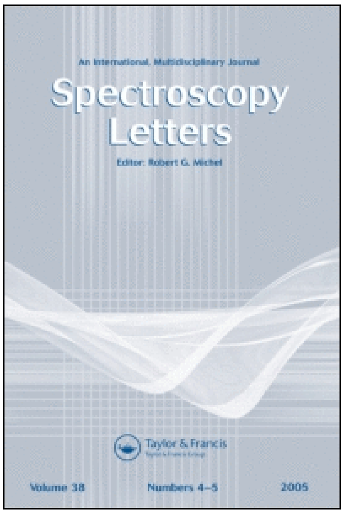

Spectroscopy Letters
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597299

Multiple Mathematical Corrections Before Quantitative Analysis of Biological Infrared Spectra

Fréedéric Cadet[a]
[a] Laboratoire de Biochimie Faculté des Sciences, Université de la Réunion, Réunion, France-Dom

To cite this Article Cadet, Fréedéric(1996) 'Multiple Mathematical Corrections Before Quantitative Analysis of Biological Infrared Spectra', Spectroscopy Letters, 29: 7, 1335 — 1351

To link to this Article: DOI: 10.1080/00387019608007127

URL: http://dx.doi.org/10.1080/00387019608007127

SPECTROSCOPY LETTERS, 29(7), 1335-1351 (1996)

MULTIPLE MATHEMATICAL CORRECTIONS BEFORE QUANTITATIVE ANALYSIS OF BIOLOGICAL INFRARED SPECTRA

Key words: Biological Mid-FTIR spectra, quantification, spectral correction.

Frédéric CADET

Laboratoire de Biochimie, Faculté des Sciences, Université de la Réunion,
15 avenue, René Cassin, BP 7151, 97715 Saint-Denis Messag
Cedex 9, Réunion, France-Dom.
Fax: Intl (262) 93 81 66.

Corresponding author:

Frédéric CADET
Laboratoire de Biochimie
Faculté des Sciences
Université de la Réunion
BP 7151, avenue René CAssin
97715 Saint Denis Messag Cedex 9
La Réunion
France-DOM

ABSTRACT

Infrared reflectance spectroscopy is often used to predict the composition of various food products. However the precision of the analysis could be improved. For this purpose, several treatements can be used ; (i) baseline correction. Baseline deformation is one of the major problem that is associated with infrared spectra; (ii) spectra can be expanded in terms of a set of orthonormal polynomials derived from the Legendre polynomials, and the leading terms of the expansion, which contains most of the baseline variation, can be removed; (iii) the classification of samples prior to analysis is one way to obtain more homogenuous families. Hence, with the aim of improving the precision of the values obtained during quantitative determination of constituants from biological samples, we have sucessively applied different mathematical treatments on their infrared spectra : baseline correction, Legendre polynomials decomposition and classification procedure. The application of these mathematical treatments improves successively the precision of the predicted sucrose content values in biological samples. The mean and standard deviation of the differences between predicted and reference values are : $1.52x10^{-2}$ and $3.17x10^{-1}$ respectively before correction and (i) after the baseline correction : $5.06x10^{-2}$ and. $3.08x0^{-1}$ % respectively, (ii) after decomposition Legendre polynomilas : $-6.13x10^{-3}$ and $2.76x10^{-1}$ % respectively, (iii) and after classification procedure according to correlation distance : $-1.00x10^{-3}$ and $2.65x10^{-1}$ % respectively. Surprisingly, when only classification procedure is used for correction, the mean and SD values are $7.43x10^{-2}$ and $2.62x10^{-1}$. Hence, it appears that the successive mathematical treatments can be advantageously substituted by a single classification procedure.

INTRODUCTION

Infrared reflectance spectroscopy is more and more used in food and agricultural industry for the quantitative determination of major biochemical constituants (Osborne and Fearn, 1986, Williams and Norris, 1987). With the recent developments in the field of computer and software engineering and the advent of new techniques such as ATR (Attenuated Total Reflectance) cells, the development and application of physical quantitative analysis methods such as Mid infrared spectroscopy have considerably expanded (Crocombe *et al.*, 1987; Van de Voort and Ismail, 1991). Principal Component Analysis showed to be very useful for the study of MIR spectra (Gillete and Koenig, 1982, Bertrand *et al.*, 1984, Cowe and Nocol., 1985).

We have previously reported the use of multidimensional statistical analysis combined with Mid-FTIR spectroscopy for the prediction of sucrose content sugar cane juices (Cadet *et al.*, 1991, Cadet and Offmann., 1995). However the precision of the analysis could be improved.

With the aim of improving the precision of the values obtained during quantitative determination of constituants from biological samples, we have directly or successively applied different mathematical treatments on their infrared spectra : baseline correction (Cadet and Offmann, 1996), decomposition Legendre polynomials (Cadet, 1996a) and classification procedure (Cadet, 1996b).

MATERIALS AND METHODS

Biological samples.

Raw sugar cane juices are obtained as previously described (Cadet and Offmann, 1995). The calibration set is constituted of 107 samples and the verification set of 30 samples.

Mid-FTIR ATR spectra

Mid-Fourier Transform Infrared (Mid-FTIR) spectra were collected on a Michelson-100 Fourier transform spectrophotometer.The data were recorded from 5000 to 500 cm^{-1} in 4 cm^{-1} increments at log(1/R), in which R is the ratio of the reflected intensity for the background to that of the sample. Attenuated total reflectance spectra were obtained with a Specac Overhead Attenuated Total Reflectance system. The crystal of the reflectance element is made from zinc selenide, a material that is quite inert to water; it is quite rapidly cleaned between samples by spraying with water and then dried with filter paper.

Mathematical treatment

All the mathematical treatments were performed on a Compaq personal computer with software written in "C" language and developped in our laboratory.

Multidimensional statistical analyses, such as principal component analyses (PCA), describe variation in multidimensional data by few synthetic variables. These synthetic variables are linear combination of all the original variables and have the advantage of having no correlation with each other. Simpler descriptions of data sets are thus obtained with minimal loss of information. These treatments were used for morphological analysis of spectra (le Nouvel., 1981) and for graphical representation of spectra similarity (Devaux *et al.*, 1988).

PCA was applied to the spectra from 800 to 1240 cm^{-1} (with 230 data points used as principal variables). Spectra were centered prior to PCA according to :

$$X_{ij} = A_{ij} - A_j - A_i + A$$

where X_{ij} = centered data ; A_{ij} = spectral data (log 1/R) of spectrum i and wavelength j; A_j = mean value of spectral data at wavelength j for every spectrum; A_i = mean value of spectral data of spectrum i for every wavelength; and A = average mean of all spectral data in the collection.

Principal component regression (PCR) was used to establish a prediction equation. PCR is basically a multilinear regression applied to scores assessed by PCA (Lebart *et al.*, 1977). Interest in the introduction of scores according to their predictive ability had already been shown (Bertrand *et al.*, 1987, Cadet & al., 1995).

Concentrations are predicted according to :

$$C_{n.1} = X_{n.k} \cdot V_{k.p} \cdot R_{p.1}$$

where C is the column vector of predicted concentrations, X is the centered matrix of spectral data, V is the matrix of latent vectors of PCA, and R is the column vector of the regression coefficients of the prediction equations. n, k, p are respectively the number of samples; the number of wavelengths; the number of significant principal components.

RESULTS AND DISCUSSION

Reference values and spectra.

The spectrum between 3800 and 700 cm-1 of a sample features three major absorption zones : 3700-2800 cm-1, 1800-1470 cm-1 and 1240-800 cm-1 (Figure 1).

The first two zones can be assigned mainly to water. The 1240-800 cm-1 zone features the characteristic C-O, O-H and C-C vibrational bands of saccharides. In the spectrum, the

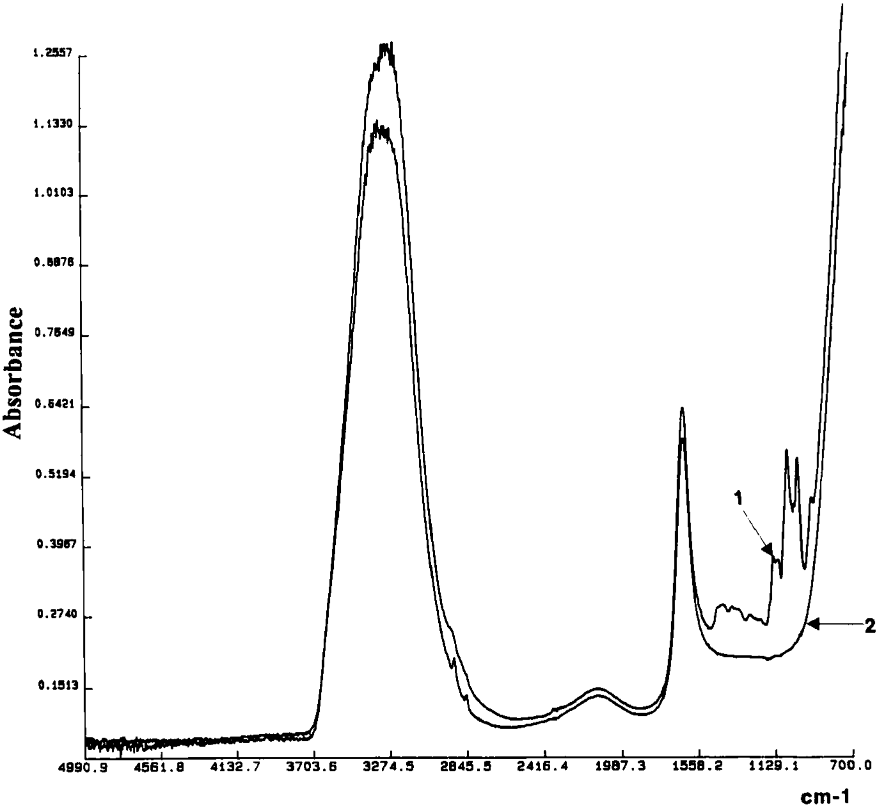

Figure 1. Mid-FTIR spectra of a biological sample (1) and of water (2).

1240-800 cm-1 zone however corresponds to the absorption zones of the three major constituants of sugar cane juice (sucrose, glucose and fructose). The reference values associated with each of the spectrum of the verification set are given in Table 1.

Baseline correction

The deformation of baselines (Powell, 1984, Hruschka, 1987) is one of the problem that face infrared spectra. In MIR

Table 1. Difference between the reference and the predicted values without correction and after simple or successives corrections: baseline correction, decomposition Legendre polynomials, classification procedure.

SAMPLES NUMBER				Mathematical Treatments									
	Initial Spectra			Baseline correction (1)		(1) + decomposition Legendre polynomials (2)		(1) + (2) + classification			Direct classification		
	Référence (a)	Predicted (b)	Deviation (b-a)	Predicted (c)	Deviation (c-a)	predicted (d)	Deviation (d-a)	Class	Predicted (e)	deviation (e-a)	Class	Predicted (f)	deviation (f-a)
1	21.029	21.195	0.166	21.201	0.172	21.108	0.079	2	21.027	-0.002	2	20.935	-0.094
2	22.144	22.092	-0.052	22.169	0.025	21.997	-0.147	2	21.95	-0.194	2	22.29	0.146
3	21.487	21.553	0.066	21.554	0.067	21.307	-0.18	2	21.098	-0.389	2	21.446	-0.041
4	21.699	21.649	-0.05	21.697	-0.002	21.798	0.099	2	21.913	0.214	2	21.755	0.056
5	23.035	23.419	0.384	23.313	0.278	23.299	0.264	2	22.986	-0.049	2	23.47	0.435
6	21.197	21.261	0.064	21.359	0.162	20.896	-0.301	2	21.136	-0.061	2	21.204	0.007
7	21.581	21.996	0.415	21.857	0.276	21.84	0.259	2	22.261	0.68	2	22.034	0.453
8	24.273	24.028	-0.245	24.086	-0.187	24.151	-0.122	2	23.791	-0.482	2	24.672	0.399
9	19.793	20.43	0.637	20.316	0.523	20.325	0.532	3	20.195	0.402	4	20.288	0.495
10	19.482	19.519	0.037	19.484	0.002	19.591	0.109	3	19.469	-0.013	4	19.63	0.148
11	21.187	21.135	-0.052	21.179	-0.008	20.997	-0.19	3	21.044	-0.143	2	20.827	-0.36
12	21.137	21.006	-0.131	20.961	-0.176	21.305	0.168	3	21.17	0.033	2	21.125	-0.012
13	20.171	20.028	-0.143	19.942	-0.229	19.972	-0.199	3	20.133	-0.038	4	19.889	-0.282
14	18.763	19.111	0.348	19.104	0.341	19.193	0.43	3	18.886	0.123	4	18.947	0.184
15	20.029	19.991	-0.038	20.146	0.117	20.058	0.029	3	19.98	-0.049	2	20.228	0.199
16	18.838	19.03	0.192	19.062	0.224	18.981	0.143	3	18.895	0.057	4	18.803	-0.035
17	19.373	19.787	0.414	19.537	0.164	19.497	0.124	3	19.56	0.187	4	19.489	0.116
18	21.017	21.37	0.353	21.211	0.194	21.489	0.472	3	21.256	0.239	2	21.259	0.242
19	20.804	21.019	0.215	20.841	0.037	20.842	0.038	3	20.872	0.068	2	21.132	0.328
20	20.884	20.478	-0.406	20.508	-0.376	20.709	-0.175	3	20.705	-0.179	2	20.665	-0.219
21	20.16	20.41	0.25	20.11	-0.05	20.163	0.003	3	20.454	0.294	4	20.37	0.21
22	19.48	18.917	-0.563	18.86	-0.62	18.872	-0.608	3	18.994	-0.486	4	19.293	-0.187
23	20.223	19.924	-0.299	19.94	-0.283	19.976	-0.247	3	19.885	-0.338	4	20.294	0.071
24	20.281	20.434	0.153	20.208	-0.073	20.569	0.288	3	20.27	-0.011	2	20.381	0.1
25	19.781	19.4	-0.381	19.354	-0.427	19.436	-0.345	3	19.541	-0.24	4	19.48	-0.301
26	20.851	21.045	0.194	20.909	0.058	21.031	0.18	3	20.909	0.058	2	21.173	0.322
27	18.169	17.913	-0.256	17.789	-0.38	17.977	-0.192	4	18.327	0.158	4	18.193	0.024
28	17.04	16.238	-0.802	16.068	-0.972	16.491	-0.549	4	16.723	-0.317	4	16.41	-0.63
29	18.311	18.238	-0.073	18.188	-0.123	18.27	-0.041	4	18.401	0.09	4	18.568	0.257
30	17.588	17.648	0.06	17.336	-0.252	17.483	-0.105	4	17.946	0.358	4	17.787	0.199
	Mean		1.52E-02		-5.06E-02		-6.13E-03			-1.80E-03			7.43E-02
	Standard deviation		3.17E-01		3.08E-01		2.76E-01			2.65E-01			2.62E-01

spectra, uncertain variations are observed and these add an arbitrary constant to every absorbance. We have shown that baseline correction by total translation with respect to an invariable zone improves sensibly the precision of the quantiative predictions. We have extensively described this method previously (Cadet and Offmann, 1996).

From the spectra constitutive of the calibration set, the mean absorbance and the standard deviation for each and every point (wavelength) have been calculated. The diagram obtained (data not shown) features the zones at which the standard deviation (SD) value is close or equal to zero that is the invariable zones. Hence this zone has been chosen as the reference zone. The mean spectrum of the collected spectra has then been calculated and has been used as reference spectrum. Each of the other spectra of the collection undergo on the whole spectrum a translation (positive or negative) so that the spectra in the invariante zone is confounded.

The corrected spectra are obtained following :

$$S_C = S - cst$$

where S_C is the corrected spectra, S the initial spectra and cst the difference in absorbance between the invariable zone of the reference spectrum and the same zone in the spectrum to be corrected. The spectra of the calibration set and those of the verification set are thus corrected from the same reference spectrum. An example is given in figure 2.

The predicted values, without correction and with correction based on the invariable zone are given in Table 1. The precision of the predicted sucrose content values is improved for the corrected samples. For samples that have not been corrected, the bias and the standard deviation values are respectively $1.52\ 10^{-2}$ and $3.17\ 10^{-1}$ % while for corrected samples (corrected from the invariable zone) the bias and SD

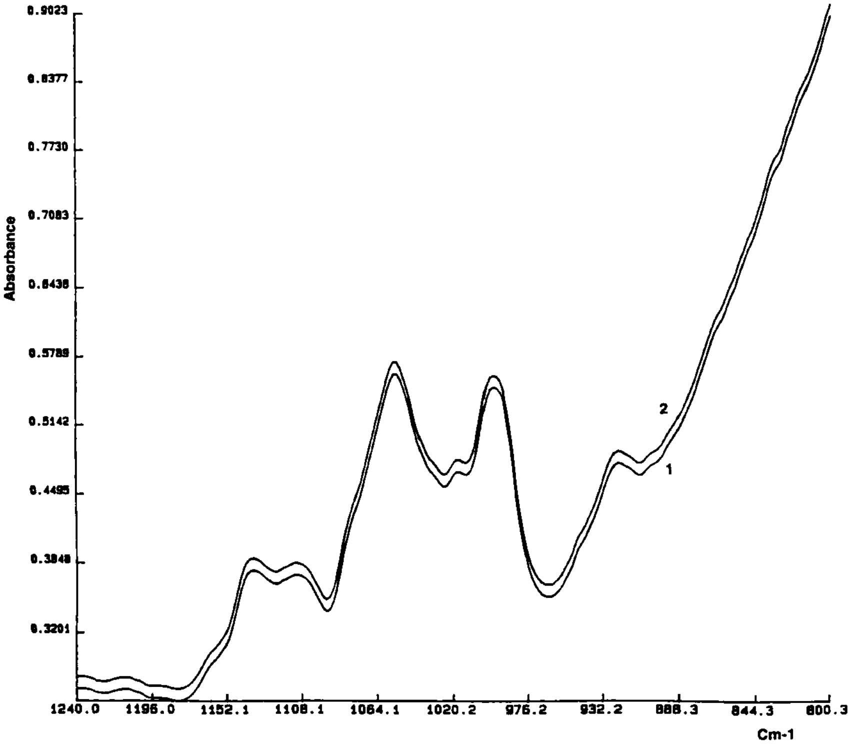

Figure 2. 800-1240 cm-1 FTIR spectra before (1) and after (2) baseline correction (cst = 0.011).

are -5.06 10^{-2} and 3.08 10^{-1} % respectively. In fact SD values significantly decreases.

Decomposition by Legendre polynomials.

Spectra can be expanded in terms of a set of orthonormal polynomials derived from the Legendre polynomials, and the leading terms of the expansion, which contains most of the baseline variation, can be removed. The visual comparaison of two spectra between α_1 cm^{-1} and α_2 cm^{-1} mainly consist in

comparing the areas beneath the spectra. A scalar product can hence be associated with the integral between α_1, and α_2.
Lipkus (1988) has introduced the use of an orthonormed polynomial Legendre system whose scalar product is given by :

$$\langle f,g\rangle = T \cdot \int_{\alpha_1}^{\alpha_2} f \cdot g \qquad (1)$$

When the orthonormed system (α_1, α_2) is considered : (P_i, i ε N) a spectrum S(α) can be decomposed into :

$$S(\alpha) = \sum_{i=0}^{+\infty} \beta_i \bullet P_i \qquad (2)$$

With the elimination of the first terms, it becomes :

$$S_0(\alpha) = S(\alpha) - \sum_{i=0}^{n} \beta_i \bullet P_i \qquad (3)$$

where $S_0(\alpha)$ represents the corrected spectrum and *n* an integer.

We have recently described this method (Cadet, 1996a). The application of Legendre correction to the spectra for the filtering parameter *n* equal to 2 is illustrated on figure 3 as an example.

The mean of the difference between reference and predicted values are 1.52×10^{-2} before correction and -6.13×10^{-1} % after correction, while standard deviation values are 3.17×10^{-1} and 2.76×10^{-1} % before and after correction (Table I). This treatment again improves the standard deviation value obtained after baseline correction.

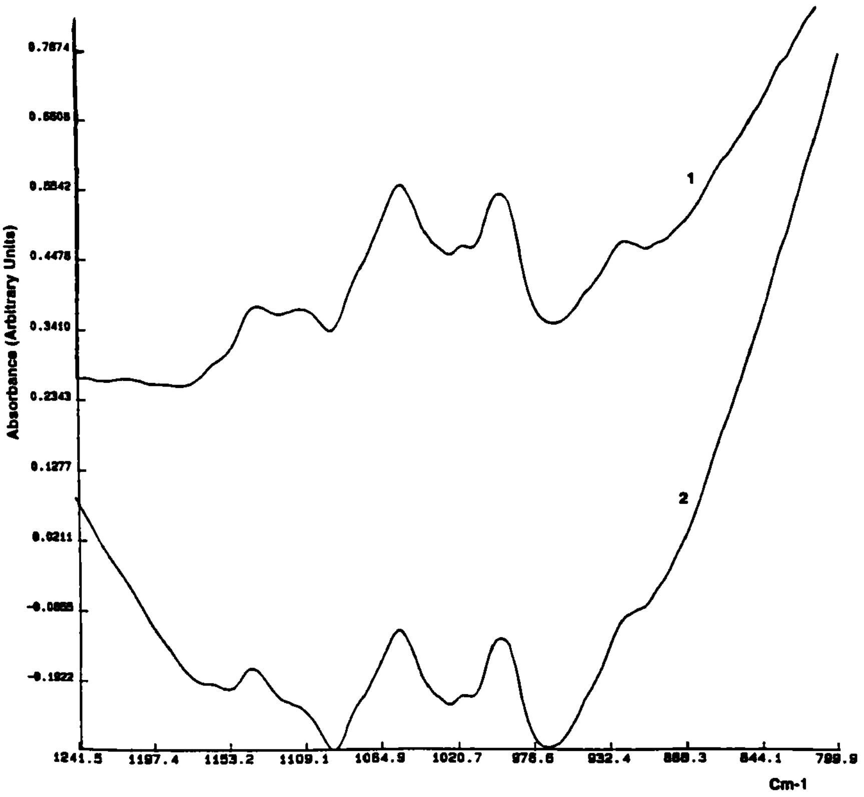

Figure 3. 800-1240 cm-1 FTIR spectra before (1) and after (2) decomposition Legendre polynomials (n=2).

Classification procedure with the use of the correlation distance

The method extensively described in a recent paper (Cadet, 1996b) consisted in isolating the two most dissimilar individuals from a large calibration family of Mid-FTIR spectra and by successive Principal Component Analysis and Principal Component Regression, a family composed of a few individuals was constituted. Each individual from this family represented the first spectrum of the corresponding classes that were ultimately formed. The classification of the remaining samples from the calibration family were carried out by the mobile centres method. The classification is done without knowing the chemical composition of samples.

Before the sucrose content value of a sample from the verification set is predicted, this sample is assigned to a class. The factorial map of the verification set (spectral collection) using the first two axes of the principal component analysis assessed on the calibration set is given in figure 4. We have shown that in a scope of quantifying sucrose in sugar cane juices the use of the euclidian distance in the classification procedure improves the results. However the correlation distance (half-defined and positive distance : the square of the sinus of two spectra) appears to be a more appropriate mathematical distance for the classification of Mid-infrared spectra of biological samples (Cadet, 1996c).

From the calibration family of 107 spectra, the four most dissimilar individuals were isolated and four classes were constituted. The classification procedure applied to each of these classes reduced the calibration family to 37 spectra that were distributed in 4 sub-classes. Each of the four sub-classes was assessed by principal component regression and four prediction equations were hence established. The mean and standard deviation of the differences between predicted and reference values were respectively 1.52×10^{-2} and 3.17×10^{-1} % before classification (one prediction equation is established with the whole calibration set) and -1.00×10^{-3} and 2.65×10^{-1} % after classification (Table 1). This procedure showed to improve, again, the precision of the predictions.

However, when the chemical values of the verification sets were directly predicted (without baseline correction and

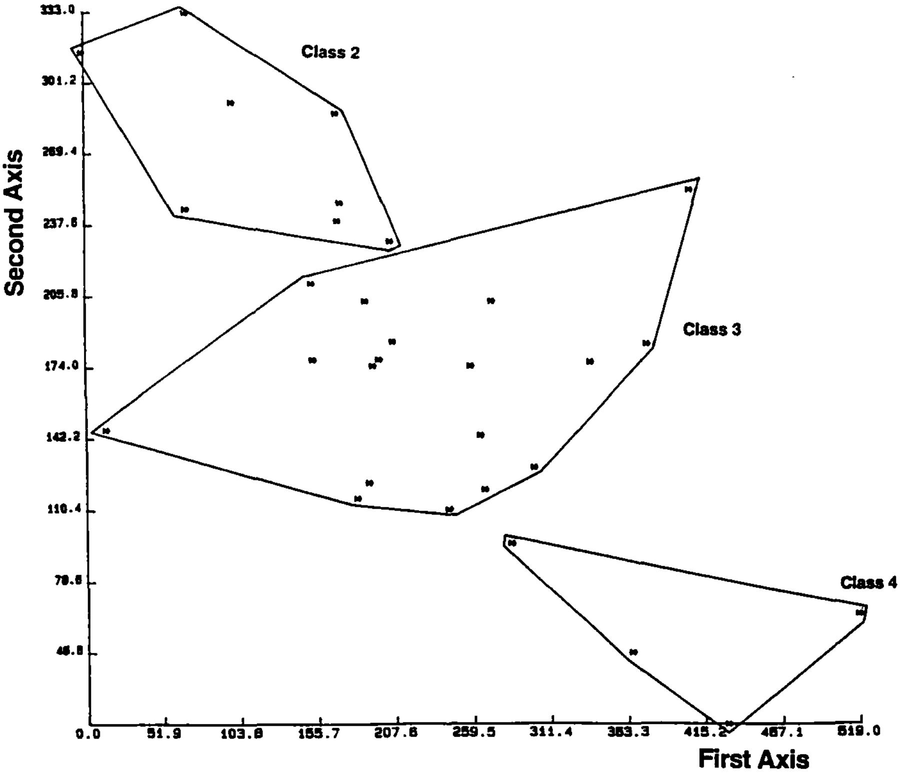

Figure 4. Classification of the verification set.

correction by Legendre polynomials decomposition) the mean and standard deviation were surprisingly 7.43×10^{-2} and 2.62×10^{-1} respectively.

CONCLUSION

The application of successive mathematical treatments to Mid-infrared spectra of biological samples (baseline correction, decomposition by Legendre polynomials and classification procedure) sensibly improves the precision of the predicted values of sucrose concentration.

The direct application of the classification procedure surprisingly gave similar results. It is known that the choice of the most appropriate mathematical distance in a classification procedure is difficult to make. We have recently shown that the correlation distance was more appropriate than the euclidian distance that is usually used (Cadet, 1996c). However the results obtained suggest that the mathematical distance used here (correlation distance) would not be the most appropriate mathematical distance. Even if the resulting classification procedure is not ideal, it appears to compensate aleatory translation errors (baseline correction) and baseline deformations (Legendre). This direct compensation is such that the result is sensibly better than that obtained after successive mathematical corrections.

These overall results show that the methods that have been proposed for the processing of Mid-IR spectra are interesting. The multiple mathematical treatment can be advantageously replaced by a single classification procedure hence simplifying considerably mathematical treatments applied to spectral data.

Acknowledgements

This work was supported by a grant from the Ministère de la Recherche et de la Technologie, the Conseil Général and the Conseil Régional de la Réunion.

REFERENCES

- Bertrand D., Robert P., Tran V., (1984), "Traitements mathématiques des spectres NIR de mélanges. 11ème Congrès de L'association internationale de chimie céréalière, Vienna, Austria, 6th june.

- Bertrand D., Lila M., Furtoss V., Robert P., and Downey G., (1987), "Application of principal component analysis to the prediction of lucerne forage protein content and in vitro dry matter digestibility by NIR spectroscopy" *J. Sci. Food Agric.*, **41**, 299-307.

- Cadet F., Bertrand D., Robert P., Maillot J., Dieudonné J., Rouch C. (1991), "Quantitative determination of sugar cane sucrose by multidimensional statistical analysis of thier Mid-Infrared attenuated total reflectance spectra", *Appli. Spectrosc.*, **45**, 2, 166-172.

- Cadet F., Wong Pin F., Rouch C., Robert C., Baret, P., (1995), "Enzyme kinetics by mid-infrared spectroscopy: ß-fructosidase study by a one step assay", *Biochim. Biophys. Acta*, 1246, 142-150.

- Cadet F., Offmann B., (1995) "Direct spectroscopic sucrose determination of raw sugar cane juices", *J. Agric. Food. Chem.,* submitted for publication.

- Cadet F., Offman B., (1996), "Baseline correction applied to a biological Mid-Infrared spectra collection", *Spectrosc. Lett.*, **29**, 3, under press.

- Cadet F., (1996a), "Application of Legendre polynomials correction before component quantification of biological Mid-Infrared spectra", *Spectrosc. Lett.*, **29**, 5, under press.

- Cadet F., (1996b), "A method of classification of biological FTIR spectra prior to quantitative analisys", *Applied Spectroscopy,* submitted for publication

- Cadet F., (1996c), "Choice of a particular mathematical distance for the classification of biological FTIR spectra", *Spectrosc. Lett.*, Submitted for publication.

- Cowe I.A. and Mc Nichol J.W., (1985), "The use of principal component in the analysis of Near Infrared spectra", *Appl. Spectrosc.*, **39**, 257-262.

- Crocombe R.A., Olson N.L., Hills S.L., (1987), "Quantitative fourier transform infrared methods for real complex samples.", American Society for testing and Materials, 95-130.

- Devaux M.F., Bertrand D., Robert P.,Qannari M., (1988) "Application of multidimensional analyses to the extration of discriminant spectral patterns from Near Spectra", *Appl. Spectrosc.*, **42**, 6, 1015-1019.

- Gillette P.C., Koenig J.L., (1982), "Noise reduction via factor analysis in FT-IR spectra", *Appl. Spectrosc.*, **36**, 5, 535-539.

- Hruschka W.R., (1987), "Data analysis : wavelength selection methods. In: "Near Infrared Technology in the agricultural and food industry", Eds Williams P.C. and Norris K.H., AACC, Saint Paul, MN (USA), 35-46.

- Lebart L. , Morineau A., and Tabard N., (1977), "Techniques de la Description Stastitique" (Dunod, Paris), 7-46 .

- Le Nouvel J., (1981), "Etude d'une famille de courbes par des méthodes d'analyse des données. Application à l'analyse morphologique de courbes provenant de données médicales", Thèse de 3ème cycle, Université de Rennes I, France.

- Lipkus A.H., (1988), "Application of Legendre polynomials to the elimination of baseline variation in biological FT-IR spectra", *Appl. Spectrosc.*, **42**, 3, 395-400.

- Osborne B.G., Fearn T, (1986) Near Infrared spectroscopy in food analysis, Longman Scientific and Technical, Wiley and Sons, NewYork.

- Powell J.R, Wasacz F.M., Jacobsen R.S., (1984), "An algorithm of the reproductible spectral subtraction of water from FT-IR spectra of proteins in dilute solutions and adsorbed monolayers," *Appl. Spectrosc.*, **40**, 3,339-345.

- Van de voort F.R, Ismail A. A, (1991), "A rapid FTIR quality control method for fat and moisture determination in butter", *Trends Food Sci. Technol.*, 13-17.

- Williams P., Norris K., (1987), " Near infrared technology in the agricultural and food industry", American Association of Cereal Chemists ed, Saint Paul, MN, USA, 330 p.

Received: March 18, 1996
Accepted: April 30, 1996