Determination of the content of six major biochemical compounds of green coffee using NIR

F. Davrieux[a], J.C. Manez, N. Durand and B. Guyot

[a] *CIRAD Centre de Coopération Internationale en Recherche Agronomique pour le Développement, TA30/A, 34398 Montpellier cedex 5, France. Fabrice.davrieux@cirad.fr*

Introduction

The most common coffee tree species are *Coffea arabica* and *Coffea canephora*, better known as Arabica and Robusta. Arabica accounts for 70% of world coffee production and Robusta around 30%.

The biochemical composition of green coffees can be used to characterise quality. Certain biochemical constituents, such as trigonelline, sucrose and chlorogenic acids, are flavour precursors, others such as caffeine play a role in the bitterness of roasted coffee and fat helps in fixing the flavour compounds formed during roasting. With knowledge of the water content, it is possible to characterise green coffee preparation and storage conditions.

Near infrared spectroscopy is a rapid and non-destructive way of predicting the biochemical composition of green coffee after calibration for each constituent involved.

Experimental procedure

For green coffee substantial variability within *Coffea* species exists as well as in the cultural and post-harvest practices used in the commodity chain.

In order to represent this substantial variability, two spectral databases corresponding to the two species were compiled from more than 4,000 spectra (3,520 Arabica and 571 Robusta) of different geographical origins, acquired by diffuse reflectance, corresponding to samples collected over six years.

The NIR calibrations were obtained on samples selected for their spectral representativeness, using partial least squares regression to determine the caffeine, chlorogenic acid, trigonelline, fat, sucrose and dry matter content of green coffees.

Methods

Near infrared spectroscopy

NIR: spectrometer NIRSystem 6500. Foss-Perstorp.

Software: ISI NIRS 2 version 4.11 (InfraSoft International).

Sample presentation: 3 g of homogenized ground coffee were analysed in diffuse reflectance from 400 nm to 2,500 nm (2 nm steps). The analysis was duplicated.

Reference analysis

After extraction and purification, caffeine, trigonelline and sucrose contents were determined by HPLC (UV or electrochemical detector), chlorogenic acid contents by UV spectroscopy and fat and moisture contents by gravimetry.

Near Infrared Spectroscopy: Proceedings of the 11th International Conference.

Results

Principal components analysis

The decision to compile two databases was imposed by the very different biochemical composition of the species, as confirmed by a principal components analysis of the spectra, which led to separation of the two species into two groups (Figure 1). This analysis was based on second derivatives of the spectra (*SNVD* corrected) for wavelengths between 900 and 2500 nm.

The first to fourth PCs explained 42.57%, 26.62%, 20.86% and 3.66% of initial inertia respectively. A factorial discriminant analysis carried out on the basis of these PCs was used to class the samples according to the two varieties, with a 99.9% success rate.

The Mahalanobis *H* distances observed for the Arabica compared to a database compiled with Robusta alone were equal to 5.12 on average, with a maximum value of 11.46 and a minimum of 3.04. Conversely, the H distances observed for the Robusta were 9.65 on average, with a maximum of 23.43 and a minimum of 3.75.

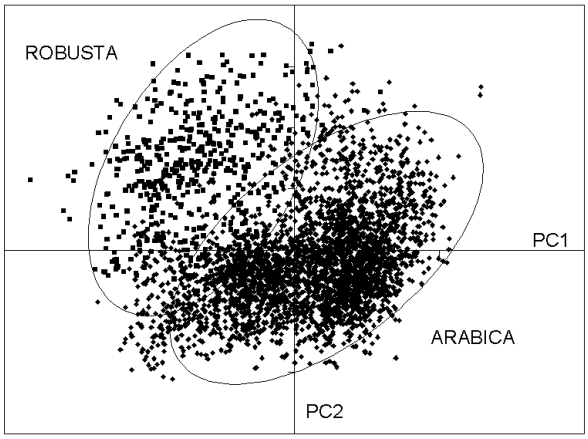

Figure 1. Scatter plot of sample scores for the first two PCs (95% confidence ellipses).

Calibration results

The performances of the predictive models (Table 1) were tested by estimating the standard errors of prediction (*SEP*) on a set of independent samples for each constituent. These *SEP* values were close to the standard errors found in the laboratory (*SEL*) for the determination of caffeine (Figure 2), dry matter and fat (arabica). These equations enabled efficient quantification of these constituents, by routine analysis, based on the spectra.

The prediction equations for sucrose and chlorogenic acid contents were less efficient, but they still made it possible to sort the samples, which is of definite interest for studying cultural practices and for genetic selection. These equations can be improved by analysing further samples, and by improving the precision of the reference method.

A single equation was developed for both species in the case of trigonelline, as the distribution (Figure 3) of contents for this constituent was Gaussian and unimodal. The *SEP* value estimated from 30 independent samples (Figure 3) was 0.054, the regression slope was equal to 1.09, the R^2 and bias

were 0.95 and 0.004, respectively. Established in this way, the equation offers better calibration (*SEC* = 0.04, 1-*VR* = 0.82) and prediction performances than the equations developed separately for each species (Table 1).

Table 1. Calibration statistics for the two species for each constituent.

	Constituent	*N*	M	*SD*	*SEC*	R^2	*SECV*	1-*VR*	*SEP*	*SEL*	No of terms PLS
Arabica	Dry matter	456	89.69	1.67	0.10	0.99	0.12	0.99	0.13	0.10	14
	Caffeine	361	2.25	0.16	0.05	0.88	0.06	0.85	0.06	0.06	4
	Trigonelline	323	0.99	0.09	0.04	0.82	0.05	0.72	0.05	0.03	13
	Fat	177	14.31	1.71	0.45	0.93	0.49	0.92	0.54	0.50	5
	Sucrose	331	7.39	0.94	0.43	0.79	0.53	0.68	0. 50	0.40	12
	Chlorogenic acids	267	7.64	0.80	0.35	0.81	0.39	0.76	0.39	0.30	8
Robusta	Dry matter	259	92.43	2.37	0.09	0.99	0.12	0.99	0.13	0.10	15
	Caffeine	330	2.56	0.43	0.08	0.96	0.08	0.96	0.08	0.06	8
	Trigonelline	200	0.79	0.09	0.05	0.62	0.06	0.57	0.06	0.03	6
	Fat	43	10.32	1.18	0.55	0.78	0.69	0.66	*	0.50	3
	Sucrose	108	4.87	0.89	0.34	0.85	0.47	0.72	0.48	0.40	8
	Chlorogenic acids	169	11.69	1.12	0.46	0.85	0.53	0.80	0.53	0.30	7
Arabica Robusta	Trigonelline	520	0.92	0.13	0.05	0.85	0.05	0.82	0.05	0.03	11

N: total number of samples used for computation; M: mean; SD: standard deviation of the concentration values; *SEC*: standard error of calibration; R^2: coefficient of multiple determination; SECV: standard error of cross validation; *SEP*: standard error of prediction; *SEL*: standard error of laboratory; 1-*VR*: estimate of the fraction of explained variance.
not enough samples to estimate *SEP*.

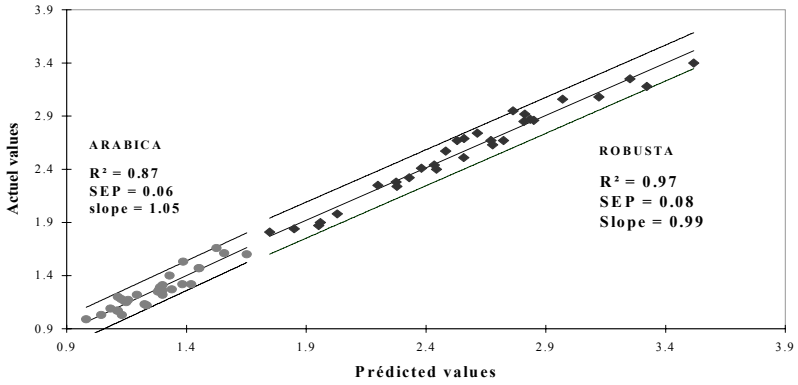

Figure 2. Correlation plot between actual reference values for caffeine and NIRS predicted values obtained on a set of 30 independent samples for each species (confidence interval 95%).

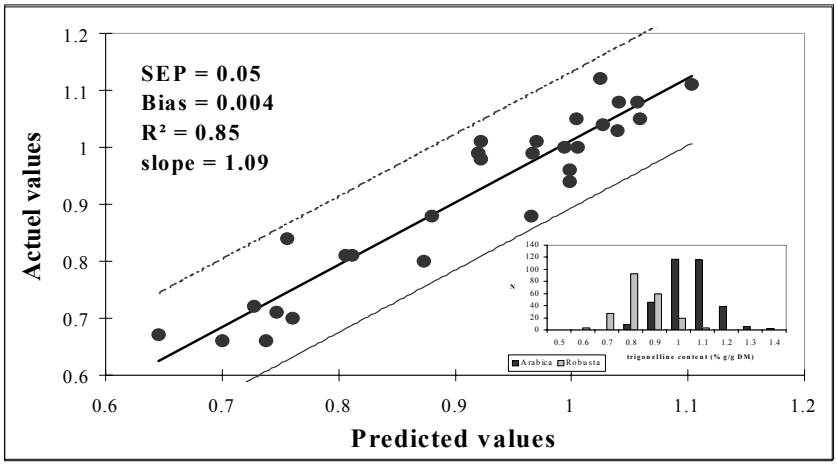

Figure 3. Correlation plot between actual reference values for trigonelline and NIRS predicted values obtained on a set of 30 independent samples (confidence interval 95%). Histogram of trigonelline content for all the samples.

Calibration transfer

The spectral databases were successfully transferred to a second monochromator, applying the multiple standardisation procedure patented by J.S Shenk and M.O Westerhaus, from a set of 30 crushed and dried samples packed in sealed sample cups.

The H distances measured on 70 samples of varied genotypes and origins acquired with the second instrument were equal to 6.50 on average before standardization and 2.8 after standardization when compared to the spectral databases developed on instrument 1.

The equations were upgraded by integrating 32 new spectra and references (caffeine, chlorogenic acids, sucrose and dry matter), and by using a repeatability file compiled from 30 green coffee spectra acquired in duplicate on each instrument. The performances of the equations upgraded in this way were similar to the original calibrations.

Conclusion

- This result made it possible to apply the appropriate calibration equation for each species, based on Mahalanobis distances, in routine analysis. The databases, compiled in this way, have been optimized in terms of prediction accuracy and spectral representation and now only 2% of green coffee samples analysed are outliers from the base.
- It is essential to improve the predictive models for certain constituents, and this means acquiring new samples and working in parallel on the reference methods.
- Optimisation of the database needs to take into account the spectral variability of coffees from emerging markets (organic coffees, *terroir* coffees) and predictive models need to be developed for whole beans.