106			Set D, n = 155		
Equations	R^2P	*SEP*	Have	*SEP*	R^2P	Have
Local equation (E_{B1}) n = 244 (Table 1) 10 factors	0.59	2.31	3.3	0.34	5.56	9.6
Global equation (E_{A3}) n = 1082 (Table 4) 16 factors	0.74	1.96	0.6	0.78	2.06	1.6
Internal calibrations on the respective independant sets	R^2CV	*SECV*		R^2CV	*SECV*	
	0.78	1.32		0.84	1.67	

values are then used as reference values to create a global equation from the full spectra.

Conclusion

This type of calibration has a certain risk and we believe that it is better to calibrate with actual reference values, when possible, instead of NIR predicted values. The success of this experiment comes from the availability of a large set of well-known parameters able to express the digestibility variation. But anyway, the results on the test sets show that the accuracy is better (lower *SEP*) when using the robust "transferred" calibration rather than a local calibration. This global artificial model is much more accurate and robust than the local one by using an extended range of the Y values and of the X values.

An ultimate solution would be to merge the spectra with actual values with those with "predicted" values to create a new calibration set. The spectra with actual values give a certain amount of variation and the spectra with "predicted" values give the missing spectral variation to develop robust equations. This has not been done in this article because new independent test sets would have been needed, but it will be performed in the future to apply a very robust equation to obtain Tilley & Terry digestibility values which are valid in any situation.

References

1. P. Dardenne, *NIR news* **7(5),** 8 (1996).
2. R. Agneessens, P. Dardenne, P. Lecomte, P. Dujeux and T. Fourneau, *Annales de Zootechnie* **44(suppl),** 39 (1995).
3. P.C. Williams, in *Near Infrared Technology in the Agricultural and Food Industries*, Ed by Phil Williams and Karl Norris. American Association of Cereal Chemists, Inc., p. 131 (1987).
4. J.S. Shenk and M.O. Westerhaus. NIRS3 Version 4.01. Routine operation, Calibration Development and Network System Management Manual. Marketed by NIRSystems, Inc.

Received: 29 October 1997
Revised: 10 March 1998
Accepted: 18 June 1998