ELSEVIER

0308-8146(95)00030-5

Food Chemistry **54** (1995) 321–326

0308-8146/95/$9.50

Analytical, Nutritional and Clinical Methods Section

Discrimination between *Coffea arabica* and *Coffea canephora* variant *robusta* beans using infrared spectroscopy

E. K. Kemsley,* S. Ruault & R. H. Wilson

Institute of Food Research, Norwich Laboratory, Norwich Research Park, Colney, Norwich NR4 7UA, UK

(Received 7 September 1994; revised version received and accepted 3 January 1995)

The seed or 'bean' of the coffee plant is an important crop, grown commercially across the world. Two species are commonly cultivated: *Coffea arabica* and *Coffea canephora* variant *robusta*. Analytical techniques for species identification, in particular of coffee products such as ground or 'instant' coffees, are of great importance. In this paper, mid-infrared spectroscopy is proposed as a rapid alternative to existing authentication methods, which are often time-consuming or difficult to implement successfully. A Fourier-transform infrared spectrometer is used for this work, equipped with a diffuse reflectance accessory. Statistical procedures comprising principal components analysis and classical discriminant analysis are applied to spectra of ground roast *arabica* and *robusta* beans, and results presented which demonstrate that the species of such samples can readily be identified.

INTRODUCTION

The coffee bean is obtained from the fruit of the coffee plant, a small evergreen shrub belonging to the genus *Coffea*, family Rubiaceae. Two species of *Coffea* have acquired worldwide economic importance: *arabica* (approximately 90% of world coffee production), and *canephora* variant *robusta* (approximately 9% of production). *Arabica* beans are valued the most highly by the trade, as they are considered to have a finer flavour than *robusta*. It is therefore important that the species of raw beans and of various coffee products can be identified. A trained inspector can easily distinguish raw *arabica* and *robusta* beans from differences in size and colour. Unfortunately, these visual indicators are eliminated by the roasting and milling processes, so that identification of ground roast and instant coffees, both of which are dark-brown powders, requires an alternative method.

Efforts have been made to characterise the two coffee species using chemical data, with some success (Clifford, 1985, 1986). Speer *et al.* (1991) used high-performance liquid chromatography (HPLC) to detect diterpene-16-*O*-methylcafestol, a compound present only in *robusta* coffees. This substance remains stable during the roasting process, so that it is a useful indicator of the presence of *robusta* beans in roast, ground or instant coffee products. However, a disadvantage of wet chemical analysis is that it is time-consuming. Sample preparation is tedious and can be difficult, involving extraction, dissolution and dilution steps. There is a need to find simple, fast and reliable methods for tackling food authentication problems. One such technique is mid-infrared spectroscopy, which has been shown to be useful for the rapid, non-invasive analysis of a wide range of foodstuffs (Wilson & Goodfellow, 1994). Reported applications encompass both qualitative and quantitative work, with perhaps the most effort expended in the area of compositional analysis.

In recent years, the rapid increase in affordable computing power has made multivariate statistical methods readily available to the spectroscopist, and these are now being used in combination with infrared spectroscopy to address increasingly complex tasks, such as classification and authentication problems. Reported work in these areas includes the discrimination between edible oils (Lai *et al.*, 1994), between cell walls from different plant species (Kemsley *et al.*, 1994) and between fruit purees of different types (Defernez *et al.*, 1995). The methodologies described in these papers are broadly similar. Infrared spectra are collected using a Fourier-transform infrared (FTIR) spectrometer, equipped with a sampling accessory appropriate for the sample morphology. Sample preparation is in general kept to a minimum. Next, data processing comprising

* To whom correspondence should be addressed.

two distinct stages is carried out: firstly, the 'data compression' step of principal components analysis (PCA), which removes redundancy in the original spectra leaving a reduced, simplified data set; secondly, discriminant analysis (DA), which uses the compressed data and known classifications to create a set of 'class means', then reclassifies existing observations to the nearest class mean in order to estimate a likely success rate for the assignment of future, unknown observations. There are many variations on this basic DA theme; a range of common methods is discussed in the book by Massart *et al.* (1988). A full description of PCA and its application can be found in the work by Jolliffe (1986).

In the present paper, the data analysis procedure outlined above is applied to infrared spectra of ground roast coffee beans, with the aim of discriminating between the *arabica* and *robusta* species. Convincing results are presented indicating that mid-infrared spectroscopy is able to distinguish between pure samples of each species, and may offer the potential for adulteration detection in the future. A conventional FTIR technique for sampling powders, diffuse reflectance (DRIFT) (Wilson & Goodfellow, 1994), was employed for spectral acquisition. To obtain a DRIFT spectrum, a powdered or ground sample is placed into a stainless-steel cup (approximately 10 mm diameter), and the sample surface flattened. The cup is placed in a DRIFT accessory, which incorporates suitable mirrors to steer the infrared beam onto the sample, to collect the portion that is diffusely reflected from the sample surface and to direct it onto the detector to record the spectrum. For certain samples, dilution in another matrix (for example, powdered infrared grade potassium bromide (KBr)) is essential to avoid optical distortion effects in the spectra. Nevertheless, sample preparation remains relatively straightforward, and FTIR combined with DRIFT can be regarded as a rapid analytical technique.

MATERIALS AND METHODS

Twenty-eight samples of whole roast coffee beans were obtained for this work; the supplier was able to guarantee their authenticity. Twenty of the samples were *arabica* beans, eight *robusta*. The *arabica* beans originated from eight different countries: Ethiopia, Brazil, Zaire, Zimbabwe, Kenya, Honduras, Costa Rica and Columbia. The *robusta* beans originated from Indonesia, Uganda, Thailand, Vietnam, Zaire and Ghana.

From each sample, approximately 15 beans were selected at random, and ground in a Krups coffee grinder. Equal quantities of the ground coffee and of infrared grade KBr were weighed into a mortar, and pounded with a pestle to reduce the mixture to a very fine powder. Infrared spectra were recorded of all 28 samples. In addition, one *arabica* mixture was selected at random and spectra obtained of five further sample cup re-loadings.

All spectra were collected using a Monit-IR (Spectra-Tech, Applied Systems Inc.) FTIR spectrometer operating in the region 800–4000 cm^{-1}, equipped with a sealed, desiccated interferometer compartment, a deuterated triglycine sulphate detector and a permanently mounted DRIFT accessory, which incorporated windows to minimise exposure of the infrared beam to the atmosphere. All spectral measurements were made at nominal 8 cm^{-1} resolution, with 64 interferograms co-added before Fourier transformation, zero-filled to give a data point spacing of ~4 cm^{-1} in the frequency domain.

All sample single-beam spectra were transformed to Kubelka–Munk units using a background spectrum of ground KBr, and truncated to 286 data points in the region 800–1900 cm^{-1}. To reduce the effect of irreproducible sample cup loading, a single-point baseline correction at 1900 cm^{-1} was performed, followed by normalisation on the integrated spectral area. The pre-treated data was processed using PCA and DA procedures. All data processing was carried out using 'Win-Discrim' (E. K. Kemsley, Institute of Food Research, Norwich), a specialised package for discriminant analysis.

RESULTS AND DISCUSSION

Typical spectra of ground *arabica* and *robusta* beans are shown in Fig. 1. The spectral quality is high: examination of the baseline in the region 1800–2400 cm^{-1} shows that detector noise is negligible. Another common source of spectral contamination is also noticeably absent: water vapour bands in the region 1500–1800 cm^{-1}. Thus, the sealed, desiccated optical bench combined with the permanently mounted DRIFT accessory offered high performance, whilst avoiding the lengthy purge times of traditional research-grade instruments. All spectra were truncated to 286 points in the region 800–1900 cm^{-1} to accommodate computer memory limitations and to avoid the inclusion of irrelevant data, which can degrade the results of subsequent analyses.

There is another source of spectral variability which is less a function of instrumental parameters, but rather an intrinsic difficulty associated with the DRIFT sampling technique: irreproducible re-loading of the sample cup. Grinding the sample to a fine powder and smoothing the sample surface in the same direction upon each loading of the cup mitigate the problem to an extent; nevertheless, the number and orientation of sample particles in the infrared beam will inevitably vary, and in turn lead to variation in the quantity of diffusely reflected radiation. This variability manifests itself as differences, not in relative intensities, but rather in overall spectral response. Baseline-correction followed by normalisation on the integrated spectral area helps remove this unwanted variability. To illustrate this effect, the replicate spectra of the randomly chosen *arabica* sample are shown before and after pre-treatment in Fig. 2.

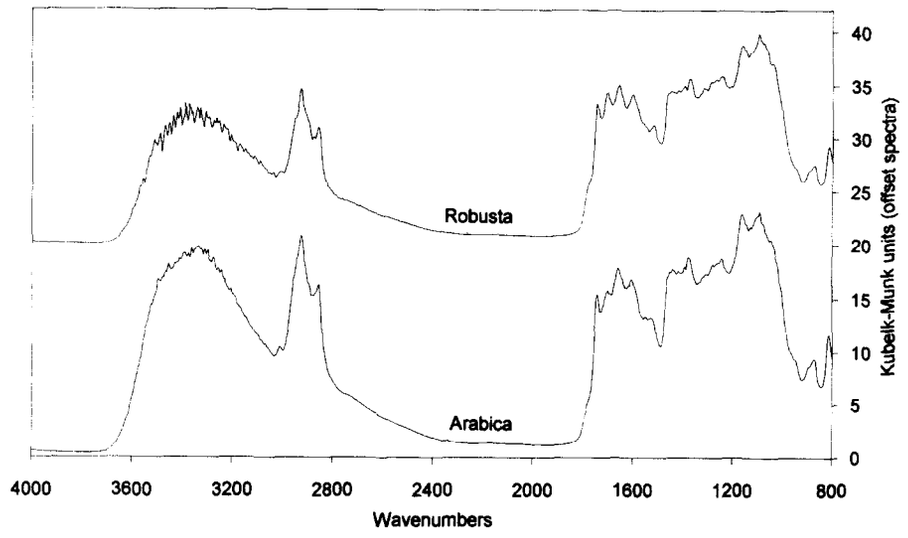

Fig. 1. Raw spectra of *arabica* and *robusta* ground coffees.

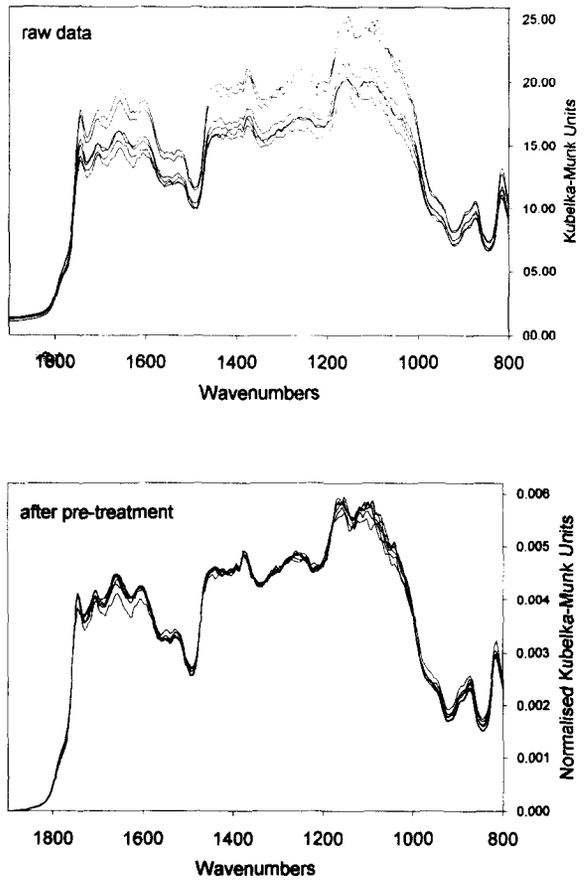

Fig. 2. Spectra of repeated sample re-loadings of single *arabica* sample, before and after baseline correction and area-normalisation.

A full assignment of the spectral bands is a challenging problem, and will not be attempted in this work. The composition of ground coffee is highly complex; in addition to a few readily identified major constituents, there are many more minor compounds, not all of which have yet been elucidated (see Clifford, 1985), but which will doubtless contribute to the infrared spectrum. Carbohydrates are the bulk constituent of roast coffee, largely present as complex polysaccharides; the simple sugars found in green coffee beans are largely consumed during roasting, with only small quantities of glucose and fructose surviving. Complete spectral assignment of such bio-polymers is notoriously difficult, and has been achieved for only very few samples. However, carbohydrates generally exhibit large features in the so-called 'fingerprint' region (900–1400 cm^{-1}), and are probably responsible for the majority of bands in the roast coffee spectrum. Other major compounds include proteins and, in particular, lipids. Green *arabica* beans contain 14–18% (w/w) crude lipid, whereas green *robusta* beans contain 9–12%. A loss of 1–2% occurs on roasting. The bulk of the crude lipid is a typical seed oil: C_{16} and $C_{18:2}$ are the dominant fatty acids, each present at around 35–40%. The balance of the seed oil consists of unsaponifiable lipids, including hydrocarbons, tocopherols, pigments, phospholipid, sterols and diterpenes. Lipids in general exhibit a characteristic band arising from the carbonyl (C=O) vibration centred at ~1744 cm^{-1}, which can be identified in both the *arabica* and *robusta* spectra. Variations in the lipid composition can subtly affect the spectral shape in and around this feature.

Pre-treated spectra of the 28 samples are shown in Fig. 3. To aid clarity, the *arabica* and *robusta* groups are offset. There is substantial within-species variation, but overall the two groups appear quite similar. The most noticeable differences occur in the bands centred at ~1744 cm^{-1} and ~1150 cm^{-1}, which tend to be relatively larger in the *arabica* spectra. The former observation can be explained by the known difference in lipid content of the two species; the latter could be due to differences in the polysaccharide composition. However, not all the *arabicas* fit this pattern, and identifying the species of individual spectra using visual inspection alone is not wholly satisfactory.

PCA was applied to the 28 spectra data set. This is an essential first step, since multivariate statistical procedures cannot be employed whilst the number of

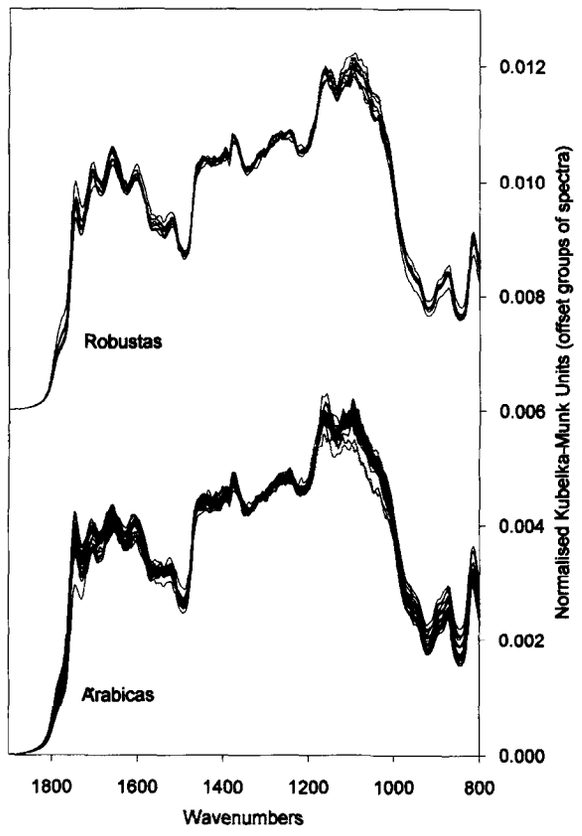

Fig. 3. Pre-treated spectra of *arabica* and *robusta* ground coffees.

spectral data points (or 'variates') exceeds the number of spectra (or 'observations'). The variates in the original data set will normally be correlated with one another, to a greater or lesser extent; PCA removes this redundancy by a linear transformation of the original data to a set of new, uncorrelated variates, termed the principal component (PC) scores. In doing so, a re-arrangement takes place, such that only the first few PC scores are required to describe most of the information contained in the many original variates. The number of significant PC scores is always less than the number of observations, and multivariate methods can proceed. A further advantage of the reduction in variates is the resultant simplification of the data set, enabling easier visualisation of relationships within the data, for example by plotting pairs of PC scores against one another.

The percentage (%) variance and cumulative % variance accounted for by each of the first 10 PC scores are presented in Table 1. The cumulative % variance rises fairly slowly. Generally, this indicates that there are a large number of independent sources of variation in the data, which is consistent with the chemical complexity of the coffee bean. Plots for the first few scores against one another were examined. Considerable grouping of the data was found in up to the first three PC scores. An example 2D plot, of the first against third PC, is shown in Fig. 4. Some division according to species is evident, but perhaps a better impression of the separation of the two groups in 'PC-space' is obtained from examining a 3D plot of the first three PC scores (Fig. 5). The overlap present in the 2D plots is now eliminated and the two classes can be entirely spatially separated. In view of these findings, the first 3 PC scores were used to perform a classical DA, using the Mahalanobis D^2 metric (Mardia, 1977). As expected from the encouraging 3D scores plot, all 28 spectra were correctly classified at the re-assignment stage: the results are presented graphically in Fig. 6. Also marked on Fig. 6 are the D^2 values, measured in the same PC-space, for the remaining replicates of the *arabica* sample. All are correctly classified, although this is not an especially useful finding, since these are not truly independent samples. However, an indication of the expected reproducibility of the method is obtained.

Table 1. Percentage variance and cumulative percentage variance for the first 10 PC scores

PC score	% Variance	Cumulative % variance
1	42·0	42·0
2	20·7	62·7
3	16·4	79·1
4	5·5	84·6
5	3·7	88·3
6	2·0	90·2
7	1·6	91·8
8	1·4	93·2
9	1·3	94·5
10	1·2	95·7

Having achieved a successful DA, one may now speculate on the basis of this discrimination. Figures 4 and 5 show that the first three PC scores all play a role in distinguishing the two species. Therefore, examination of the PC loadings (sometimes known as 'factor spectra') may be worthwhile. Since the PCA transformation reported in this work employed the correlation matrix method (see Jolliffe, 1986), the loadings resemble variance-scaled spectra and can be somewhat difficult to interpret. In order to assist interpretation, an 'inverse variance-scaling' process was employed, consisting of

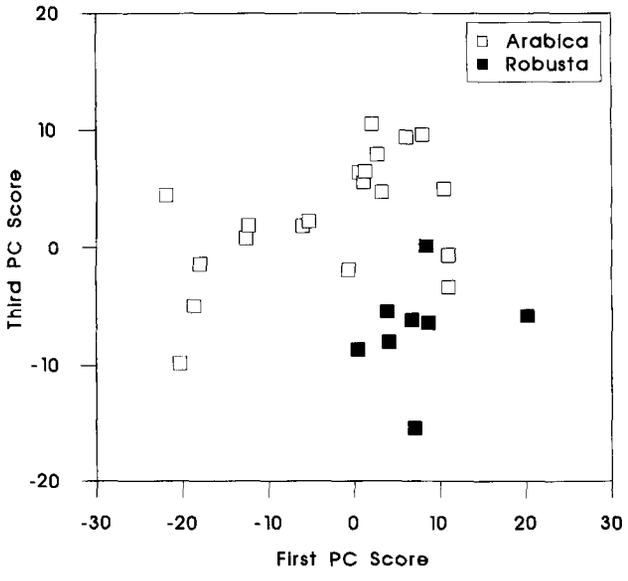

Fig. 4. Plot of first versus third PC scores.

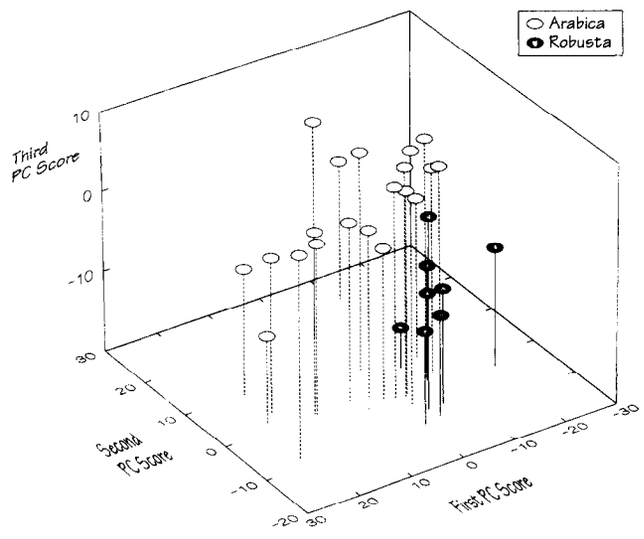

Fig. 5. 3D plot of first versus second versus third PC scores.

multiplying the loadings by the data standard deviation. This gives the loadings a more familiar appearance, allowing specific spectral features to be identified. The first three loadings are shown, offset and inverse variance-scaled, in Fig. 7. The first loading is relatively featureless, apart from a fairly broad band at ~1100 cm^{-1}, perhaps indicating that it represents largely a baseline or overall intensity shift not completely removed by the data pre-treatment. The most pronounced features are in the second and third loadings, in particular a negative feature again centred at ~1100 cm^{-1}, and a sharper band centred at ~1744 cm^{-1}. The former occurs in the spectral region associated strongly with carbohydrates, and could be due to differences in the polysaccharide composition. It has been suggested (Clifford, 1985) that such differences may arise as the result of different behaviour during roasting, although not all researchers share this view (Bradbury, 1987). The feature at 1744 cm^{-1} can be identified with a similar band in the

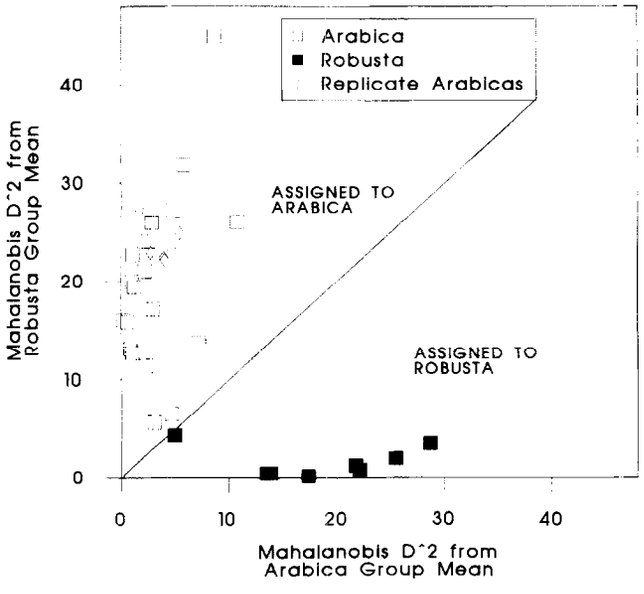

Fig. 6. Plot of Mahalanobis D^2 values in 3D PC-space of each sample from *Arabica* and *Robusta* group means.

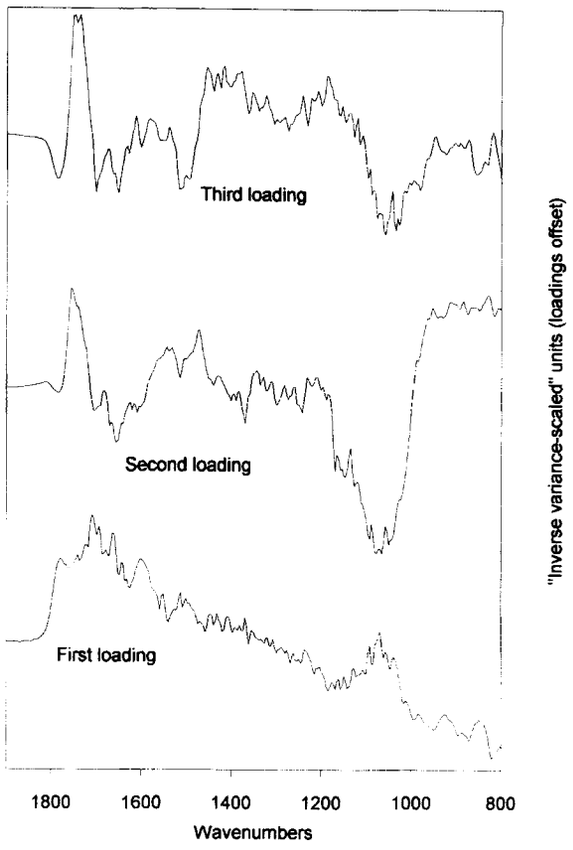

Fig. 7. First three PC loadings, offset and 'inverse variance-scaled' to aid interpretation.

original spectral data, where it was found to be relatively more intense in *arabica* than *robusta*, and arises from the lipid content of the beans. Since the PC scores represent the relative importance of the corresponding loadings in explaining the original data, it is not surprising that the second and third scores are also larger for *arabica* compared with the *robusta*. We conclude therefore that the basis of the discrimination is due at least in part to differences in the lipid composition of the two species, although the minor structure in the loadings suggests that this is not the whole story, and a full interpretation remains a complex problem.

CONCLUSIONS

This work has demonstrated that FTIR spectroscopy combined with the DRIFT sampling technique may be used for the species discrimination of *arabica* and *robusta* ground roast coffees. Minimal sample preparation, and a sealed desiccated spectrometer requiring no lengthy purge time, mean that analysis is extremely rapid in comparison with wet chemical methods. Data pre-treatment minimised the sampling irreproducibility. Data processing comprised PCA, which revealed clear grouping of the spectra according to species, followed by classical DA based on the PC scores, which yielded 100% successful discrimination.

The only qualifier to this success is the relatively small sample size (20 *arabicas*, 8 *robustas*). Clearly, work of this kind is of limited value unless certifiably authentic samples are used, and such samples are hard to obtain in any quantity. However, further samples are likely to become available to us in the near future, which will be used to extend the database. In addition to identifying pure samples of each species, future work will explore whether it is possible to detect blends of *robusta* and *arabica*, as this may offer a new, fast method for the detection of adulteration.

ACKNOWLEDGEMENT

The authors would like to thank Mr G. Downey, of Teagasc, National Food Centre, Dublin, Ireland for supplying the samples used in this work.

REFERENCES

Bradbury, A. (1987). *Proceedings of Douzieme Colloque International sur la Chimie des Cafes Verts, Terrefies et Leurs Derives.* Association Scientifique Internationale du Cafe, Montreux, France.

Clifford, M. N. (1985). Chemical and physical aspects of green coffee and coffee products. In *Coffee: Botany, Biochemistry and Production of Beans and Beverage*, ed. M. N. Clifford & K. C. Wilson. Croom Helm, London, UK pp. 305–74.

Clifford, M. N. (1986). Physical properties of the coffee bean. *Tea Coffee Trade J.*, **May**, 30–3.

Defernez, M., Kemsley, E. K. & Wilson, R. H. (1995). The use of infrared spectroscopy and chemometrics for the authentication of fruit purees. *J. Agric. Food Chem.*, **43**, 109–13.

Jolliffe, I. T. (1986). *Principal Components Analysis.* Springer-Verlag New York, USA.

Kemsley, E. K., Belton, P. S., McCann, M. C., Ttofis, S., Wilson, R. H. & Delgadillo, I. (1994). A spectroscopic method for the authentication of vegetable matter. *Food Control.*, **5**(4), 241–3.

Lai, Y. W., Kemsley, E. K. & Wilson, R. H. (1994). The potential of FTIR spectroscopy for the authentication of vegetable oils. *J. Agric. Food Chem.*, **42**, 1154–9.

Mardia, K. V. (1977). Mahalanobis distances and angles. In *Multivariate Analysis* (ed. P. R. Krishnaiah, North-Holland, Amsterdam, The Netherlands, pp. 495–511.

Massart D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y. & Kaufman L. (1988). *Data handling in Science and Technology* (Vol. 2: *Chemometrics, a Textbook*), ed. B. G. M. Vandeginste & L. Kaufman. Elsevier Science Publishers, Amsterdam, The Netherlands.

Speer, K., Tewis, R. & Montag, A (1991). *Quatorzieme Colloque Scientifique International sur le Cafe,* Proceedings of an ASIC Conference, San Francisco, CA, USA, 14–19 July, pp. 237–44.

Wilson, R. H. & Goodfellow, B. J. (1994). Infrared spectroscopy. In *Spectroscopic Techniques for Food Analysis.* VCH, New York, USA, Chapter 3.