NIRS predictions on heterogeneous databases

L. Thuriès[1,2], D. Bastianelli[2], L. Bonnal[2], F. Davrieux[2]

1- Phalippou-Frayssinet S.A.,Organic Fertilisers, La Mothe, F81240 Rouairoux, France. thuries@cirad.fr

2- CIRAD. Centre de Coopération Internationale en Recherche Agronomique pour le Développement, TA40/01, F34398 Montpellier cedex 05, France.

Introduction

Faced to heterogeneous database questions, a NIR user is often adviced to work on more homogeneous datasets. However, as heterogeneity and variability are widespread among agriculture areas, it is not always possible to have subsets which are at the same time homogeneous and large enough for calibration. It is therefore interesting to try calibration on heterogeneous databases before saying it is impossible …

The major objective of this study was to compare different strategies for NIR predictions. On one hand, build models from a dataset comprising different data-subsets, and on another hand, compare them to models based on the 'pure' datasubsets.

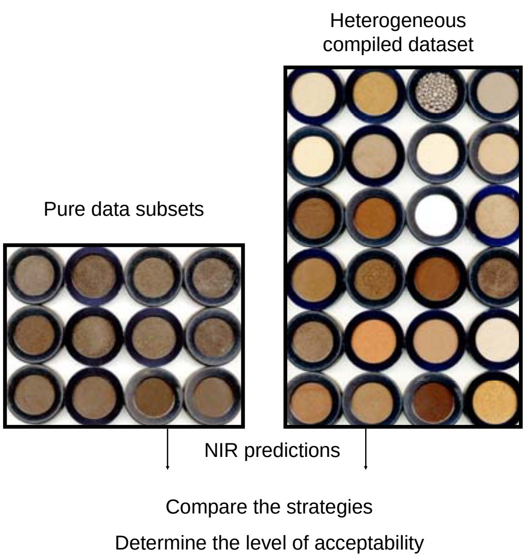

Fig 1. : NIR predictions on pure vs heterogeneous datasets

Materials and methods

The organic materials originated from

(i) industrially pre-processed plant residues, principally collected in the largest organic fertiliser factory in France

(ii) and tropical plant residues samples collected on-field in Brazil and Kenya, potentially utilisable in composting.

Pure datasets were: (a) wet grape skins, (b) dry grape skins, (c) de-oiled grape pips, (d) coffee cake, (e) cocoa cake, (f) olive pulp, (g) tropical plant residues samples. The compiled dataset comprised all these subsets.

The parameters measured were Organic Matter (OM, by loss on ignition), and Total Nitrogen Kjeldahl (TN).

Due to the heterogeneity of fresh materials, samples were dried (40°C) ground (<1mm sieve) before being scanned on a NIRS 6500 (Foss NIRSystems) in duplicate in ring cups. Spectra acquired in reflectance were corrected with SNVD 2,5,5 (WIN-ISI) mathematical treatment. Calibrations were performed using a modified partial least square regression (mPLS, WIN-ISI).

Results and discussion

The models developed with the compiled dataset were accurate for both parameters (Tab.1, Tab.2). For such an heterogeneous database, the R² equalled or overpassed 0.9, and the RPD were around 3.

Table 1: Performance of OM calibration models for the compiled and pure datasets

		population		calibration statistics			
material	**n**	**mean**	**SD**	**SEC**	**R²**	**SECV**	**RPD**
compiled dataset	309	93.2	2.96	0.94	0.90	1.07	2.8
wet grape skins	54	92.1	1.67	0.73	0.81	1.27	1.3
dry grape skins	47	92.4	1.64	0.59	0.87	0.91	1.8
de-oiled grape pips	40	95.8	0.94	0.47	0.75	0.59	1.6
coffee cake	26	98.8	0.77	0.28	0.86	0.44	1.7
defatted cocoa cake	49	90.9	1.26	0.75	0.64	0.86	1.5
olive pulp	46	91.4	1.78	0.57	0.90	0.78	2.3
tropical residues	43	93.4	3.59	0.41	0.99	0.92	3.9

SD, Standard Deviation of parameter in the population
SEC, Standard Error of Calibration
SECV, Standard Error of Cross-Validation
RPD = SD / SECV

Table 2: Performance of TN calibration models for the compiled and pure datasets

		population		calibration statistics			
material	**n**	**mean**	**SD**	**SEC**	**R²**	**SECV**	**RPD**
compiled dataset	272	2.30	0.54	0.16	0.91	0.17	3.1
wet grape skins	53	2.56	0.36	0.10	0.92	0.17	2.1
dry grape skins	50	2.31	0.17	0.10	0.63	0.12	1.4
de-oiled grape pips	44	1.99	0.26	0.12	0.79	0.14	1.9
coffee cake	32	0.46	0.46	0.11	0.94	0.17	2.6
defatted cocoa cake	48	2.84	0.64	0.15	0.95	0.18	3.7
olive pulp	43	1.83	0.18	0.10	0.69	0.12	1.5

Both OM and TN were generally better predicted for pure datasets (Tab.1, Tab.2, Figure 2). For example, the SEC for OM were 1/3 to ¾ that of the compiled dataset. The corresponding SECV were also lower, excepted for wet grape skins, probably due to the inner heterogeneous nature of these samples resulting in inappropriate reference values. On average, the SECV were 0.82% dry weight for OM.

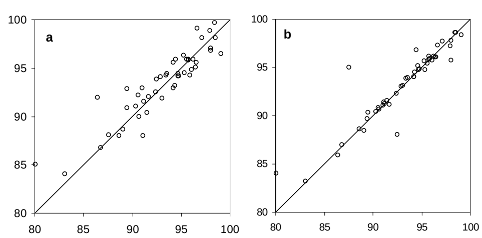

Figure 2. OM predictions for the "tropical" dataset with (**a**) the compiled equation and (**b**) tropical equation.

For TN, SEC were all lower than the compiled dataset's. SECV were also under or equalled the compilation's, excepted for cocoa where some outliers raised the SECV. On average, the SECV were 0.15% dry weight for TN.

For pure datasets (cocoa excepted), SECV were in general largely higher than SEC, whereas SECV were close to SEC for the compiled dataset. This result tend to indicate that the models developed for the compiled dataset were more stable than those for the pure ones.

The SECV were far under the normative tolerances (max 3.0 g $100g^{-1}$ bulk weight for OM, and min-max 0.2 – 0.3 g $100g^{-1}$ bulk weight for NT) for organic soil improvers (French Norm NFU#44051).

Conclusions and perspectives

Calibrations on pure datasets seem to perform slightly better (SECV) than that of the compilation. Nevertheless, models developed on the global dataset (made by compilation of the subsets, thus heterogeneous) had an acceptable predictive capacity and this strategy is therefore very useful.