Understanding the "*H*" statistic during routine analysis of animal fats

Juan García-Olmo,[a,b] Ana Garrido-Varo[a] and Emiliano De Pedro[a]

[a]*Department of Animal Production, Faculty of Agriculture and Forestry Engineering, University of Córdoba, PO Box 3048, E-14080 Córdoba, Spain*

[b]*NIR/MIR Unit, SCAI, University of Cordoba, Spain*

Introduction

To ensure the correct usage of a predictive analytical technique such as near infrared (NIR) reflectance spectroscopy, outlier detection should be included as an integral part of instrument operation.[1]

In routine analysis, using NIR reflectance spectroscopy, detection of erroneous or abnormal samples should be based on the spectral information obtained. The calculation of statistics such as the Mahalanobis distance,[2] levarage[3] or the "*H*" statistic,[4] is of great value to detect sample spectra that differ widely from spectra in the calibration set, thus avoiding an extrapolation of the calibration model. In routine analysis, ISI software is especially suitable for this purpose, since it provides both predicted values and the "*H*" statistic of a sample immediately after scanning its NIR spectrum.[4]

However, correct use and interpretation of *"H"* values is crucial, because samples with higher-than-recommended *"H"* values should be sent to the laboratory for reference analysis. Given that reference analysis is usually expensive, it is important to differentiate between real and false *"H"* values in samples that appear to be outliers.

Previous studies have shown that NIR calibration equations, with a precision similar to that of the reference method,[5–7] can be obtained for predicting fatty acids in liquid samples of pig fat. However, over a four-year period spent in obtaining equations, certain uncontrolled variations have been observed in the "*H*" values for predicted samples that hinder the adoption of an outlier detection strategy.

The purpose of this study was to achieve an enhanced understanding of the "*H*" statistic during routine analysis of samples of liquid animal fat.

Material and methods

Samples

Three different validation sample sets were used (see Table 2). Validation set A consisted of the spectra of 20 Iberian pig fat samples obtained in May 1998 not included in the calibration file. Validation set B was the spectra of 30 Iberian pig fat samples obtained in November 1999 included in the calibration file. Validation set C consisted of 150 spectra of one pig fat sample (designated as the fat check sample), which was representative of the mean fatty acid composition values in the calibration file. This check sample had been analysed three times a week from June 1999 to July 2000.

Near Infrared Spectroscopy: Proceedings of the 10th International Conference

NIR analyses and reference data

Validation sets were analysed by gas chromatography (GC) and NIR in the same way as the calibration set.[7] The percentage by weight of the main fatty acids in liquid fat samples (palmitic acid or C16 : 0, stearic acid or C18 : 0, oleic acid or C18 : 1 and linoleic acid or C18 : 2) was determined by gas chromatography. NIR data were recorded from 400 to 2500 nm using a Foss NIRSystems 6500 scanning monochromator equipped with a spinning module. Samples were analysed by folded transmission using a ring cup with a pathlength of 0.1 mm (ref. IH-03459). Only the NIR range (from 1100 to 2500 nm) was used in the analysis. Spectra were collected and processed by the ISI Ver. 3.11 software (Infrasoft International, Port Matilda, PA, USA).

Monitoring NIR equations

The three validation sets were used to monitor the equations developed earlier and reported in a previous work.[7] Shenk *et al*.[8] have outlined a procedure for monitoring predictions obtained by NIR equations. A monitoring test should be made whenever there is any reason to believe that an equation may not be predicting adequately. This procedure, when applied to a validation sample set ($n > 9$), established limits for bias ($0.6 \times SEC$), for unexplained error or $SEP(C)$ ($1.3 \times SEC$) and for the "*H*" statistic ("*H*"< 3). When bias, $SEP(C)$ or "*H*" statistics of a validation set exceed confidence limits, samples should be sent to a laboratory for reference analysis.

Results and discussion

Table 1 shows fatty acid composition for the three validation sets. Mean values for composition of validation sets A and B and the composition of the sample represented in set C were similar to the mean fatty acid values of the calibration set (Table 2).

Table 1. Weight percentage of fatty acids for validation sets A, B and C.

Constituent	Set A ($n = 20$)			Set B ($n = 30$)			Set C ($n = 1$)
	Mean	*SD*	Range	Mean	*SD*	Range	
C16 : 0	20.13	1.63	17.40–23.40	21.08	1.98	17.90–24.60	20.40
C18 : 0	9.54	1.28	7.60–12.40	10.44	1.66	7.70–14.90	10.00
C18 : 1	53.56	2.70	47.40–58.10	52.40	3.11	46.30–57.50	52.60
C18 : 2	10.11	1.15	8.60–12.40	9.61	1.69	6.90–13.50	10.40

Table 2. Statistics for NIR equations predicting weight percentage of fatty acids in Iberian pig fat.

Constituent	Mean	*SD*	Range	*SECV*	r^2	*SEL*[a]
C16 : 0	21.00	1.39	16.83–25.18	0.26	0.97	0.26
C18 : 0	10.62	1.31	6.68–14.56	0.24	0.97	0.22
C18 : 1	52.24	2.37	45.13–59.36	0.26	0.99	0.25
C18 : 2	9.39	1.30	5.48–13.30	0.15	0.99	0.15

[a]Standard laboratory error calculated from 20 samples analysed in duplicate using the reference method (GC)

Table 3. *SEP*, bias and *SEP*(*C*) values for validation sets A, B and C.

	Set A			Set B			Set C		
	SEP	Bias	*SEP*(*C*)	*SEP*	Bias	*SEP*(*C*)	*SEP*	Bias	*SEP*(*C*)
C16 : 0	0.22	–0.03	0.22	0.29	–0.13	0.26	0.21	–0.01	0.21
C18 : 0	0.21	–0.09	0.20	0.28	0.08	0.27[a]	0.38	0.25[a]	0.29[a]
C18 : 1	0.37	0.29[a]	0.24	0.43	–0.33[a]	0.27	0.25	–0.05	0.24
C18 : 2	0.14	0.03	0.14	0.19	0.12[a]	0.16	0.24	–0.02	0.24[a]

Bias limits: C16 : 0 = 0.14, C18 : 0 = 0.12, C18 : 1 = 0.15 and C18 : 2 = 0.08
SEP(*C*) limits: C16 : 0 = 0.30, C18 : 0 = 0.26, C18 : 1 = 0.31 and C18 : 2 = 0.17
[a]Bias and *SEP*(*C*) values higher than confidence limits

It should be stressed that all validation sets were scanned after taking the spectra of the calibration set. The calibration set was obtained between January 1997 and April 1998, validation set A was collected just after the calibration set (in May 1998) and validation set B was scanned 19 months later (in November 1999). Validation set C was obtained over a period of 13 months (from June 1999 to July 2000), one year after scanning the calibration samples.

Table 3 shows *SEP*, bias and *SEP(C)* values obtained for each fatty acid during routine analysis of the validation sets. As can be seen from this table, *SEP* values were very low and similar to the *SECV* values. In addition, with only a few exceptions, bias and *SEP*(*C*) values for the three data sets were lower than the confidence limits established for each fatty acid. These data confirm the robustness and the high precision of the NIR equations obtained previously.[7]

However, the mean "*H*" values for each validation set differed. This statistic measures the mean distance of validation set spectra to the centroid of the principal component space defined by calibration set spectra. Thus, the "*H*" statistic provides information about the distance between the validation set spectra predicted by calibration equations and the spectra belonging to the calibration set. As Table 4 shows, the "*H*" value for validation set A ("*H*" = 2.63) was below three. Thus, spectra from validation set A can be considered similar to calibration set spectra.

However, the mean "*H*" value for validation set B ("*H*" = 25.44) and Set C ("*H*" = 33.38) were much higher than the maximum value recommended.[8] This means that spectra for validation sets B and C must be considered as outliers, despite their excellent *SEP* values. These anomalous "*H*" statis-

Table 4. Mean "*H*" statistic values for validation sets A, B and C using different principal component spaces.

	"*H*" calculated using different spectra sets	
	Calibration set	Calibration set file + validation set C
Set A	2.63	2.19
Set B	25.44	2.10
Set C	33.38	1.17

tics for validation sets B and C are difficult to explain, since validation set B contained replicate spectra for the same samples scanned during calibration development and Set C had spectra of a single sample, which was similar to the mean spectrum of the calibration set. Validation set C, the check spectra of a sample analysed three times weekly over several months, should reflect the influence of day-by-day variations of the instrument and/or the environment in a fat sample representative of the calibration set.

In order to model these day-by-day instrumental and/or environmental variations, the principal component space used to calculate "*H*" values was reconstructed using not only the calibration set but also validation set C. "*H*" statistics were again calculated by projecting validation sets onto the new principal component space. Mean "*H*" statistic values for validation sets are shown in Table 4. The mean "*H*" value for validation set A was similar using both principal component spaces (2.63 vs. 2.19). However, mean "*H*" statistic values for validation set B and C decreased with the new principal component space (25.44 *vs* 2.10 for validation set B and 33.38 vs 1.17 for validation set C).

When a principal component space taking into spectra of both the calibration set and validation set C is constructed, mean "*H*" statistic values for all validation sets were lower than three. Thus, there is no reason to consider the validation sets as outliers.

Conclusions

Results show that the *"H"* statistic is very useful for detecting outlier fat samples by using NIR spectral information projected onto the principal component space. To construct this multivariate space, spectra of a fat sample that models day-by-day instrumental and/or environmental variations must be included.

Acknowledgments

NIR data were obtained using NIR hardware and software at the NIR/MIR Spectroscopy Unit (SCAI) of the University of Cordoba. GC data were obtained at the Laboratorio Agrario de Córdoba (Junta de Andalucía). Special thanks to Ms Paquita Baena, Mr Antonio López and Mr Alberto Sánchez de Puerta for laboratory assistance.

References

1. H. Martens and T. Næs, *Multivariate Calibration*. John Wiley & Sons, Chichester, UK (1991).
2. R. De Maesschalck, D. Jouan-Rimbaud and D.L. Massart, *Chemom. Intell. Lab. Syst.* **50,** 1 (2000).
3. *Unscrambler User´s guide*, version 5.0. Programme package for multivariate calibration. CAMO A/S, Trondheim, Norway (1993).
4. J.S. Shenk and M.O. Westerhaus, *Routine operation, calibration, development and network system management manual*. NIRSystems Inc., 12101 Tech Road, Silver Spring, MD 20904, USA, (1995).
5. 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, p. 341 (1992).
6. J. Garcia-Olmo, A. Garrido and E. De Pedro, in *Near Infrared Spectroscopy: Proceedings of the 9th International Conference*, Ed by A.M.C. Davies and R. Giangiacomo. NIR Publications, Chichester, UK, p. 253 (2000).
7. J. Garcia-Olmo, A. Garrido and E. De Pedro, *J. Near Infrared Spectrosc*. **9,** 49 (2001).
8. J.S. Shenk, J.J. Workman and M.O. Westerhaus, in *Handbook of near infrared analysis, Ed by* P.A. Burns and E.W. Ciurczak. Marcel Dekker, NY, USA, p. 383 (1992).