# Local vs Global methods applied to large NIR databases covering high variability

Olivier Minet [1]*, J.A. Fernández Pierna[1], B. Lecler[1],V. Baeten [1] and P. Dardenne[1]

[1]Walloon Agricultural Research Centre (CRA-W), Valorisation of Agricultural Products Department

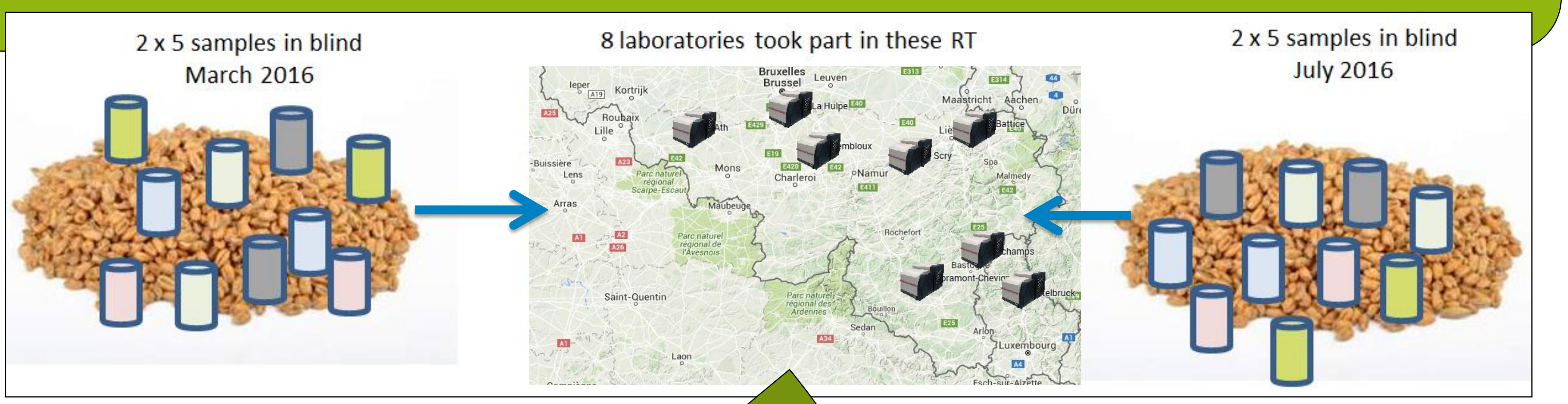
Chaussée de Namur, 24, B-5030 Gembloux, Belgium;

*Contact person: o.minet@cra.wallonie.be; FoodFeedQuality@cra.wallonie.be

## Context

REQUASUD is a network of laboratories in Belgium equipped with Foss XDS NIR spectrometers involved in agricultural products analysis. In this context, twice a year, ring tests (RT) are organized to evaluate the performances of each participant laboratory.

One of these RT consisted in the analysis of 5 samples of wheat grain sent in duplicate and in blind to 8 laboratories. The data used come for the RT of 2016 where the same samples were sent in March (RT1) and in July (RT2) with the pairs randomly changed. The samples have been analyzed by wet chemistry for protein content (Dumas method).

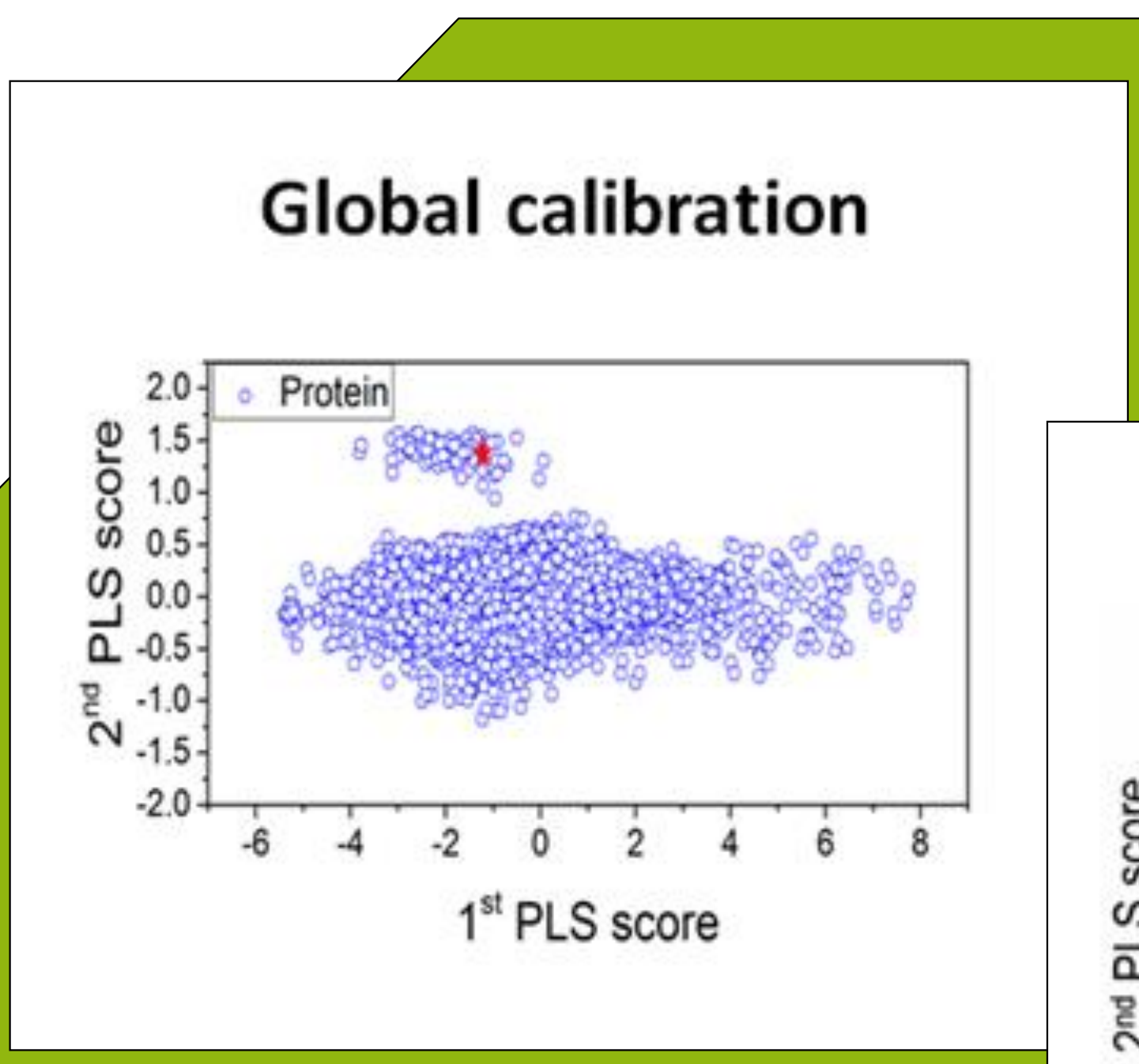


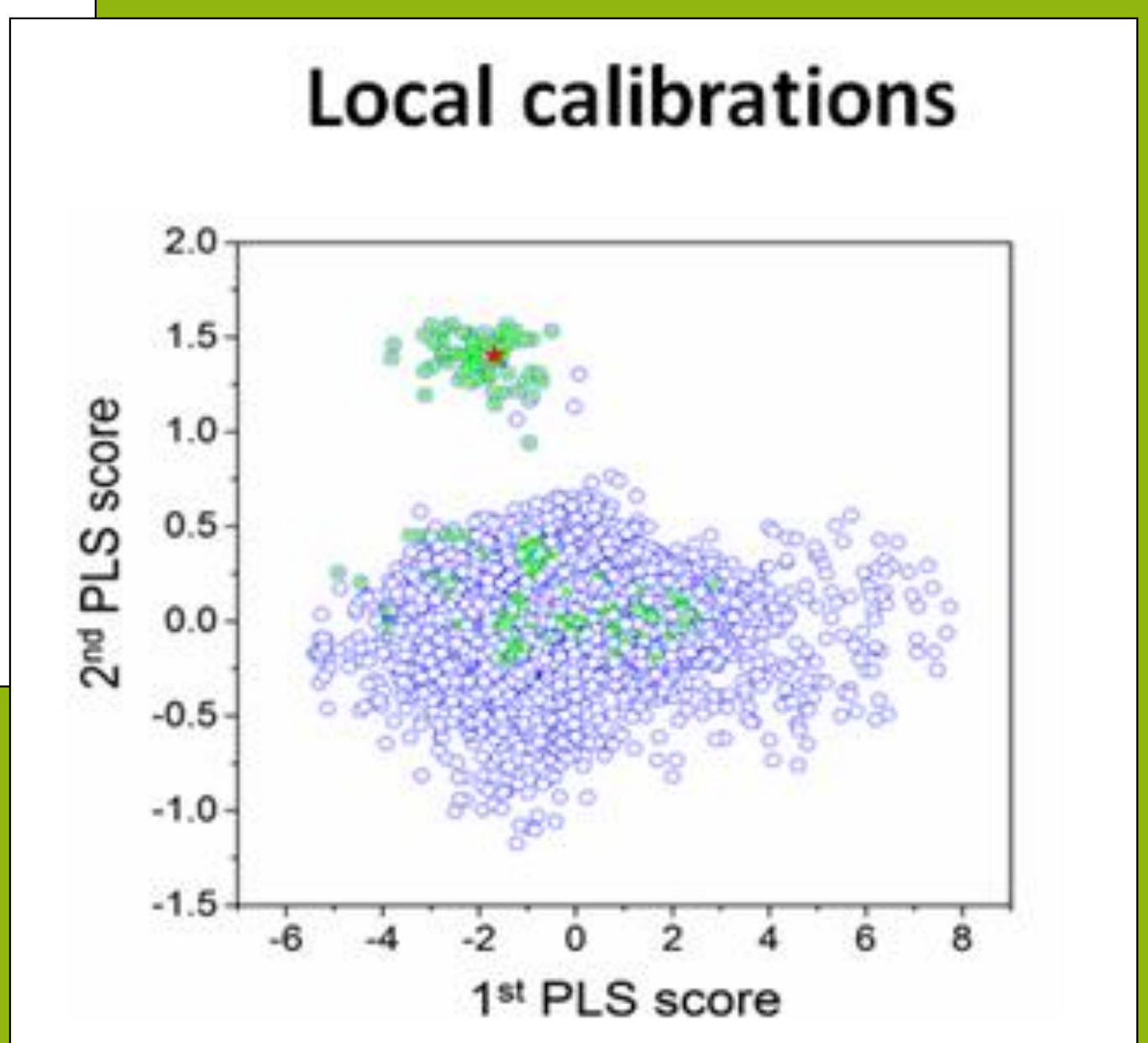
## Aim

The purpose of this study is not to compare the results of each lab but to compare different methods of regression :

- Global PLS
- Local PLS (Shenk algorithm)
- The new Local Calibration by Customized Radii Selection (LCCRS/RADIUS).

## Global vs Local modelling





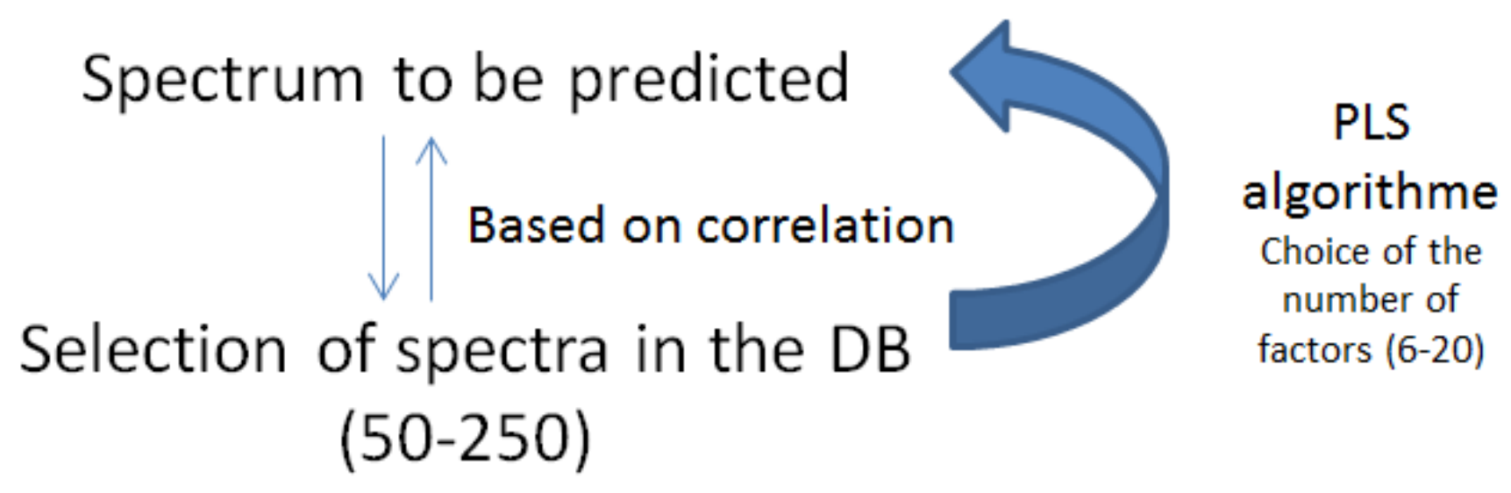
**Global calibrations** in principle are expected to be very robust to sample composition variation. However, in practice the prediction accuracy (in terms of RMSEP) usually decreases when the database gets larger.

**Local methods** present the main advantage to build specific calibration with spectra which are very similar to the spectrum of the sample to be predicted. A specific model is usually more accurate than a global model.

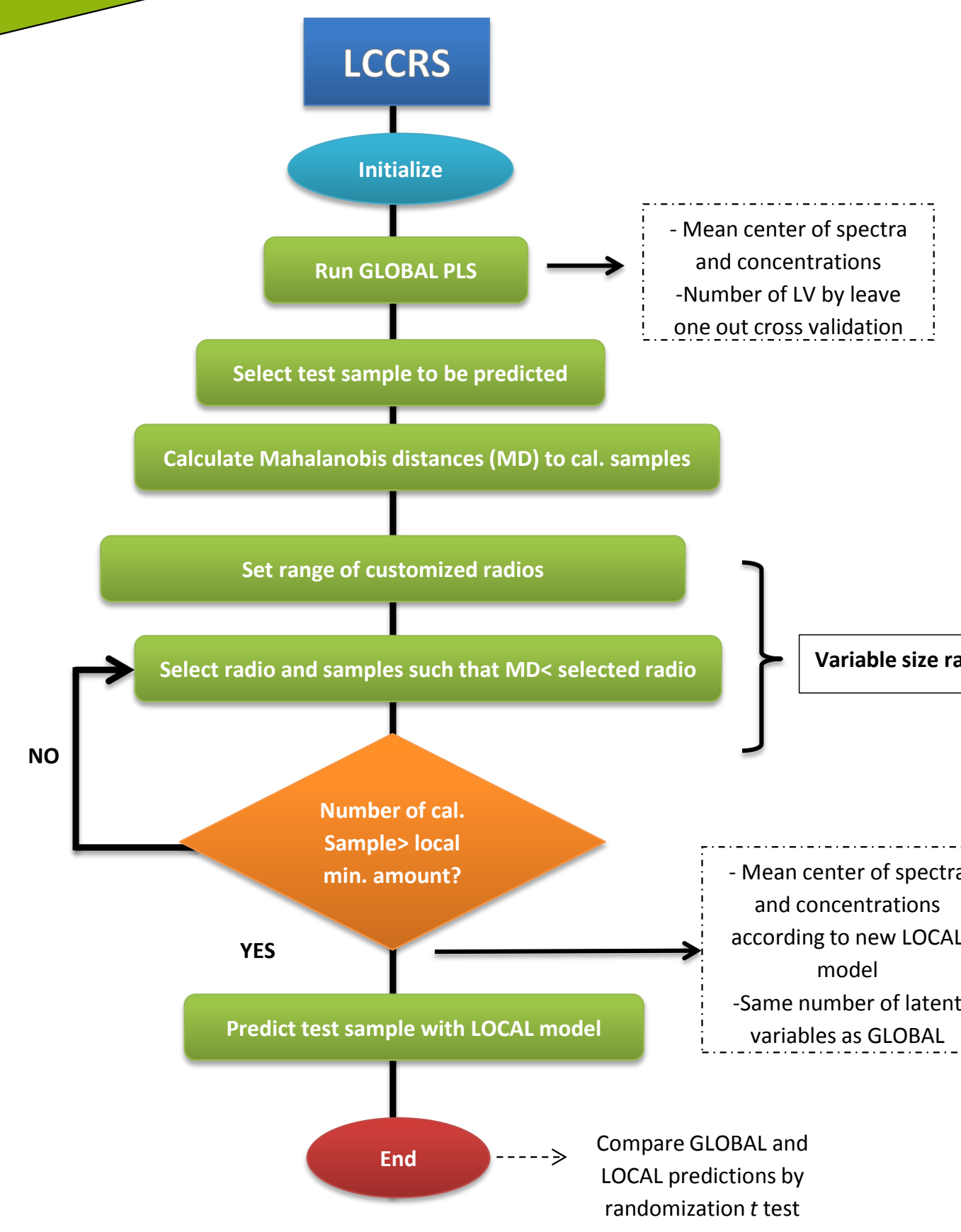
## Two local methods

### LOCAL

*J.S. Shenk, P. Berzaghi & M.O. Westerhaus, J. Near Infrared Spectrosc. 5, 223-232 (1997)*



In the Shenk algorithm the selection of calibration samples is controlled by the value of the correlation coefficient between the spectrum of the unknown sample and those of the database. Then a PLS regression is applied on the selected spectra. The Shenk algorithm is linked to the Foss instruments.


LCCRS
Initialize
Run GLOBAL PLS
- Mean center of spectra and concentrations
-Number of LV by leave one out cross validation
Select test sample to be predicted
Calculate Mahalanobis distances (MD) to cal. samples
Set range of customized radios
Select radio and samples such that MD< selected radio
Variable size radio
NO
Number of cal. Sample> local min. amount?
YES
- Mean center of spectra and concentrations according to new LOCAL model
-Same number of latent variables as GLOBAL
Predict test sample with LOCAL model
End
Compare GLOBAL and LOCAL predictions by randomization t test

### Local Calibration by Customized Radii Selection (LCCRS)

*F. Allegrini; J.A. Fernández Pierna;W.D. Fragoso; A.C. Olivieri; V. Baeten & P. Dardenne. Analytica Chimica Acta 933, 50-58 (2016)*

In the LCCRS method the number of samples selected in order to build each local model is automatically fitted and, it operates on the PLS scores space, meaning that the distance between samples is measured considering spectral similarities but also reference values coincidences.

## Results

A total of 160 spectra (5 samples x 2 (duplicate) x 2 RT x 8 labs) from the network have been collected and in order to predict protein content with the three regression methods using the CRA-W database including data from 1990 to 2015 (>4000 spectra). Those 160 samples have been also analyzed according the DUMAS method (ISO 16634-2 : 2016).

The reference values used to evaluate the NIR predictions (RMSEP) are the mean results obtained by each lab for the two RT. No outliers values have been detected by statistical tests (Grubbs, Cochran and Z-scores).

### By lab

RMSEP

| | RMSEP | | | | | |
|---|---|---|---|---|---|---|
| | NIR global PLS | | NIR local Shenk | | NIR local radius | |
| Laboratory | PLS_RT1 | PLS_RT2 | Local_winisi_RT1 | Local_winisi_RT2 | Local_radius_RT1 | Local_radius_RT2 |
| 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 |

### By sample

Protein content

| | % Protein | | | |
|---|---|---|---|---|
| | Reference values | Global PLS | Local (Shenk) | Local (Radius) |
| 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 |

The RMSEP values decrease drastically when all the spectra are averaged. The local methods, and especially the radius local, give the best accuracy. Anyway the three methods give very low values of RMSEP provided the number of scan for each sample is important (in this case 32).

### By averaging

RMSEP

| Algorithm | RMSEP |
|---|---|
| Global PLS | 0.12 |
| Local (Shenk) | 0.11 |
| Local (Radius) | 0.08 |

## Conclusion

It emerges from this study that local techniques are a good tool when dealing with large databases covering a high variability in the data. In this case, LCCRS gives better results than the Shenk local algorithm. Moreover, it presents the advantage to work without being associated to any specific software and independently of the instrument used. To achieve a good level of accuracy, it is necessary to scan several times every sample.

## Acknowledgements

This study was carried out through the work of the CRA-W staff of the Unit 15 for the preparation of the samples constituting the ring tests.

Special thanks also to REQUASUD for the use of the data.

Walloon Agricultural Research Centre

Wallonie