Comparison of linear and non-linear NIR calibration methods using large forage databases.

PAOLO BERZAGHI*[1,2], PETER C. FLINN[3], PIERRE DARDENNE[4], MARTIN LAGERHOLM[5], JOHN S. SHENK[6], MARK O. WESTERHAUS[6], and IAN A. COWE[5].

[1] University of Padova, Agripolis, 35020 Legnaro Italy.
[2] University of Wisconsin, 1925 Linden Dr., 53706 Madison, WI, USA.
[3] Agriculture Victoria, Pastoral and Veterinary Institute, Private Bag 105, Hamilton, Victoria 3300, Australia.
[4] Centre de Recherches Agronomiques de Gembloux – CRAGx, 24, Chaussee de Namur, 5030 Gembloux, Belgium.
[5] Foss Tecator AB, Box 70, SE-263 21 Höganäs, Sweden.
[6] Infrasoft International, 109 Sellers Lane, 16870 Port Matilda, PA, USA.

Introduction

Forages represent about 50% of the diets fed to dairy cow and information about their chemical composition is necessary to correctly balance nutrients in the diet. However, chemical and nutritional composition of forages is highly variable. Major sources of variation include botanical family (e.g. legumes vs. grasses), stage of maturity at harvest, method of conservation (e.g. hay vs. silage) and climatic conditions. Because of these sources of variation, commercial forage testing labs have been using several different NIR calibrations to cover the analysis of all forages. Type and source of the sample is critical for the selection of the appropriate calibration equation and this information is often missing or incorrect. Forage NIR analysis would be simplified by using few or even only one NIR calibration for all of the forages. However, the large source of variation that the calibration data set must include may cause problems of non-linear relationship between spectral and chemical information resulting in lower accuracy of prediction.

Alternatives to multivariate calibration methods that can handle non-linear relationship are artificial neural network (ANN)[1] and local PLS calibrations (LOCAL).[2] Although these methods are not new, they have only recently introduced in practical application and they were not tested with large forage database. The aim of this study was to compare the performances of modified PLS (MPLS) calibration to ANN and LOCAL calibrations for the prediction of a large forage data set.

Materials and Methods

The study used forage samples (n=25,977) from Australia, Europe (Belgium, Germany, Italy and Sweden) and North America (Canada and U.S.A) with chemistry data relative to moisture (DM), crude protein (CP) and neutral detergent fibre (NDF) content. The spectra of the samples were collected with 10 different Foss NIRSystems instruments, which were either standardized or not standardized to one master instrument. The spectra were trimmed to a wavelength range between 1100 and 2498 nm.

Two data sets, one standardized (IVAL) and the other not standardized (SVAL) were used as independent validation sets, but 10% of both sets were omitted from the validation sets and they were use for later expansion of the calibration database. The remaining samples were combined into one database (n=21,696), which was split into 75% calibration (CALBASE) and 25% validation (VALBASE).

Modified PLS equations were developed using WinISI (Infrasoft International LLC, USA). Pre-defined spectra math treatments were first derivative, 4 data points skipping gap and smoothing with SNV-Detrend scatter correction. Local PLS calibrations were also developed under WinISI software. In this case 2 settings were defined. The first was decided prior to the trial (LOCAL1), while the second (LOCAL2) was optimized for the prediction of CALBASE. Also for ANN there were 2 methods (ANN1 and ANN2) both developed under Mathlab (The Mathworks Inc., USA).

The chemical components in the 3 validation data sets were predicted with each model derived from CALBASE using the calibration database before and after it was enhanced with 10% of the samples from IVAL and SVAL data sets. Calibration performances were evaluated using standard error of prediction (SEP), bias, SEP corrected for bias (SEP(C)), slope and R^2.

Results

Regardless of calibration method, prediction of VALBASE (data not shown) had smaller SEP(C) and bias values than for IVAL (Table 1) and SVAL (Table 2). This was not surprising as VALBASE was selected from the calibration database and it had a sample population similar to CALBASE, whereas IVAL and SVAL were completely independent validation sets. Part of the problem may be caused by differences in wet chemistry methods as indicated for example by the large bias of DM in SVAL or NDF in IVAL.

None of the models developed before enhancements appeared to be consistently better for the 2 independent validation sets. However, LOCAL and ANN had lower SEP and SEP(C) than MPLS for all the 3 variables evaluated in VALBASE. This is consistent with previous studies that found LOCAL[3] and ANN [4] being able to handle data sets with large source s of variation.

In most cases, LOCAL and ANN models, but not modified PLS, showed considerable improvement in the prediction of IVAL (Table 1) and SVAL (Table2) after the calibration database had been expanded with the 10% samples of IVAL and SVAL reserved for calibration expansion. The addition of only 439 samples from the 2 independent sets to the 16272 sample of VALBASE greatly reduced bias, SEP and

SEP(C) of LOCAL and ANN of IVAL and SVAL. Under a practical point of view, the expansion of a database to predict new forage products will require fewer samples and result in better accuracy using either LOCAL or ANN than using MPLS calibrations.

The effects of sample processing, instrument standardization and differences in reference procedure were partially confounded in the validation sets, so it was not possible to determine which factors were most important.

Conclusions

Compared to MPLS, Local and ANN improved accuracy of predictions of forage samples similar to those in the calibration data set. The accuracy of prediction of complete independent data sets was unacceptable for all the models, but LOCAL and ANN were able to reduce SEP, BIAS and SEP(C) after updates using a small number of samples. LOCAL and ANN were able to manage large source of variations adding the flexibility of rapid and inexpensive expansion to new forage data sets.

Further work on the development of large databases must address the problems of standardization of instruments, harmonization and standardization of laboratory procedures and even more importantly, the definition of the database population.

References

Table 1: Prediction performances of the different calibration methods for the independent set from Italy (IVAL)

	Enhancement	SEP	SEP(C)	Bias	Slope	R^2
DM (no=1885)						
MPLS	Before	1.341	1.328	0.195	1.184	0.789
LOCAL1	Before	1.542	1.439	0.557	1.16	0.743
LOCAL2	Before	1.533	1.431	0.552	1.135	0.743
ANN1	Before	1.379	1.355	0.256	1.088	0.765
ANN2	Before	1.334	1.315	0.228	1.076	0.778
MPLS	After	1.333	1.322	0.174	1.165	0.787
LOCAL1	After	1.122	1.122	0.022	1.062	0.838
LOCAL2	After	1.074	1.074	0.006	1.055	0.851
ANN1	After	1.342	1.34	0.078	1.198	0.786
ANN2	After	1.317	1.315	0.087	1.177	0.792
CP (no=1846)						
MPLS	Before	1.821	1.329	-1.245	0.881	0.959
LOCAL1	Before	2.123	1.519	-1.484	0.866	0.947
LOCAL2	Before	1.913	1.44	-1.26	0.869	0.954
ANN1	Before	2.149	1.535	-1.504	0.845	0.955
ANN2	Before	2.001	1.408	-1.423	0.865	0.958
MPLS	After	1.739	1.31	-1.144	0.886	0.96
LOCAL1	After	1.259	1.151	-0.512	0.963	0.958
LOCAL2	After	1.143	1.102	-0.303	0.971	0.961
ANN1	After	1.189	1.043	-0.571	0.953	0.967
ANN2	After	1.062	0.987	-0.393	0.967	0.969
NDF (no=1912)						
MPLS	Before	4.619	3.473	3.047	1.019	0.926
LOCAL1	Before	5.527	4.145	3.658	0.988	0.894
LOCAL2	Before	5.356	3.817	3.758	1	0.91
ANN1	Before	4.635	3.556	2.975	1.059	0.925
ANN2	Before	4.983	3.46	3.586	1.048	0.928
MPLS	After	4.195	3.439	2.402	1.035	0.928
LOCAL1	After	3.402	3.198	1.162	1.015	0.937
LOCAL2	After	3.149	2.958	1.082	1.011	0.946
ANN1	After	3.154	3.058	0.774	1.041	0.944
ANN2	After	3.01	2.926	0.707	1.046	0.949

Table 2: Prediction performances of the different calibration methods for the independent set from Sweden (SVAL)

	Enhancement	SEP	SEP(C)	Bias	Slope	R^2
DM (no=1861)						
MPLS	Before	3.853	2.414	-3.003	-0.26	0.117
LOCAL1	Before	3.08	2.456	-1.86	-0.151	0.048
LOCAL2	Before	3.202	2.433	-2.083	-0.113	0.028
ANN1	Before	3.3	2.405	-2.26	-0.195	0.071
ANN2	Before	3.515	2.487	-2.484	-0.232	0.105
MPLS	After	2.868	2.353	-1.641	-0.277	0.12
LOCAL1	After	0.899	0.896	-0.079	0.808	0.532
LOCAL2	After	0.895	0.891	-0.092	0.795	0.544
ANN1	After	0.819	0.816	-0.072	0.906	0.593
ANN2	After	0.658	0.656	-0.055	0.976	0.733
CP (no=1860)						
MPLS	Before	1.009	0.738	0.688	0.967	0.974
LOCAL1	Before	1.342	1.06	0.824	0.982	0.944
LOCAL2	Before	1.555	1.318	0.826	0.957	0.915
ANN1	Before	1.207	0.698	0.985	0.944	0.979
ANN2	Before	1.268	0.719	1.044	0.959	0.976
MPLS	After	0.852	0.738	0.426	0.969	0.974
LOCAL1	After	0.739	0.74	0.002	0.997	0.973
LOCAL2	After	0.72	0.72	-0.011	0.993	0.974
ANN1	After	0.705	0.694	0.124	0.977	0.977
ANN2	After	0.674	0.668	0.094	0.973	0.978
NDF (no=1660)						
MPLS	Before	2.596	2.387	-1.022	1.057	0.924
LOCAL1	Before	4.462	3.982	-2.015	1.025	0.78
LOCAL2	Before	3.631	3.602	-0.465	1.057	0.822
ANN1	Before	2.897	2.482	-1.494	1.063	0.918
ANN2	Before	2.53	2.525	-0.159	1.069	0.915
MPLS	After	2.268	2.267	0.08	1.036	0.93
LOCAL1	After	2.236	2.23	-0.182	1.034	0.932
LOCAL2	After	2.17	2.164	-0.169	1.041	0.936
ANN1	After	2.199	2.196	-0.129	1.027	0.934
ANN2	After	2.049	2.042	-0.179	1.025	0.943

1. T. Næs,K. Kvaal,T. Isaksson and C. Miller, *J. Near Infrared Spectrosc.* **1**, 1 (1993).
2. J. S. Shenk,P. Berzaghi and M. O. Westerhaus, *J. Near Infrared Spectrosc.* **5**, 223 (1997).
3. P. Berzaghi,J. S. Shenk and M. O. Westerhaus, *J. Near Infrared Spectrosc.* **8**, 1 (2000).
4. N. B. Buchmann, in *Near Infrared Spectroscopy: The Future Waves*, NIR Publications, Chichester, p. 479 (1996).