Local vs global methods applied to large near infrared databases covering high variability

O. Minet, V. Baeten, B. Lecler, P. Dardenne and J.A. Fernández Pierna*

Walloon Agricultural Research Centre (CRA-W), Valorisation of Agricultural Products Department, Food and Feed Quality Unit, Henseval' Building, 24 Chaussée de Namur, 5030 Gembloux, Belgium. E-mail: foodfeedquality@cra.wallonie.be

The purpose of this study was to evaluate two different locally based regression methods (LOCAL and Local Calibration by Customized Radii Selection) and compare their performance to the classical global PLS for large NIR data. The data used in this study came from two inter-laboratory studies for wheat grain analysis organized in 2016 in the framework of the REQUASUD network. The results showed that improved predictions in terms of prediction errors can be obtained using local approaches compared to the classical global PLS. Moreover, the study highlighted clear differences between inter-laboratory studies and participating laboratories, which were even more evident when working with local procedures.

Introduction

Analyses performed by near-infrared (NIR) spectroscopy have become increasingly common in agriculture and the food industry. In several cases they even have replaced traditional chemical analysis. This is due to a large number of advantages (speed, cost-effectiveness, simultaneous determination of several constituents, little or no preparation of the sample, reduced use of chemical reagents etc.), but also to improvements to NIR instrumentation, allowing easier and faster spectrum acquisition at-line and on-line, as well as improvements in computer technology which have led to the development of more sophisticated new chemometric tools.[1]

Over the years, NIR databases have grown in terms of number of spectra and reference values. Especially in agriculture, extensive spectral databases containing thousands of NIRS reflectance spectra have been created during the last 30 years. The most extensive of these relate to feed/grain/silage and milk monitoring. Such databases have been used to build global calibrations, mainly PLS-based, which conceptually are expected to be very robust to sample composition variation. However, as databases get larger, this increases the complexity in terms of variability, and although this is normally seen as an advantage in global calibrations, in practice it creates a problem because prediction accuracy decreases.[2–4]

One possible solution would be to build specific calibration equations for small groups of similar samples, but this is cumbersome in practice and increases the complexity of the analytical systems. Another approach for modeling a complex calibration function that is much simpler and more flexible is local regression. Instead of constraining the calibration function to have a parametric form, it assumes that the data can be locally, i.e. around some neighborhood of the spectra, approximated by a parametric form, namely low-order polynomials. In other words, it computes a specific calibration equation for each sample analyzed using reduced calibration data extracted from a large library.[5] In practice this involves a group of methods based on selecting from a large database, a set of samples spectrally similar to an unknown sample whose properties are to be predicted. Following this strategy, a specific local model is then developed for that sample using the previously selected "neighborhood" samples as a calibration set. Figure 1 shows an example of data where global and local approaches are applied. The number of samples selected by a local approach is drastically reduced. This means that each sample is predicted with a different

Correspondence
J.A. Fernández Pierna (foodfeedquality@cra.wallonie.be)
doi: 10.1255/nir2017.045
Citation: O. Minet, V. Baeten, B. Lecler, P. Dardenne and J.A. Fernández Pierna, "Local vs global methods applied to large near infrared databases covering high variability", in *Proc. 18th Int. Conf. Near Infrared Spectrosc.*, Ed by S.B. Engelsen, K.M. Sørensen and F. van den Berg. IM Publications Open, Chichester, pp. 45–49 (2019). https://doi.org/10.1255/nir2017.045

ISBN: 978-1-906715-27-4

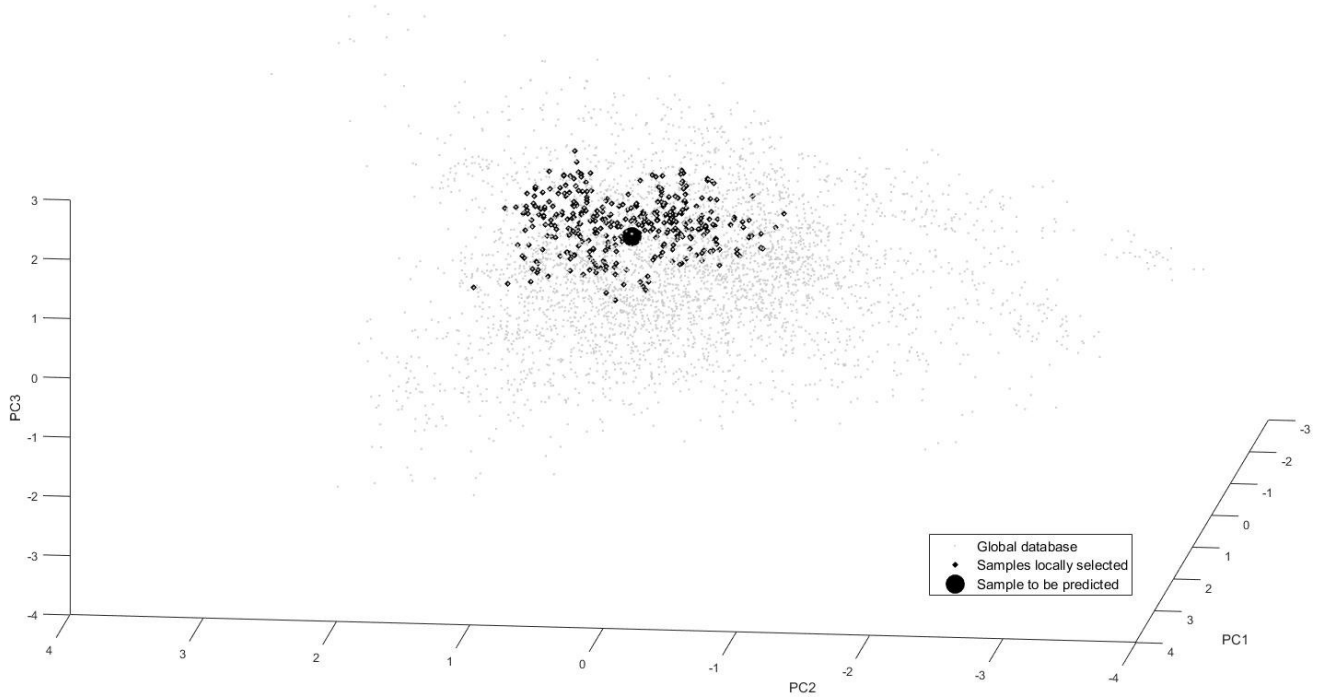

Figure 1. PC1 vs PC2 vs PC3 plot showing the difference between global vs local approaches on a PCA view: in the figure, the sample represented by the big black point will be predicted by all the grey points representing the whole database when using global PLS. However, when using local methods, the sample will be predicted by the black points (diamonds), representing only samples in the database which are similar to the sample for which a prediction is required. In the global method all the samples are selected to generate the model although some of them are spectrally different. In local methods, only similar spectra to the one for which a prediction is required are selected to build a specific calibration.

calibration equation. At present, several local regression-based methods exist, such as Locally Weighed Regression (LWR),[6] CARNAC,[7] LOCAL[2,8] and, most recently, Local Calibration by Customized Radii Selection.[9]

In local approaches the samples selected have to be similar enough to the sample to be predicted. The algorithm of prediction must be configured in a way that makes it possible to set up a threshold (based on GH criterion[10] or Mahalanobis distance, for instance) in order to select only the most similar spectra in the database. The database also has to be "clean", meaning that all outliers, especially property (reference values) outliers, have been removed from the dataset. Moreover, the database has to be large enough and cover most of the possible spectral variability encountered in routine analysis. The structure and distribution of the data have to be as homogenous as possible to avoid unbalanced data.

The purpose of this study was to evaluate two different locally based regression methods (LOCAL and Local Calibration by Customized Radii Selection) and compare their performance to the classical global PLS for a large NIR dataset. The data used in this study came from two inter-laboratory studies for wheat grain organized in the framework of the REQUASUD network in 2016.[11]

Material and methods

A first dataset of 4392 spectra of wheat grain covering crops of the last 25 years (1990–2015) was used as historical data to build the calibration PLS models or as clean data in local methods for predicting protein content. A second, completely independent dataset of 160 spectra (2 × 5 samples × 8 labs × 2 ILSs) was then used to compare the performance of the different methods. This second dataset was composed of spectra of samples coming from two inter-laboratory studies (ILSs) organized in March and July 2016 respectively within the REQUASUD NIR instrumentation network established in Wallonia (Belgium).

In this network, the eight participant laboratories were each equipped with a Foss XDS NIR spectrometer,

standardized yearly to the same master device located in the reference laboratory at CRA-W. An ILS consisted of two sets of five blind samples sent to each laboratory; the samples were selected in order to cover as large a range as possible of the protein content. The samples used for the two ILSs were the same. However, the laboratories were not aware of this fact. Each laboratory analyzed the samples blind by NIR but also by wet chemistry, and the assigned chemical values corresponded to the average values of all the results of the laboratory.

Three chemometric methods were used in this study:

i) MPLS: Modified PLS with 11 PLS factors and using the WinISI™ software;
ii) LOCAL: In this method, the selection of calibration samples is controlled by the value of the correlation coefficient between the spectrum of the unknown sample to be predicted and those of the available database. LOCAL using the WinISI™ software has been developed by Infrasoft International LLC (Foss) and it is described elsewhere.[4,8] In this study, LOCAL was set up to select between 50 and 250 samples from the clean dataset and between 6 and 20 PLS factors.
iii) LCCRS (Local Calibration by Customized Radii Selection): In this recent locally-based method, the number of samples selected to build each local model was automatically fitted and was based on the space between PLS scores: the distance between samples was measured considering spectral similarities but also reference value coincidences. The number of samples selected in the clean database and the number of PLS factors were optimized and could be different for each constituent and each product. The method has been described elsewhere.[9] This algorithm was executed with programs developed in Matlab 2015b (The Mathworks, Inc., Natick, MA, USA).

In all cases, the data were pretreated using SNV and first derivative Savitzky–Golay with a window of nine points and a polynomial of the second degree.[12]

Results

After application of the global and the two local approaches, the results for the 160 spectra were expressed in terms of Root Mean Square Error for Prediction (RMSEP). As previously explained, the assigned values were the mean of the results obtained by the labs for each ILS.

Table 1 shows the RMSEP values calculated for each laboratory and for each ILS separately. Table 2 shows the statistics extracted from Table 1, mainly in terms of standard deviation (SD) and coefficients of variation (CV) of the RMSEP obtained for each technique. As can be observed, the variation for RMSEP from the classical PLS was significant by a factor of 1 to 3.4. The coefficients of variation (CV) were high but lower for both local methods.

Table 3 shows the results when all predictions are averaged per sample, independently of the lab or the ILS.

From this table, a mean RMSEP value by method was calculated (Table 4). The local methods give the best accuracy, although all three techniques give very low RMSEP values provided the number of analyses for each sample is high (in this case 32). For comparison, the tolerance used for reference analysis in the international ring test of BIPEA is 2.8% of the assigned value. For a common sample of wheat in Belgium, the protein content is gener-

Table 1. RMSEP values obtained by each laboratory and for each ILS.

	RMSEP					
	Global		LOCAL		LCCRS	
Laboratory	ILS 1	ILS 2	ILS 1	ILS 2	ILS 1	ILS 2
1	0.30	0.34	0.24	0.26	0.21	0.36
2	0.44	0.37	0.45	0.46	0.39	0.42
3	0.23	0.33	0.24	0.32	0.18	0.28
4	0.51	0.27	0.41	0.29	0.40	0.32
5	0.16	0.33	0.29	0.27	0.19	0.22
6	0.15	0.31	0.28	0.20	0.21	0.28
7	0.26	0.32	0.27	0.36	0.33	0.27
8	0.34	0.38	0.41	0.35	0.35	0.31

Table 2. RMSEP statistics obtained for the three algorithms.

	RMSEP		
	Global	LOCAL	LCCRS
Min	0.15	0.20	0.18
Max	0.51	0.46	0.42
Range	0.36	0.26	0.24
Mean	0.32	0.32	0.30
SD	0.09	0.08	0.08
CV	29.14	24.88	26.32

Legend : Min = RMSEP minimum, Max = RMSEP maximum, Range = Min – Max, Mean = Mean of the RMSEP, SD = standard deviation, CV = coefficient of variation (SD/Mean*100)

ally close to 12 %. In this case, the tolerance is ±0.33. This demonstrates that, in some cases, NIR analysis can be as accurate as reference analysis provided the number of measurements is high enough to represent the variability of the sample.

Conclusion

Both local approaches proved to be efficient alternatives to global models for optimizing predictions, allowing the RMSEP to be reduced when dealing with large databases with high data variability. In addition to this feature, the simplicity and speed of the local approaches, mainly based on correlations or Mahalanobis distance measured in the PLS scores space, allows their application to on-line predictions. The results for both local approaches being similar, LCCRS presents the advantage of working without being associated with any specific software and independently of the instrument used.

Table 4. RMSEP obtained for each algorithm when all the spectra are averaged by sample.

Algorithm	RMSEP
Global	0.12
LOCAL	0.11
LCCRS	0.08

In both cases, as the system selects only the most similar spectra, the complexity of the dataset used to predict unknown samples is reduced by comparison with the entire database, and the number of PLS factors can be reduced for a simpler, more robust calibration with low coefficients without excessive noise. Moreover, having a specific database can help with problems of nonlinearity.

The use of local techniques will motivate the development of unique databases in which spectra and reference values of different products of different kinds could be merged together. Such unique databases are easier to manage than individual databases for different products. They can be easily updated by adding new samples and ensuring that the reference values are correct.

References

1. P. Dardenne, G. Sinnaeve and V. Baeten, "Multivariate calibration and chemometrics for near infrared spectroscopy: which method", *J. Near Infrared Spectrosc.* **8,** 229–237 (2000). https://doi.org/10.1255/jnirs.283
2. V. Baeten, H. Rogez, J.A. Fernández Pierna, P. Vermeulen and P. Dardenne, "Vibrational spectroscopy methods for the rapid control of agro-food products", in *Handbook of Food Analysis* (3rd Edn), Ed

Table 3. Results when all predictions are averaged per sample, independently of the lab or the ILS for the three algorithms.

	% Protein			
	Reference values	Global	LOCAL	LCCRS
Sample 1	10.85	10.70	10.70	10.74
Sample 2	11.96	11.99	11.82	12.06
Sample 3	12.31	12.47	12.22	12.29
Sample 4	12.89	13.04	12.93	12.95
Sample 5	10.96	10.89	10.88	10.86

by L.M.L. Nollet and F. Toldra. Volume II, Chapter 32, pp. 591–614 (2015). https://doi.org/10.1201/b18668

3. O. Abbas, P. Dardenne and V. Baeten, "Near-infrared, mid-infrared, and Raman spectroscopy", in *Chemical Analysis of Food: Techniques and Applications*, Ed by Y. Picó. Elsevier Science, pp. 59–89 (2012). https://doi.org/10.1016/B978-0-12-384862-8.00003-0
4. J.S. Shenk, P. Berzaghi and M.O. Westerhaus, "Investigation of a LOCAL calibration procedure for near infrared instruments", *J. Near Infrared Spectrosc.* **5,** 223–232 (1997). https://doi.org/10.1255/jnirs.115
5. R.P. Hafen, *Local Regression Models: Advancements, Applications and New Methods*. Purdue University, West Lafayette, Indiana (2010).
6. F.E. Barton, J.S. Shenk, M.O. Westerhaus and D.B. Funk, "The development of near infrared wheat quality models by locally weighted regressions", *J. Near Infrared Spectrosc.* **8,** 201–208 (2000). https://doi.org/10.1255/jnirs.280
7. A.M.C. Davies, H.V. Britcher, J.G. Franklin, S.M. Ring, A. Grant and W.F. McClure, "The application of Fourier-transformed near-infrared spectra to quantitative analysis by comparison of similarity indices (CARNAC)", *Microchimica Acta* **94(1–6),** 61–64 (1988). https://doi.org/10.1007/BF01205839
8. P. Berzaghi, J.S. Shenk and M.O. Westerhaus, "LOCAL prediction with near infrared multi-product databases", *J. Near Infrared Spectrosc.* **8,** 1–9 (2000). https://doi.org/10.1255/jnirs.258
9. F. Allegrini, J.A. Fernández Pierna, W.D. Fragoso, A.C. Olivieri, V. Baeten and P. Dardenne, "Regression models based on new local strategies for near infrared spectroscopy", *Anal. Chim. Acta* **933,** 50–58 (2016). https://doi.org/10.1016/j.aca.2016.07.006
10. J.A. Guthrie, *Robustness of NIR Calibrations for Assessing Fruit Quality*. PhD thesis, Central Queensland University, Rockhampton (2005).
11. O. Minet, F. Ferber, L. Jacob, B. Lecler, R. Agneessens, T. Cugnon, V. Decruyenaere, V. Genot, S. Gofflot, E. Pitchugina, V. Planchon, M. Renesson, G. Sinnaeve, B. Wavreille, P. Dardenne and V. Baeten, *La Spectrométrie Proche-Infrarouge : Une Technologie Rapide, Precise et* Écologique *pour Determiner la Composition et la Qualité des Produits Agricoles et Alimentaires* (2016).
12. A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures", *Anal. Chem.* **8,** 1627 (1964). https://doi.org/10.1021/ac60214a047