Chemometrics and Intelligent Laboratory Systems xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A case study of extrapolation in NIR modelling — A chemometric challenge at 'Chimiométrie 2009'

Juan Antonio Fernández Pierna [a], Hélène Duval [b], Patricia Valderrama [c], Douglas N. Rutledge [c], Vincent Baeten [a], Pierre Dardenne [a,*]

[a] Food and Feed Quality Unit, Valorisation of Agricultural Products Department, Walloon Agricultural Research Centre (CRA-W), Henseval Building, Chaussée de Namur no. 24, 5030 Gembloux, Belgium
[b] Rue d'Inkermann, 59000 Lille, France
[c] AgroParisTech/INRA, 16 rue Claude Bernard, 75005 Paris, France

ARTICLE INFO

Article history:
Received 23 February 2010
Received in revised form 25 March 2010
Accepted 1 April 2010
Available online xxxx

Keywords:
NIR
Challenge
Variable selection
Rapeseed

ABSTRACT

Since 2004, a challenge is proposed at the "Chimiométrie" conference organized by the GFC ("Groupe Français de Chimiométrie"). The annual congress was held in Paris 30 November and 1 December 2009. The data are still available on the GFC website (www.chimiometrie.org/). The data for this session were extracted from the CRA-W (Walloon Agricultural Research Centre, www.cra.wallonie.be/) spectral data base. The calibration set and the validation set (the reference values for the latter were not available to the participants) were set up to have an obvious case of extrapolation. The data sets were from whole rapeseed samples and the parameter to be calibrated was the total glucosinolate content. Seven participants reported results and, according to the highest R^2 of prediction, three were asked to present their approach during the conference. These approaches and the ones of the challenge organizers are presented in this paper.

1. Introduction

For the fifth consecutive year and due to the success of the chemometric contests organized within the framework of previous congresses [1–4], another data set has been proposed at the 'Chimiométrie 2009' conference (http://www.chimiometrie.org/) held in Paris, France (30/11–01/12/2009). This data set was selected in order to test the ability of the participants to apply regression methods to NIR data in the case of extrapolation, i.e. to data sets which contain *Y*-variable values outside the range of the calibration set. The aim of the Challenge 2009 was to develop the most robust and precise calibration model. A blind validation set, presented without the reference values to the participants, had the characteristic that all the reference values were higher than the reference values of the calibration set indicating a clear case of extrapolation.

Seven participants decided to investigate the proposed data. The evaluation of the results was made on the basis of the best validation according to the R^2 obtained on the predicted values for the validation set. The three best approaches were presented during the conference and are summarized here in this paper together with the challenge organizer's approaches.

2. Material

From the CRA-W historical data set of whole rapeseed spectra (1300 spectra) obtained on several Foss NIRsystem 5000 instruments (spectra were collected in the 1100–2498 nm range with a spectral resolution of 2 nm), a subset of 288 spectra was randomly selected. The analyzed parameter is the content of total glucosinolate [5,6], which is a group of organic compounds containing sulphur and nitrogen and which are derived from glucose and an amino acid. They are present in almost all plants of the Brassicales order. The units are micromoles per gramme (μmol/g) on a Dry Matter basis. Fig. 1 reports the histogram of the glucosinolate content for both the calibration and the validation sets. The calibration set contained 118 samples (in black in the figure) and had a mean of 15.1 μmol/g and a standard deviation of 6.04. The validation set was composed of 170 samples (in grey in the figure) and had a mean of 58.6 μmol/g with a standard deviation of 22.70.

Fake visible (400–780 nm) and NIR regions (780–1098 nm) were added to the spectra. These fake regions were estimated by applying a PLS2 model constructed using another available set of rapeseed samples scanned on a monochromator instrument from 400 to 2498 nm. Then, the PLS2 model was applied to the 288 selected samples to estimate that visible region and part of the NIR region (400–1098) from the rest of the NIR region (1100–2498 nm). Therefore, this new region does not carry any additional information because the absorbances are just linear combinations of the near infrared region. The 25 first and 25 last data

* Corresponding author. Tel.: +32 0 81620354; fax: +32 0 81620388.
E-mail address: dardenne@cra.wallonie.be (P. Dardenne).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.04.001

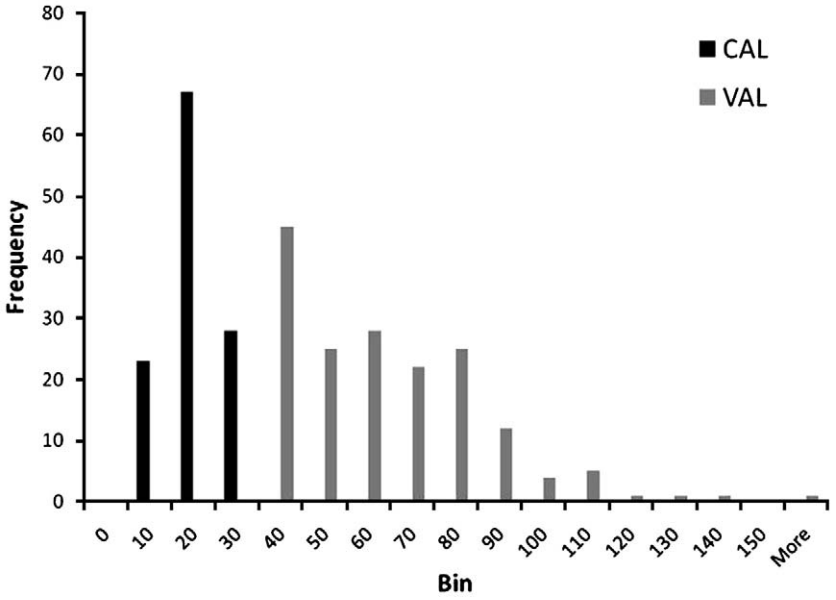

Fig. 1. Histogram of the glucosinolate content — calibration and validation set.

points (dp) were removed to produce spectra of 1000 dp from 450 nm to 2448 nm. The link between data points and wavelengths is: dp = (nm − 448)/2. The reverse is nm = dp*2 + 448. Fig. 2 presents the log (1/*R*) spectra of the calibration set.

3. Results

3.1. Participant no. 1

The calibration data set was preprocessed by SNV in order to correct the effects of offset. Principal Component Analysis (PCA) was applied as an exploratory technique, but it did not highlight possible outliers. Then, PLS was applied to construct a calibration model. In order to have a first idea on the capacity of the PLS model to adapt itself to new samples, the data was split into two groups: after ordering the samples by increasing order, one out of three is used in the validation set and two out of three in the training set.

The interval Partial Least Square regression (iPLS) (a graphically oriented local modelling procedure) method [7] was used to select the optimum wavelength range according to the following criteria: 10 latent variables maximum and a size of the interval of 10 variables. The local character of iPLS avoids the use of noisy or irrelevant spectral regions for prediction. The region obtained with this algorithm is 591–600 data points (1630–1648 nm) and after optimization, the spectral zone extends from 515 to 623 data points (1478–1694 nm). A PLS model was constructed in this region. According to the curves of RMSEC and RMSECV (Leave-One-Out Cross-Validation), the optimal number of latent variables was seven. Fig. 3 presents the linear regression with the predicted and the measured values for the internal validation set.

The characteristics of the model PLS are as follows: $R^2 = 0.87$, RMSEC = 2.3 and RMSECV = 3.0. The internal validation gives a value of SEP = 2.3.

3.2. Participant no. 2

PLS models were developed with 20 latent variables and different preprocessing methods: mean center (RMSEC = 2.70; $R^2 = 0.83$); mean center and first derivative using the Savitski and Golay algorithm [8] (RMSEC = 1.30; $R^2 = 0.96$); mean center and baseline correction (RMSEC = 2.80; $R^2 = 0.82$); mean center and smoothing (RMSEC =

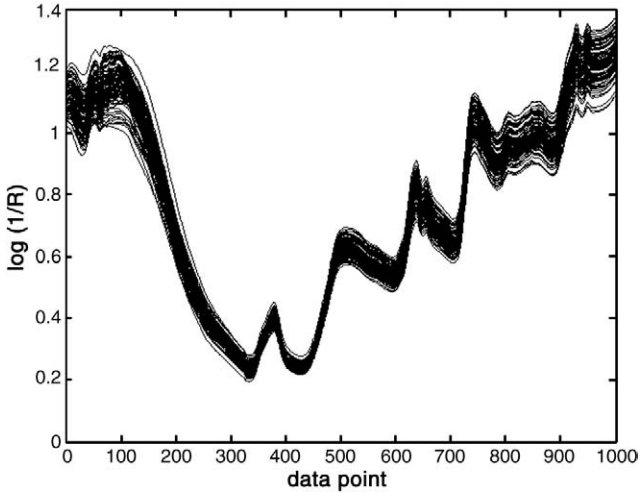

Fig. 2. Whole rapeseed spectra — calibration set between 450 and 2448 nm.

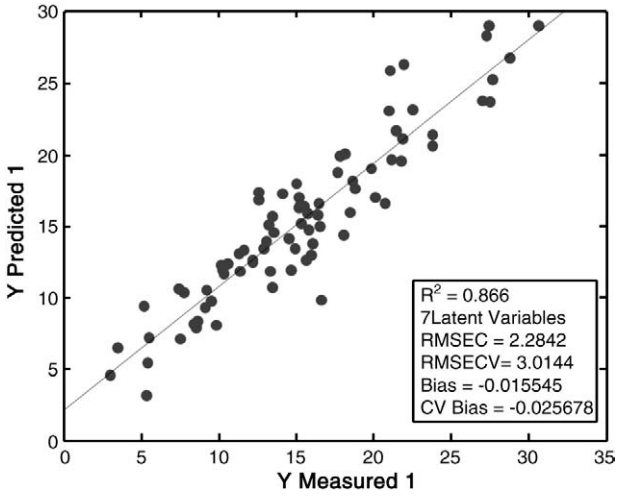

Fig. 3. Linear regression model for participant 1.

Please cite this article as: J.A. Fernández Pierna, et al., A case study of extrapolation in NIR modelling — A chemometric challenge at 'Chimiométrie 2009', Chemometrics and Intelligent Laboratory Systems (2010), doi:10.1016/j.chemolab.2010.04.001

2.78; $R^2 = 0.83$); mean center, smoothing and first derivative (RMSEC = 1.70; $R^2 = 0.93$). Then, the best model was selected as the one that showed the lowest Root Mean Square Error of Calibration (RMSEC) and the highest correlation coefficient (R^2), i.e. the model obtained with mean center and first derivative as preprocessing.

In this model, variable selection was applied using the genetic algorithm [9] with 20 latent variables (RMSEC = 1.29; $R^2 = 0.96$) and interval Partial Least Square (iPLS) [7] with intervals of ten (RMSEC = 3.71; $R^2 = 0.66$), five (RMSEC = 2.02; $R^2 = 0.91$), three (RMSEC = 1.64; $R^2 = 0.94$) and two (RMSEC = 1.25; $R^2 = 0.96$). The model was improved with iPLS and using only the interval between 501 and 1000 data points (1450–2448 nm) and the number of latent variables was 16.

In order to improve the model, outliers were detected based on different criteria [10–12]: Extreme leverage; un-modeled residuals in spectra and un-modeled residuals in dependent variables. Seven outliers were detected and removed. Then a better optimized model was obtained (RMSEC = 0.83; $R^2 = 0.98$).

To validate the best optimized model, different figures of merit such as accuracy, fit, sensitivity, analytical sensitivity, selectivity, linearity, limit of detection, limit of quantification and signal-to-noise ratio, were calculated (not shown) [13–19]. The model was validated using these figures of merit. The model showed a large sensitivity capacity, differentiating samples with a low difference of concentration. The values for accuracy and others figures of merit gave promising results, indicating that the developed model can be used as an alternative to estimate total glucosinolates. Fig. 4 shows the goodness of fit of the model. When looking at the plot of the residuals (not shown) of the calibration samples, the distributions of the errors present a random behavior that indicates the linearity of the multivariate model [10].

3.3. Participant no. 3

After examining the spectra, it was decided as a first step to calculate the 1st and 2nd derivatives using a 9-point Savitsky–Golay filter [8] in order to reduce baseline shift and enhance fine structure. The calculation of the derivatives revealed an irregularity in the spectra near 1098 nm, which was assumed to be due to an instrumental artifact. 8 points on each side of the breakpoint were removed in all three cases (raw data, 1st and 2nd derivatives). Variations in global signal intensity and baseline were then reduced by Standard Normal Variate pre-treatment [8]. Plots of the PC1 and PC2 scores of both the calibration and validation data sets revealed only one potential outlier, which was seen to present a baseline shift most visible in the zone 400–1098 nm. A plot of the y calibration values revealed no particular outliers. All samples were therefore retained and the raw spectra and derivatives were concatenated column-wise into a single calibration data matrix.

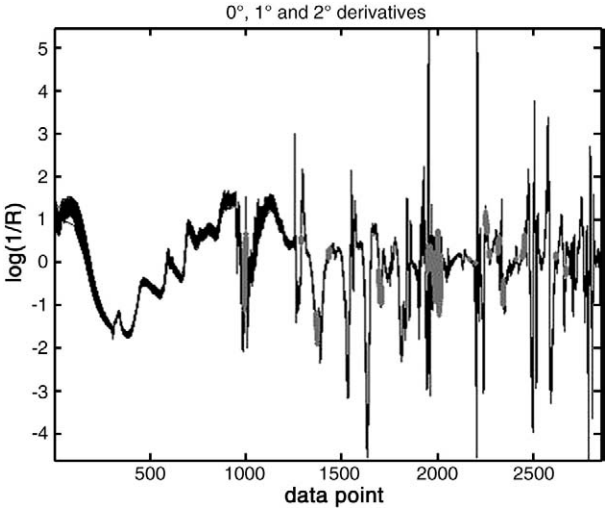

Fig. 5. Selected variables of participant 3.

Representative subsets of the calibration data matrix, to develop the model (80 spectra) and then test it (38 spectra), were selected using the Kennard & Stone method [8].

Informative variable selection was done on the model-development subset using the CLV method (Clustering of variables around Latent Variables) [20,21], where related variables were grouped together into 10 groups based on a Hierarchical Classification procedure around Latent Variables calculated by applying a PCA to each group of variables. The most informative group of variables was selected by doing a PLS regression, with Leave-One-Out Cross-Validation, between each group of variables and the y calibration values. The selected variables (in grey) are plotted in Fig. 5, where it can be seen that no variables were selected in the original spectral zone while similar variables were selected in the 1st derivative and 2nd derivative zones.

The Leave-One-Out Cross-Validation PLS regression on these variables gave an RMSECV = 2.84 for 13 Latent Variables, and an RMSEC = 1.81. The same PLS model applied to the model testing data set gave an RMSEV = 2.54. Fig. 6 shows the goodness of fit of the best model ($R^2 = 0.906$) for the calibration plus test set.

The spectra in the validation set were pre-treated in the same way and the same variables were retained to predict the corresponding unknown y values. A plot of these predicted values showed that 3

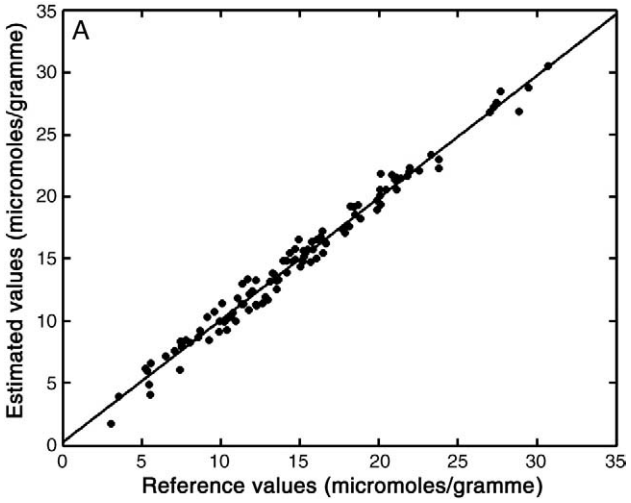

Fig. 4. Goodness of fit of the model for participant 2.

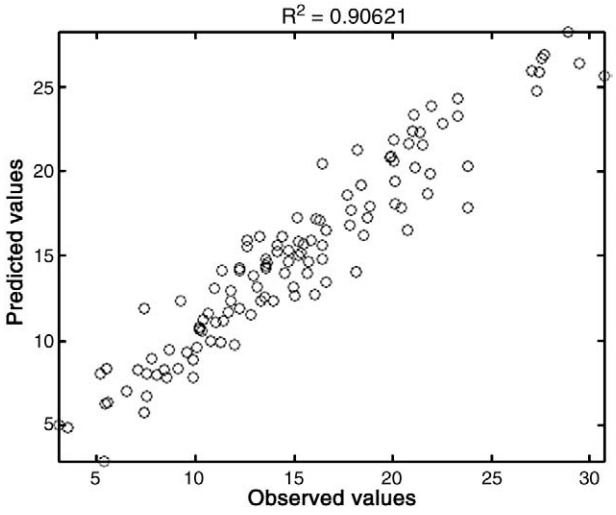

Fig. 6. Goodness of fit of the model for participant 3.

Please cite this article as: J.A. Fernández Pierna, et al., A case study of extrapolation in NIR modelling — A chemometric challenge at 'Chimiométrie 2009', Chemometrics and Intelligent Laboratory Systems (2010), doi:10.1016/j.chemolab.2010.04.001

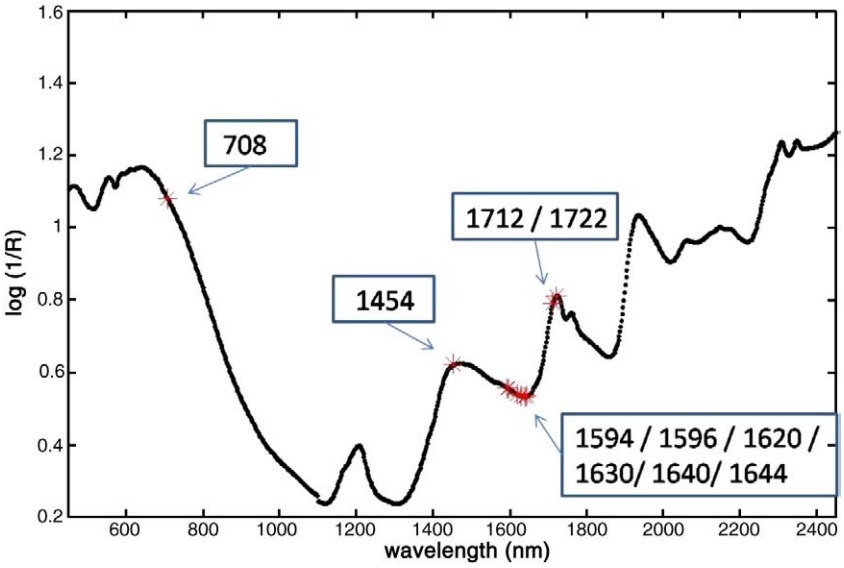

Fig. 7. Selected variables for the challenge organizer — approach 1.

samples gave particularly high values and that the range of predicted *y* values was much greater than that of the calibration samples.

3.4. Challenge organizers' approach

Two different approaches have been used. For the first approach, the data was first split into two subsets using the duplex algorithm [22], in total 65% of the samples were used for model construction and 35% for optimization. Then, a selection of the most important wavelengths was performed using PPLS (powered PLS) with Leave-One-Out Cross-Validation. PPLS [23] is a modification of the PLS algorithm. Stepwise optimization over a set of candidate loading weights obtained by taking powers of the *y*–*X* correlations and *X* standard deviations generalizes the classical PLS based on *y*–*X* co-variances. In total, 10 variables were retained (Fig. 7). Then MLR on the retained variables was performed and the final model gave a RMSEC of 2.29 and a R^2 of 0.86.

The second approach is based on work done in the 80s [5,6]. It was decided to use MLR since several wavelengths had been found to be very specific for glucosinolates. An automatic sequence search (Infrasoft International Ltd., DOS4.3) was used to calculate models with 60 different pre-treatments. The final optimal model was obtained using second derivative by the Savitski and Golay algorithm as pre-processing and retaining only 4 wavelengths (1270 nm, 1488 nm, 1616 nm and 1632 nm). Fig. 8 shows a PCA performed using only those variables. The final model gave a RMSEC of 2.36 and a R^2 of 0.85.

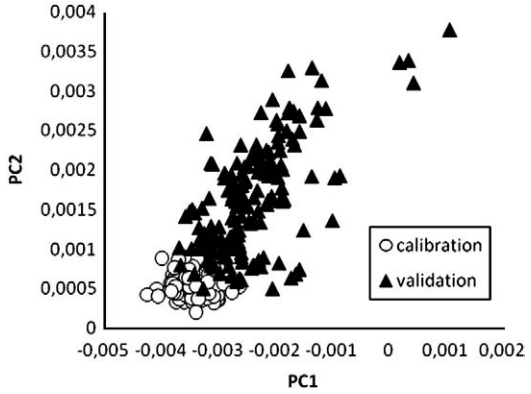

Fig. 8. PCA performed using only the 4 retained wavelengths (1270 nm, 1488 nm, 1616 nm and 1632 nm) for the challenge organizer — approach 2.

4. Final results

It was possible to detect the extrapolation 'trap' by applying any model and looking at the distribution of the predicted data. A simple PCA (log (1/*R* − 1100–2450)) showed also that 50% of the spectra had a GH (Global H = normalized Mahalanobis distance) higher than 3 (maximum GH = 20). These 2 elements could have been used to find this extrapolation problem. The spectral variation due to glucosinolates is tiny regarding the main sources of physico-chemical variation (particle size, moisture, fat, protein …).

The evaluation of the approaches was performed on the basis of the best results obtained on the predicted values for the blind test set. The reference values obtained by the reference method for these spectra were not communicated to the participants. The results for the different approaches presented in this paper are summarized in Table 1. For the evaluation, only the R^2 was retained.

5. Conclusion

A large diversity of results (R^2 from 0.0 to 0.89) was obtained during the challenge. However, when evaluating together all 7 answers received from the participants, it is interesting to note that with this Challenge 2009 we prove that it is still possible to perform a robust and precise calibration model when using only few specific wavelengths. The participants who succeeded in getting good results used variable selection algorithms. This particular case shows the danger of using full spectral models and sophisticated approaches when extrapolation occurs. The so called "analytical" model using specific regions or individual wavelengths has a much higher robustness to predict "unknown" new spectra which are outside the calibration range.

Table 1
Results for the blind test using the different approaches presented in this paper.

	Participant 1	Participant 2	Participant 3	Challenge org 1	Challenge org 2
RMSEP	9.73	15.71	20.09	9.47	10.42
Bias	−2.33	12.34	16.05	0.31	−1.6
Slope	0.8	1.21	1.31	0.79	0.77
Intercet	10.02	2.41	3.01	12.55	12.19
R^2	0.88	0.84	0.76	0.89	0.87
RDP	2.92	2.51	2.02	2.98	2.77

RPD (ratio of standard error of prediction to sample standard deviation.

Please cite this article as: J.A. Fernández Pierna, et al., A case study of extrapolation in NIR modelling — A chemometric challenge at 'Chimiométrie 2009', Chemometrics and Intelligent Laboratory Systems (2010), doi:10.1016/j.chemolab.2010.04.001

Acknowledgments

We would like to thank all the participants who spent time to treat the data and present their results. Apart from the authors, the other participants to the challenge were Jean-Claude Boulet (INRA, France), Jean-Philippe Vert (Mines ParisTech, France), Mélanie Dumaret (Vrije Universiteit Brussel, Belgium) and Adrian Tanesescu (Ritme, France).

References

[1] P. Dardenne, J.A. Fernández Pierna, A NIR data set is the object of a Chemometric contest at 'Chimiométrie 2004', Chemometrics and Intelligent Laboratory Systems 80 (2006) 236–242.
[2] J.A. Fernández Pierna, P. Dardenne, 'Chemometric contest at 'Chimiométrie 2005': a discrimination study', Chemometrics and Intelligent Laboratory Systems 86 (2007) 219–223.
[3] J.A. Fernández Pierna, P. Dardenne, Soil parameter quantification by NIRS as a Chemometric challenge at 'Chimiométrie 2006', Chemometrics and Intelligent Laboratory Systems 91 (2008) 94–98.
[4] J.A. Fernández Pierna, F. Chauchard, S. Preys, J.M. Roger, O. Galtier, V. Baeten & P. Dardenne (2010). 'How to build a robust model with only a few reference values: a chemometric challenge at 'Chimiométrie 2007'', Accepted for publication in *Chemometrics and Intelligent Laboratory Systems.*
[5] R. Biston, P. Dardenne, M. Cwikowski, J.P. Wathelet, M. Severin, 'Analysis of quality parameters of whole rapeseed by NIRS', Proceedings of the C.E.C. meeting on glucosinolates in rapeseeds, Gembloux-Belgium, 1–3 October, 1986, pp. 163–172.
[6] R. Biston, P. Dardenne, M. Cwikowski, M. Marlier, M. Severin, J.P. Wathelet, Fast analysis of rapeseed glucosinolates by near infrared reflectance spectroscopy, J. Amer. Oil. Chem. Soc. 65 (9) (1988) 1599–1600.
[7] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-square regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Applied Spectroscopy 54 (3) (2000) 413–419.
[8] M. Zeaiter, D. Rutledge, in: S. Brown, R. Tauler, R. Walczak (Eds.), 'Preprocessing Methods' in Comprehensive Chemometrics, Volume 3, Elsevier, 2009, pp. 121–231.
[9] J.H. Holland, Genetic algorithms, Sci. Am. 267 (1) (1992) 66–72.
[10] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1996.
[11] Annual Book of ASTM Standards, 'Standards practices for infrared, multivariate, quantitative analysis', E1655, vol 03.06, ASTM International, West Conshohocken, Pennsylvania, USA, 2000.
[12] P. Valderrama, J.W.B. Braga, R.J. Poppi, Variable selection, outlier detection, and figures of merit in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy, J. Agric. Food Chem. 55 (21) (2007) 8331–8338.
[13] J. Riu, F.X. Rius, Assessing the accuracy of analytical methods using linear regression with errors in both axes, Anal. Chem. 68 (11) (1996) 1851–1857.
[14] A.C. Moffat, A.D. Trafford, R.D. Jee, P. Graham, Meeting the International Conference on Harmonisation's guidelines on validation of analytical procedures: quantification as exemplified by a near-infrared reflectance assay of paracetamol in intact tablets, Analyst 125 (7) (2000) 1341–1351.
[15] P. Valderrama, J.W.B. Braga, R.J. Poppi, Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy, Journal of the Brazilian Chemical Society 18 (2) (2007) 259–266.
[16] A.C. Olivieri, N.K.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Uncertainty estimation and figures of merit for multivariate calibration, Pure and Applied Chemistry 78 (3) (2006) 633–661.
[17] J. Vessman, R.I. Stefan, J.F.V. Staden, K. Danzer, W. Lindner, D.T. Burns, A. Fajgelj, H. Miller, Selectivity in analytical chemistry — (IUPAC Recommendations 2001), Pure and Applied Chemistry 73 (8) (2001) 1381–1386.
[18] L.A. Currie, Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995), Pure and Applied Chemistry 67 (10) (1995) 1699–1723.
[19] A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations, Anal. Chem. 58 (6) (1986) 1167–1172.
[20] E. Vigneau, E.M. Qannari, Clustering of variables around latent components, Communications in statistics 32 (2003) 1131–1150.
[21] E. Vigneau, K. Sahmer, E.M. Qannari, D. Bertrand, Clustering of variables to analyse spectral data, J. Chemometr. 19 (2005) 122–128.
[22] R.D. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428.
[23] U. Indahl, A twist to partial least squares regression, J. Chemometr. 19 (2005) 32–44.

Please cite this article as: J.A. Fernández Pierna, et al., A case study of extrapolation in NIR modelling — A chemometric challenge at 'Chimiométrie 2009', Chemometrics and Intelligent Laboratory Systems (2010), doi:10.1016/j.chemolab.2010.04.001