

Advantages and disadvantages of multiple linear regression and partial least squares regression equations for the prediction of fatty acids

J. García-Olmo, A. Garrido-Varo and E. De Pedro

Dpto. Producción Animal. ETSIAM. University of Córdoba, PO Box 3048, E-14080, Córdoba, Spain.

Introduction

There are several methods for the multivariate calibration of near infrared (NIR) spectroscopic data. During the early stages of NIR development, multiple linear regression (MLR) models were used most. Later, the developments in instrumentation and chemometric methods led to the increase in the use of the so-called full regression models like principal component regression (PCR) and partial least squares (PLS) regression. Several authors compare results obtained by MLR and PLS regression in different products.¹⁻⁶ The majority of previous papers concluded that PLS and MLR, essentially, give the same predictions errors. Where one method prevails, it is usually by small differences in the coefficient of determination (r^2) and *SEP*.

Based on our experience working on the development of calibration equations for the prediction of fatty acids in Iberian pig fat,^{7,8} optimum predictions could be obtained with PLS models. In order to develop models that could be used with several instruments distributed in many different Iberian pigs industries and Regulatory Laboratories, it could be of interest to make models as simple as possible for two main reasons. The first is that less sophisticated chemometric models will be easier to transfer from one instrument to another. The second is that it could be possible to develop less sophisticated and cheaper instrumentation, once it was proved that the precision obtained with a simple model (using a few wavelengths) is satisfactory for industry requirements.

Material and methods

Samples

A total of 372 samples of Iberian pig fat, belonging to animals reared during 1997 and 1998 and produced under the guidelines of the Designation of Origin “Jamón de Huelva”, were used. A total of 352 samples were used as the calibration set (Table 1) and 20 samples of Iberian pig fat reared during 1998 (Table 2) were used for validation of the chemometric models.

Reference analysis

The determination of the fatty acid content of fat samples was performed by gas chromatography (GC).⁶ The methyl esters of fatty acids were extracted with hexane and were determined using a Perkin Elmer Sigma 3D chromatograph with FID detector. Calibration equations were only obtained for the

Table 1. Calibration set ($n = 352$).

	Mean	SD	Range
%C16 : 0	21.1	1.5	17.4–25.3
%C18 : 0	10.7	1.4	7.7–14.9
%C18 : 1	52.3	2.5	45.0–58.1
%C18 : 2	9.4	1.3	6.8–13.5

Table 2. Validation set ($n = 20$).

	Mean	SD	Range
%C16 : 0	20.1	1.6	17.4–23.4
%C18 : 0	9.5	1.3	7.6–12.4
%C18 : 1	53.6	2.7	47.4–58.1
%C18 : 2	10.1	1.2	8.6–12.4

main fatty acids of Iberian pig fat: % palmitic acid (% C16:0), % stearic acid (% C18:0), % oleic acid (% C18:1) and % linoleic acid (% C18:2).

NIR analysis

NIR data were recorded from 400 to 2500 nm using a Foss-NIRSystem 6500 spectrophotometer (Foss NIRSystems, Silver Springs, MD, USA) equipped with a spinning module. Samples were analysed by folded transmission using a sample cell of 0.1 mm pathlength (ref. IH- 0345). Spectra were collected with ISI NIRS 3 software ver. 3.11 (Infrasoft International, Port Matilda, PA, USA).

This software was used for the chemometric analysis of NIR data, with methodology for development and validation of NIR calibrations described by Mark and Workman⁹ and Shenk and Westerhaus.^{10,11} Statistics used to evaluate calibration equations were: standard error of calibration (*SEC*), standard error of prediction (*SEP*) and coefficient of determination (R^2).

Two regression methods were used: MLR and Modified PLS. The MLR method chose the maximum number of wavelengths selected for each variable by a stepwise process. It was performed by selecting the wavelength that was most highly correlated with the reference values (measured by *F* statistic value). Once a wavelength was added to the equation, it was evaluated against all the wavelengths of the spectra. The process stopped when the added wavelength showed little or no improvement in the model.

The PLS method performs a principal component analysis decomposition in such a way that reference data is used for an optimal decomposition of NIR data and then performs regression equations. Modified PLS (MPLS) method standardises residuals values before calculate next regression term.

All the equations were obtained using the standard normal variate and detrending (*SNVD*) method¹² and different derivative math treatments.

Results

Table 3 summarises statistical values for the best calibration equations obtained with the MPLS regression method.⁸ The coefficients of determination (R^2) for all the fatty acids (% C16:0, % C18:0, % C18:1 and % C18:2) were excellent, reaching values close to 1 in all cases (Table 3). The *SEC* values for all the fatty acids were very low, ranging from a minimum of 0.15 (% C18:2) to a maximum of 0.24 (% C16:0).

MLR equations were also obtained by using the same calibration set. An important parameter for the development of the MLR model is to set the number of wavelengths per variable that should use the model. MLR equations with 10 wavelengths selected per variable had similar R^2 and *SEC* values (Table 4) as MPLS calibration equations (Table 3). Table 5 shows the 40 wavelengths selected for the four MLR calibration models. MLR equations with two wavelengths selected per variable were also obtained and they had lower precision (Table 6) than MPLS models (Table 3). However, all the coefficients of determination were excellent ($R^2 > 0.91$) and *SEC* values were ranging from a minimum of

Table 3. Statistics of MPLS equations.

	<i>N</i>	<i>SEC</i>	R^2
%C16 : 0	321	0.24	0.97
%C18 : 0	320	0.19	0.98
%C18 : 1	311	0.22	0.99
%C18 : 2	331	0.15	0.99

Table 4. Statistics of MLR equations (10 wavelengths per variable).

	<i>N</i>	<i>SEC</i>	R^2
%C16:0	310	0.23	0.98
%C18:0	319	0.27	0.96
%C18:1	308	0.26	0.99
%C18:2	324	0.16	0.99

Table 5. Selected wavelength for MLR equations.

	Selected wavelengths (nm)									
	1 st	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th	10 th
%C16 : 0	2426	2132	1784	1516	2220	2190	2244	1744	1654	2458
%C18 : 0	1230	1734	1680	2250	1646	2370	2436	2286	1756	2142
%C18 : 1	2412	1568	2422	1176	2204	1008	2450	1750	2222	1708
%C18 : 2	1150	2226	1794	2120	2176	2040	2284	1668	2456	1872

Table 6. Statistics of MLR equations (2 wavelengths per variable).

	<i>N</i>	<i>SEC</i>	<i>R</i> ²
%C16:0	327	0.42	0.92
%C18:0	328	0.39	0.91
%C18:1	328	0.48	0.96
%C18:2	321	0.28	0.96

Table 7. Selected Wavelength for MLR equations.

	Selected wavelength (nm)	
	1 st	2 nd
%C16:0	1654	2458
%C18:0	1746	2456
%C18:1	1708	2450
%C18:2	2190	2450

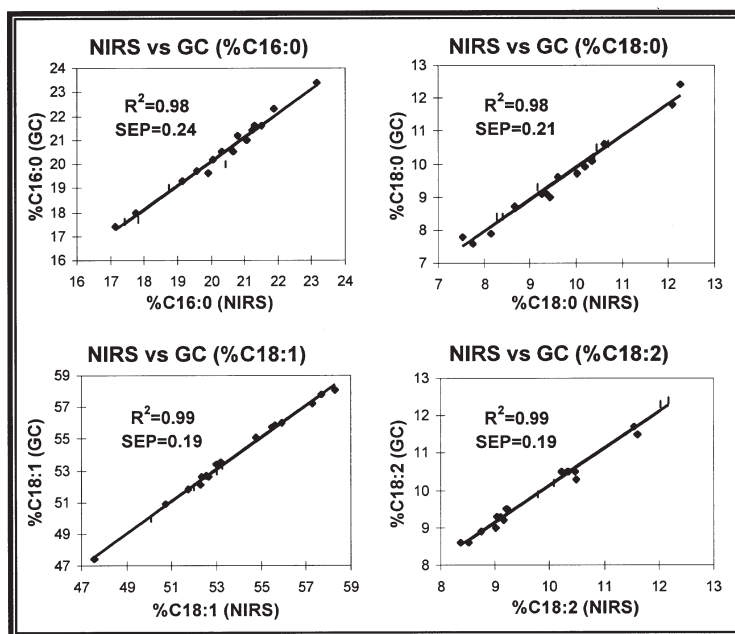


Figure 1. Validation results with MPLS equations.

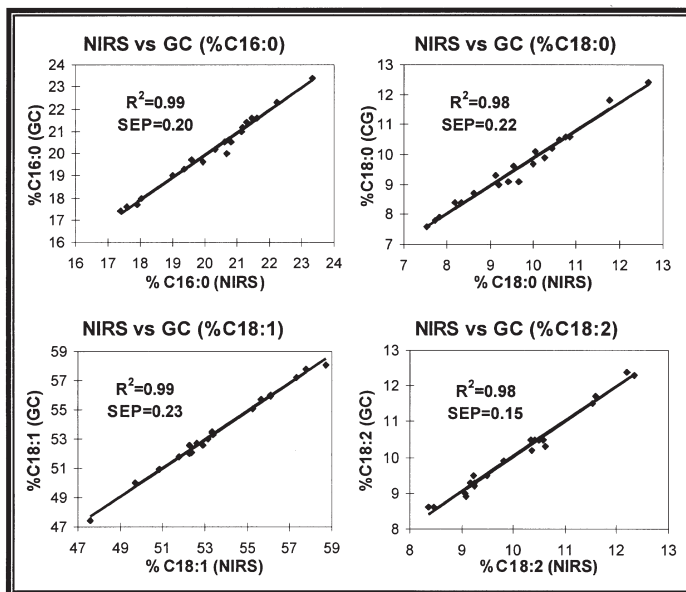


Figure 2. Validation results with MLR equations (10 wavelengths per variable).

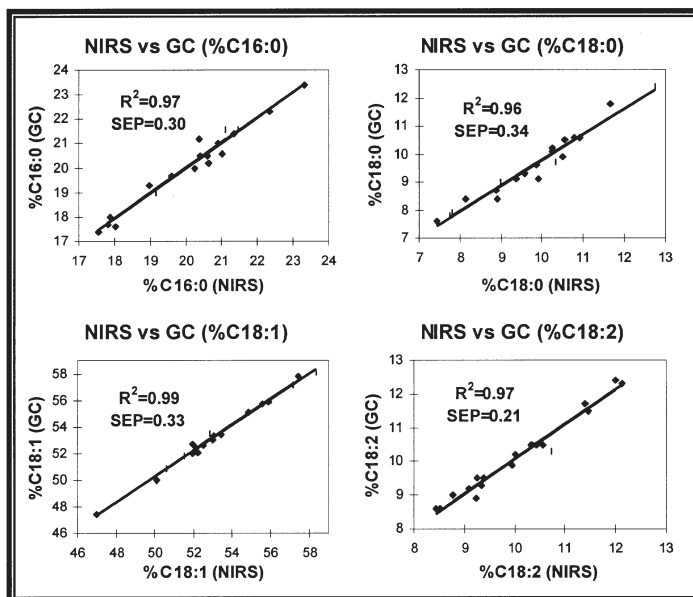


Figure 3: Validation results with MLR equations (2 wavelengths per variable).

0.28 (% C18:2) to a maximum of 0.48 (% C18:1). Table 7 gives the 8 wavelengths selected for the four MLR calibration models.

Once the calibration equations with the different regression models were obtained, they were applied to an external validation set of 20 samples not included in the calibration set and belonging to animals reared during 1998 (Table 2). As can be seen in Table 2, the fatty acid composition of samples of the external validation set was similar to the composition of samples of the calibration set (Table 1). Figures 1 to 3 show statistics (coefficient of determination and *SEP* values) and NIR predicted values v. GC values of each fatty acid with MPLS and MLR equations.

The comparison of validation statistics (R^2 and *SEP*) between MPLS and MLR equations (with ten wavelengths selected per variable) confirmed that both models had a similar precision. MLR equations (with two wavelengths selected per variable) had *SEPs* values slightly higher than those obtained with MPLS equations.

Conclusions

For the development of calibration equations for the prediction of fatty acids of Iberian pig fat, spectral and reference data were used. MLR models with ten wavelengths per variable predict fatty acid content of Iberian pig fat with the same precision as an MPLS model. An MLR model with two wavelengths per variable is precise enough for the quality control of Iberian pig industries.

Acknowledgements

NIR data were obtained using the NIR hardware and software of the Centralised NIR Unit of the University of Cordoba. GC data were obtained at Laboratorio Agrario Córdoba (Junta de Andalucía). Special thanks to Antonio López and Alberto Sánchez de Puerta for their help in sample preparation.

References

1. W.R. Windham and P.C. Flinn, in *Near Infrared Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications*, Ed by K.I. Hildrum, T. Isaksson, T. Næs and A. Tandberg. Ellis Horwood, Chichester, UK, p. 459 (1992).
2. D. Jouan-Rimbaud and D.L. Massart, in *Near Infrared Spectroscopy: The Future Waves*, Ed by A.M.C. Davies and P. Williams. NIR Publications, Chichester, UK, p. 194 (1996).
3. B. Rosental and P. Williams, in *Near Infrared Spectroscopy: The Future Waves*, Ed by A.M.C. Davies and P. Williams. NIR Publications, Chichester, UK, p. 1 (1996).
4. M. Blanco, J. Coello, H. Iturriaga, S. Maspocho and R. Gonzalez, *Mikrochim. Acta* **128**, 235 (1998).
5. J. Lu, W.F. McClure, F.E. Barton, D.S. Himmelsbach, *J. Near Infrared Spectrosc.* **6**, 77 (1998).
6. Y.B.C. Man and M.H. Moh, *J. Amer. Oil Chem. Soc.* **75**(5), 557 (1998).
7. E. De Pedro, A. Garrido, I. Bares, M. Casillas and I. Murray, in *Near Infrared Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications*, Ed by K.I. Hildrum, T. Isaksson, T. Næs and A. Tandberg. Ellis Horwood, Chichester, UK, pp. 345–348 (1992).
8. J. García-Olmo. *I Accésit del VIII Premio FEDNA para Jóvenes Investigadores, convocado por la Fundación Española para el Desarrollo de la Nutrición Animal* (1998).
9. H. Mark and J. Workman, *Statistics in Spectroscopy*. Academic Press, NY, USA (1991).
10. J.S. Shenk and M.O. Westerhaus, *Routine operation, calibration, development and network system management manual*. NIRSystems, Silver Spring, MD, USA (1995).
11. J.S. Shenk and M.O. Westerhaus, in *Near Infrared Spectroscopy: The Future Waves*, Ed by A.M.C. Davies and P.C. Williams. NIR Publications, Chichester, UK, p. 198 (1996).
12. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc.* **43**, 772 (1989).