

Near infrared spectroscopy predictions on heterogeneous databases

Thuriès, L.^{a, b}; Bastianelli, D.^c; Bonnal, L.^c and Davrieux, F.^d

^a *Phalippou-Frayssinet S.A., Organic Fertilisers, 81240 Rouairoux, France. E-mail: thuries@cirad.fr*

^b *CIRAD, Laboratoire Matière Organique Sols Tropicaux, UPR078, TA 40/01, 34398 Montpellier Cedex 5, France*

^c *CIRAD, Laboratoire d'Alimentation Animale, TA 30/A Baillarguet, 34398 Montpellier Cedex 5, France*

^d *CIRAD, TA 80/16, 34398 Montpellier Cedex 5, France*

Keywords: near infrared spectroscopy, calibration, heterogeneity, plant material, organic matter

Introduction

Faced with heterogeneous database questions, the user of near infrared (NIR) spectral databases is often advised 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 with hundreds, and even thousands, of samples for a local calibration [1]. 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 spectroscopy predictions. This involved comparing the performance of models developed from “pure” datasets containing relatively small numbers of samples with a model developed from a larger combined dataset encompassing a wide variability.

Materials and methods

Organic materials

The raw materials originated from (i) industrially pre-processed plant residues, principally collected in the largest organic fertiliser factory in France or from other sources; and (ii) tropical plant residues samples collected from the field in Brazil and Kenya as parts of trees, shrubs, crops and cover crops. The tropical material included total above ground material, roots, stems, twigs, pods, leaves and litters. Pure datasets were (a) wet grape skins, (b) dry grape skins, (c) de-oiled grape pips, (d) coffee cake, (e) de-fatted cocoa cake, (f) olive pulp and (g) tropical plant residues samples. The combined dataset comprised all seven “pure” subsets.

Sample preparation and reference analyses

Each sample was analyzed for its moisture content by drying to constant weight in an oven at 105°C. Subsets of samples were measured for organic matter (OM) content by subtracting the ash content (weight remaining after ignition at 525°C overnight) from the original dry weight of sample and for total nitrogen (TN) content (Kjeldahl method). Due to the heterogeneity of fresh materials, samples were rapidly dried in an aerated oven at 40°C to prevent nitrogen volatilization and Maillard reactions[2], and ground to pass a 1 mm sieve.

Sample scanning and data analysis

Each ground sample was scanned on a NIRS 6500 (Foss NIRSystems, Silver Spring, MD, USA) in duplicate in ring cups. Spectral data were collected every 2 nm from 400 to 2,498 nm. Individual spectra, each consisting of the average of 32 scans, were stored as log (1/reflectance), and corrected with a standard normal variate and detrend (2,5,5) (Win-ISI, Infrasoft International, Port Matilda,

PA, USA) mathematical treatment [3]. Visible wavelengths were discarded as they introduced instability in the models. Calibrations of the parameters studied were performed using a modified partial least square regression (WIN-ISI, Infrasoft International, Port Matilda, PA, USA) [4]. The standard error of calibration (SEC), the coefficient of determination (R^2), and the standard error of cross-validation (SECV) were calculated. In order to minimize overfitting of the equations, cross-validation was used as internal validation during calibration development.

Results and discussion

Both OM and TN were generally better predicted for pure datasets than for the combined dataset (Table 1, Table 2, Figure 1). The models developed with the combined dataset were accurate for both parameters. For such a heterogeneous database, the R^2 equalled or surpassed 0.9, and the RPD were around 3.

The OM predictions for the tropical residue dataset were dispersed around the 1:1 line when calculated with the combined dataset equation (Figure 1). When calculated with the equation specially dedicated to the tropical residue dataset, the OM contents were better predicted.

In general, the SEC for OM were one third to three quarters that of the combined dataset (Table 1, Table 2). On average, the SECV were 0.82% dry weight for OM. The corresponding SECV values were also lower, excepted for wet grape skins, probably due to the inner heterogeneous nature of these samples resulting in inappropriate reference values. Even if the grape residues originated from several varieties, growing regions or have been submitted to different types of wine elaboration processes, each data subset for grape parts could be considered more homogeneous than the tropical residue dataset. The same remark can be made for coffee with different varieties, origins and roasting procedures, for cocoa with different varieties and origins, and for olive residues with different varieties, origins and oil extraction procedures, they can be considered more homogeneous than the tropical residue dataset.

For TN (Table 2), SEC values were all lower than that of the combined dataset. On average, the SECV values were 0.15% dry weight for TN. The SECV values were also under or equal to that for the combined dataset, excepted for cocoa where some outliers raised the SECV.

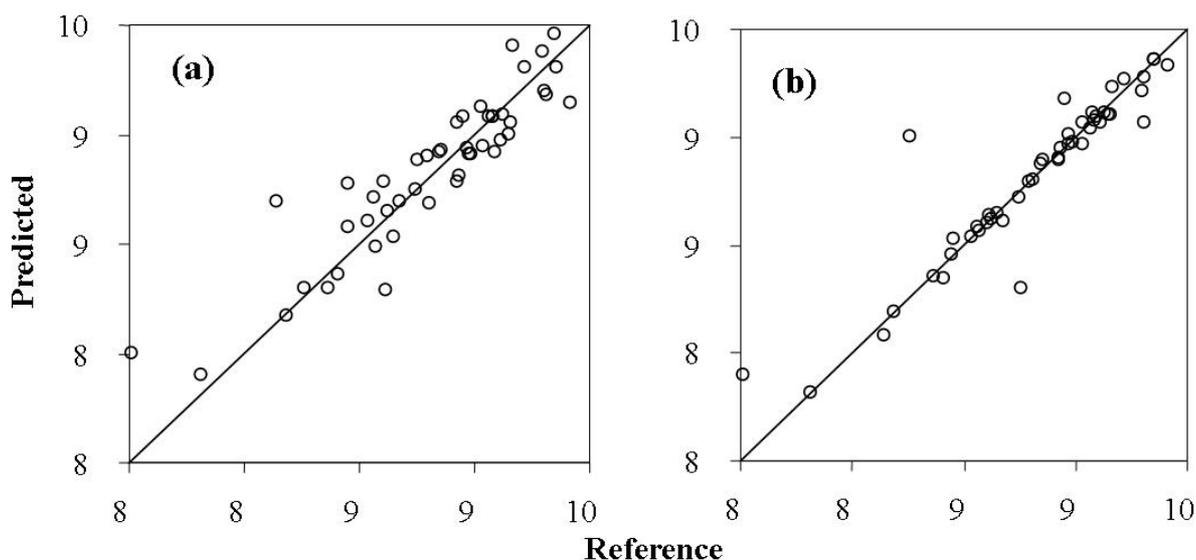


Figure 1. Organic matter (OM) predictions ($\text{g}\cdot 100\text{ g}^{-1}$ dry matter) for the tropical residue dataset with (a) the combined equation, and (b) the tropical residue equation.

Table 1. Performance of organic matter (OM) calibration models constructed using partial least squares procedures and spectra corrected with a standard normal variate and detrend 2,5,5 for the combined and pure datasets.

Material	Population			Calibration statistics			
	n	Mean	SD	SEC	R ²	SECV	RPD
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
De-fatted cocoa cake	49	91.0	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
Combined dataset	309	93.2	2.96	0.94	0.90	1.07	2.8

n: number of samples

SD: standard deviation of parameter in population

SEC: standard error of calibration

R²: coefficient of determination of calibration

SECV: standard error of cross-validation

RPD: ration of performance to deviation (SD·SECV⁻¹)

Table 2. Performance of total nitrogen (TN) calibration models constructed using partial least squares procedures and spectra corrected with a standard normal variate and detrend 2,5,5 for the combined and pure datasets.

Material	Population			Calibration statistics			
	n	Mean	SD	SEC	R ²	SECV	RPD
Wet grape skins	53	2.6	0.36	0.10	0.92	0.17	2.1
Dry grape skins	50	2.3	0.17	0.10	0.63	0.12	1.4
De-oiled grape pips	44	2.0	0.26	0.12	0.79	0.14	1.9
Coffee cake	32	2.1	0.46	0.11	0.94	0.17	2.6
De-fatted cocoa cake	48	2.8	0.64	0.15	0.95	0.18	3.7
Olive pulp	43	1.8	0.18	0.10	0.69	0.12	1.5
Combined dataset	272	2.3	0.54	0.16	0.91	0.17	3.1

n: number of samples

SD: standard deviation of parameter in population

SEC: standard error of calibration

R²: coefficient of determination of calibration

SECV: standard error of cross-validation

RPD: ration of performance to deviation (SD·SECV⁻¹)

For pure datasets, the SECV values were in general largely higher than the SEC values, whereas the SECV values were close to the SEC values for the combined dataset. With the exception of cocoa, the SECV values for the OM models were as high as almost twice the corresponding SEC values, particularly for wet and dry grape skins and tropical residue. The differences between SEC and SECV values for the TN models were also relatively large with +70% for wet grape skins and +20% for olive pulp. This result tends to indicate that the models developed for the combined dataset were more stable than those for the pure datasets.

As the SECV values were far under the normative tolerances of a maximum of 3.0 g.100g⁻¹ bulk weight for OM, and a range of 0.2 to 0.3 g.100g⁻¹ bulk weight for TN for organic soil improvers (French Norm NFU44-051) [5], all the models developed here could be used for quality control on-site in the organic fertilizer factory.

Conclusions

Calibrations on pure datasets seem to perform slightly better with a lower SECV than calibrations on a combined database. Nevertheless, models developed on a global dataset made by combining many subsets, had an acceptable predictive capacity. Using one unique combined model would be easier to use than maintaining six models dedicated to particular materials. The risk of making an error in prediction for an “out of range” or atypical material would then be reduced. When a local calibration is impossible due to a reduced number of samples or when calibrations dedicated to a unique type of material are not economically viable a combined approach can be used with confidence..

References

1. G. Sinnaeve, P. Dardenne and R. Agneessens, *J. Near Infrared Spectrosc.* **2**, 163 (1994).
2. P.J. Van Soest and V.C. Mason, *Anim. Feed Sci. Technol.* **32**, 45 (1991).
3. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc.* **43**, 772 (1989).
4. J.S. Shenk and M.O. Westerhaus, *Crop Sci.* **31**, 469 (1991).
5. NFU 44-051 *Amendements organiques*, Association Française de Normalisation (AFNOR), Paris, France, p. 694. (1981).