6

CONCLUSIONES GENERALES

6.1 CONCLUSIONES 149

CONCLUSIONES GENERALES

Las conclusiones que se pueden extraer del trabajo recogido en esta tesis, en relación a los objetivos inicialmente planteado, son las siguientes:

1. Se han establecido metodologías analíticas precisas y exactas para la determinación y seguimiento de las principales especies que caracterizan la fermentación alcohólica llevada a cabo por *Saccharomyces cerevisiae*; glucosa, etanol, biomasa, glicerina y acidez. Las determinaciones, realizadas con estos métodos, han sido utilizadas como "valores de referencia" para construir los modelos de calibración multivariantes para los diferentes analitos citados.

2. Se han fijado las condiciones de trabajo que permiten el desarrollo de un proceso de fermentación alcohólica. A pesar de la complejidad del medio en el que transcurren las fermentaciones alcohólicas, debido a la acumulación de biomasa y de productos procedentes del metabolismo de la levadura, se ha establecido una metodología de registro de espectros NIR, tanto en discontinuo (*at-line*) como en continuo (*in-line*), que ha permitido la monitorización y seguimiento de los cambios acaecidos durante el proceso de fermentación.

3. La utilización de herramientas multivariantes, por reducción de variables, ha posibilitado correlacionar la información espectral registrada con las determinaciones analíticas realizadas, de esta manera, se han creado modelos de calibración que permiten el seguimiento y la caracterización del proceso fermentativo en tiempo real. Por tanto, la espectroscopia en el infrarrojo cercano, como técnica analítica, en conjunción con PLS, como herramienta regresiva y predictiva, permiten el seguimiento, la monitorización y la obtención de información en tiempo real de los cambios acaecidos durante el proceso fermentativo.

4. La utilización de herramientas de factorización y resolución ha permitido el seguimiento de la evolución de la concentración de los analitos principales involucrados en la fermentación alcohólica: glucosa, etanol y biomasa. Por tanto, queda comprobado que el método de resolución MCR-ALS, utilizado para extraer información útil en diferentes problemas y situaciones analíticas, es también una herramienta apta para resolver los perfiles de las principales especies que interviene en un proceso de fermentación.

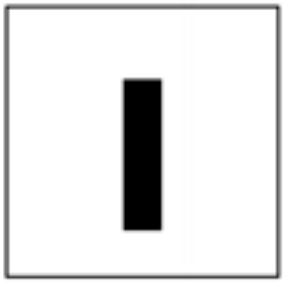
5. La utilización de modelos fermentativos empíricos, tradicionalmente usados para modelar bioprocesos y diseñar reactores biológicos, ha permitido refinar y mejorar los modelos de resolución creados. El uso conjunto de ambos ha permitido dilucidar los cambios producidos en procesos de fermentación alcohólica llevados a cabo en diferentes condiciones de temperatura y pH inicial.

6. Se han construido estructuras de datos tridimensionales, a partir de espectros NIR, utilizando la temperatura como dimensión adicional. Este proceder amplia el ámbito de utilización de los métodos de calibración *3-way*, poseedores de propiedades matemáticas favorables, pero de aplicación limitada debido a la falta de estructuras tridimensionales apropiadas, especialmente en el campo de la química analítica.

7. La utilización conjunta de los métodos PARAFAC y MLR, como estrategia sinérgica de creación de modelos, ha demostrado ser útil para eliminar el efecto distorsionador introducido por los cambios no modelados de temperatura. Esta asociación ha proporcionado buenos resultados, tanto en condiciones de extrapolación como de interpolación, cuando se ha aplicado a dos sistemas diferentes.

Con en esta tesis se amplía la línea de trabajo del grupo de investigación al seguimiento en continuo de procesos biológicos y a la utilización de herramientas de modelado por resolución y multi-way.

8. La experiencia y conocimientos adquiridos en el estudio del proceso de fermentación alcohólica permitirá seguir trabajando y profundizando en el conocimiento de nuevos bioprocesos. En este sentido, el proceso más inmediato, relacionado con el tema aquí presentado y del que ya se tiene alguna experiencia anterior, es la utilización de compuestos procedentes de desechos agrícolas y/o industriales pero susceptibles de ser fermentados y de ser sustrato de partida para la producción de etanol.
9. La experiencia adquirida en el análisis de información a través de herramientas de resolución ha permitido abrir las puertas al estudio y conocimiento de diferentes tipos de procesos y sistemas en evolución desde una perspectiva muy ventajosa, ya que este tipo de herramientas tienen unas demandas de información química de referencia nulas o muy bajas.



ANALYTICAL MONITORING OF ALCOHOLIC FERMENTATION USING NIR SPECTROSCOPY

Blanco M., Peinado A. C., Mas J.

Biotechnology and Bioengineering , **2004**, 88(4), 536-542

Analytical Monitoring of Alcoholic Fermentation Using NIR Spectroscopy

Marcel Blanco,[1] Antonio C. Peinado,[1] Jordi Mas[2]

[1]Department of Chemistry, Faculty of Sciences, Universidad Autónoma de Barcelona, 08193 Bellaterra, Barcelona, Spain; telephone: +34-93581-1367; fax: +34-93581-2379; e-mail: marcel.blanco@uab.es
[2]Department of Biology, Faculty of Sciences, Universidad Autónoma de Barcelona, Barcelona, Spain

Received 29 December 2003; accepted 4 June 2004

Published online 6 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/bit.20214

Abstract: Alcoholic fermentation under *Saccharomyces cerevisiae* yeasts is governed largely by glucose uptake, biomass formation, ethanol and glycerin production, and acidification. In this work, PLS calibration models were developed with a view to determining these analytical parameters from near infrared spectra and analytical data provided by the corresponding reference methods. The models were applied to a set of samples obtained from various fermentation processes. The glucose, ethanol, and biomass values predicted by the models exhibited a high correlation with those provided by the reference method.

Keywords: near infrared spectrsocopy; alcoholic fermentation; multivariate calibration

INTRODUCTION

After the oil crisis of 1973, the world production of bioethanol (a renewable and alternative source of energy to traditional fossil fuels) has grown steadily due to its use in the formulation of liquid fuels as carburant as well as antidetonant agent. The world figure for the year 2006 is expected to be in the region of 35 gigaliters and thus 75% greater than that for 2001 (20.3 Gl; www.distill.com/world_ethanol_production.html). Industrial ethanol is obtained by chemical synthesis (7%) or alcoholic fermentation of various glucose-rich substrates (93%) (www.distill.com/berg/).

Saccharomyces cerevisiae is the microorganism most widely used in industrial alcoholic fermentations; it is a nonpathogenic GRAS (generally regarded as safe) microbe. Metabolically, *S. Cerevisiae* is a facultative aerobic microorganism; in the presence of glucose concentrations above 9 g/l, however, it metabolizes glucose via a fermentation pathway (Ribéreau-Gayon et al., 2000). Although alcoholic fermentation yields ethanol as the main product, the process involves additional metabolic pathways, the most important of which is glyceropyruvic fermentation; this produces glycerol and pyruvic acid, the latter evolving to various compounds including succinic and lactic acids, acetone and 2,3-butanediol. *S. cerevisiae* yeasts can thus be used for the industrial production not only of ethanol (Arikava et al., 1999; Overkamp et al., 2002), but also, upon genetic modification, of other foodstuffs (Hansen et al., 2001), with improved sensory properties and an increased added value.

The increasing demand for ethanol, its fermentative origin, and the wide variety of intermediate metabolites and end coproducts obtained during alcoholic fermentation have raised the need for analytical methods capable of providing information about the status of fermentation processes in real time.

A number of mathematical models have been used to monitor alcoholic fermentation under different batch (Birol et al., 1998) and continuous (Larsosson and Enfors, 1999) operating conditions; such models, however, are subject to many severe restrictions and fail when the conditions imposed are not met, which is frequently the case, as a result of composition nonuniformity in the raw material (Sainz et al., 2003) or the variability inherent in the fermentation process (Cramer et al., 2002), for example. In addition, the models are constructed from data obtained after the alcoholic fermentation has completed or once the sample, withdrawn by using an appropriate offline reference method (Salmon, 1998), has been analyzed, so the analytical datum lacks temporal value as fermentation continues while the sample is being processed. However, near infrared (NIR) spectroscopy allows processes to be monitored and analytical information withdrawn from them in real time (Blanco et al., 2000; Adamopoulos et al., 2001; Cimander et al., 2002).

The aim of this work was to develop calibration models for the different analytical parameters involved in alcoholic fermentation under *S. cerevisiae* yeasts, namely, glucose, ethanol, glycerin, biomass, and acidity. The models should

Correspondence to: Marcel Blanco
Contract grant sponsor: Ministerio de Ciencia y Tecnología (MCyT), Spain
Contract grant number: BQU2003-04247

effectively allow alcoholic fermentation to be monitored in real time.

EXPERIMENTAL

Microorganisms, Media, and Growth Conditions

Saccharomyces cerevisiae ATCC 1326 strain was obtained from the American Type Culture Collection. The strain was maintained on YPD agar medium (1% w/v yeast extract, 2% w/v peptone, 2% w/v glucose) at 4°C. Precultures of yeast cells were grown in 250 ml Erlenmeyer flasks containing 50 ml of Wickerman medium (0.5% m/v peptone, 0.3% m/v yeast extract, 0.3% m/v malt extract) supplemented with 20% w/v glucose and stirred at 400 rpm at 25°C for 48 hr, after which they were used to inoculate a bioreactor containing 2.75 liters of the same medium as the preculture. Glucose was obtained from Panreac, and peptone, yeast extract, and malt extract from ADSA Micro. Fermentation runs were conducted in a 3 liter bioreactor (New Brunswick Scientific Bio-Flo model) equipped with temperature and stirring controls. The operating temperature and stirring rate were 25°C and 400 rpm, respectively.

Samples

The samples used to develop and validate the calibration models were obtained in two different ways.

Fermentation Samples

These were directly withdrawn from the fermentation system. Overall five fermentation runs were conducted. Each fermentation sample was split into five aliquots that were used for the following determinations: glucose and ethanol (following dilution), biomass (following centrifugation), glycerin (upon dilution), acidity (expressed as acetic acid), and NIR spectrum.

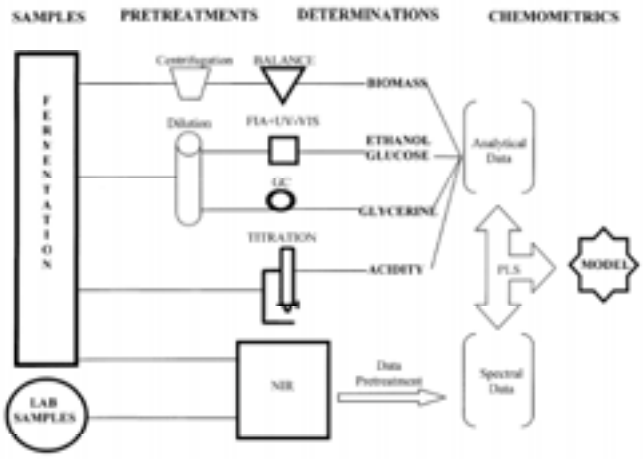


Figure 1. Schematic description of the procedure used to obtain spectral and analytical data for constructing the PLS models.

Laboratory Samples

These were prepared by mixing accurately weighed amounts of glucose, ethanol, acetic acid, glycerin, and biomass in proportions differing from those obtained in the fermentation process. The NIR spectra for these samples were included in the spectral data matrix used to construct the calibration models. Figure 1 depicts the techniques used to obtain the reference values used in conjunction with the NIR spectra to construct the partial least squares (PLS) models.

Reference Methods

Glucose and ethanol were individually determined in the flow injection analysis (FIA) manifold of Figure 2. The determination of glucose was based on the following coupled enzymatic reactions:

$$\text{Glucose} + O_2 + H_2O \xrightarrow{\text{GOD}} \text{Gluconic acid} + H_2O_2$$

$$2\ H_2O_2 + \text{4-aminophenazone} + \text{phenol} \xrightarrow{\text{POD}} \text{p-benzoquinone-4 phenazone} + 4\ H_2O$$

The latter reaction gave a colored product that was monitored spectrophotometrically at 506 nm. The composition of the reagents was as follows. The buffer was a 0.1 M aqueous solution Na_2HPO_4 (Merck) adjusted to pH 7.0. The enzymatic reagent was prepared in the previous buffer and contained 12.5 U GOD/ml, 5 U POD/ml, 1.5 mM 4-aminophenazone (Sigma), and 9.3 mM phenol (Fluka). GOD (glucose oxidase E.C. 1.1.3.4) and POD (peroxidase E.C. 1.11.1.7) were both obtained from Sigma.

The determination of ethanol relied on the following enzymatic reaction:

$$CH_3CH_2OH + NAD^+ \xrightarrow{\text{ADH}} CH_3CHO + NADH + H^+$$

The NADH thus formed was monitored spectrophotometrically at 340 nm. All chemicals used in this reaction were reagent-grade. The composition of the reagents was as

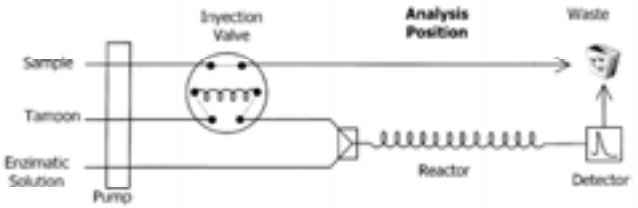


Figure 2. Elements and placement of the FIA manifold used to determine ethanol and glucose.

Table I. Samples included in the calibration and external validation sets for the different analytes.

		Number of samples for				
Set	Sample type	Glucose	Ethanol	Acetic acid	Glycerine	Biomass
Calibration	Laboratory	12	12	8	31	9
	Fermentation	14	14	20	6	12
	Total	26	26	28	37	21
	Concentration range	0–225 g/L	0–18%	0–50 g/L	0–10 g/L	0–14 g/L
Validation	Laboratory	5	5	5	8	0
	Fermentation	18	12	12	9	11
	Total	23	17	17	17	11

follows. The buffer was an aqueous solution containing 75 mM sodium pyrophosphate, 75 mM semicarbazide hydrochloride (Fluka), 21 mM glycine, and 0.15 M sodium chloride and adjusted to pH 8.0 with sodium hydroxide. The enzymatic reagent was prepared in the previous buffer and contained 130 U ADH/l and 1 g NAD^+/l. Both ADH (alcohol dehydrogenase E.C. 1.1.1.1.) and NAD^+ were obtained from Sigma.

One aliquot of each sample withdrawn from the fermentation reactor was diluted 100–200 times in order to accommodate its results within the linear determination range. Biomass concentration was measured using the dry weight method. The protocol was extracted from Rhodes and Stanbury (1997). Pyrex tubes were dried to 105°C, cooled in a dessicator, and tared. Five milliliter samples (by triplicates) were obtained and centrifuged at 10,000*g* at 4°C for 15 min. After washing with deionized water, the biomass was separated by recentrifugation. The resulting residue was dried at 105°C until constant weight was achieved.

The glycerin concentration was determined by gas chromatography according to Savchuk et al. (1999). A Hewlett-Packard HP 5890 Series II gas chromatograph equipped with a flame ionization detector and a Supelco SPB-1701 fused silica capillary column (1.5 m × 0.25 mm i.d., 0.25 µm film thickness) was used for this purpose. Acidity was determined by titrating a 5 ml sample from the fermentation system with 1 M NaOH in the presence of phenolphthalein as indicator. Total acidity was expressed as acetic acid.

Apparatus and Software

Spectra were recorded on a FOSS NIRSystems 6500 spectrophotometer (Raamsdonksveer, The Netherlands) equipped with a rapid content analyzer (RCA) module that was furnished with a gold reflector having an optical spacing of 0.5 mm. The instrument was governed and the data were acquired using the Vision 2.51 software package, also from NIRSystems. PLS models were constructed using Unscrambler 8.0 from CAMO (Trondheim, Norway).

An appropriate amount of sample was placed in a flat-bottom quartz cuvette and the gold reflector positioned in its bottom. Each recorded spectrum was the average of 32 scans performed at 2 nm intervals over the wavelength range of 1,100–2,500 nm.

Data Processing

Spectral data were subjected to various treatments, including the standard normal variate (SNV) and the first and second derivatives, which were obtained using the Savitzky-Golay algorithm (Savitzky and Golay, 1964) with a second-order polynomial and a window size of 11 points.

Preprocessed data provided by the reference methods were modeled using the PLS1 algorithm (Geladi and Kowalski, 1986). PLS1 models were constructed by cross-validation using the leave-one-out procedure. The optimum number of PLS1 components was taken to be that minimizing the sum of residuals,

$$\text{PRESS} = \sum_{i=1}^{m} (\hat{y}_{\text{NIR}} - y_{\text{REF}})^2$$

where m is the number of samples used to construct the model, y_{REF} the reference value, and $\hat{y}_{\text{NIR}}$ that provided by the model.

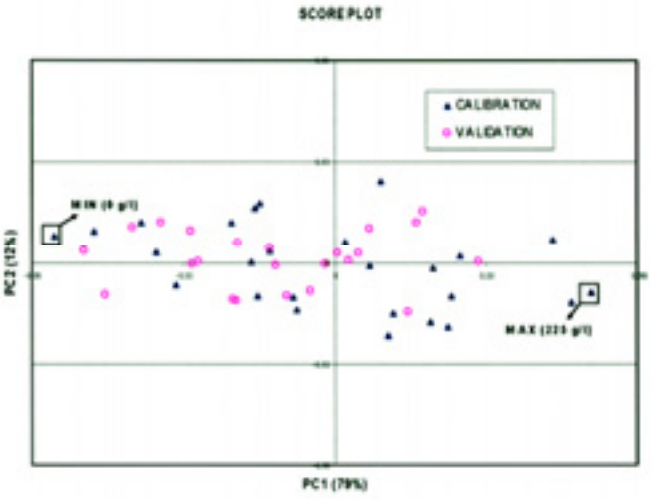


Figure 3. Plot of scores for selected samples in the calibration and validation sets for glucose. The samples within each square bound the two concentration extremes.

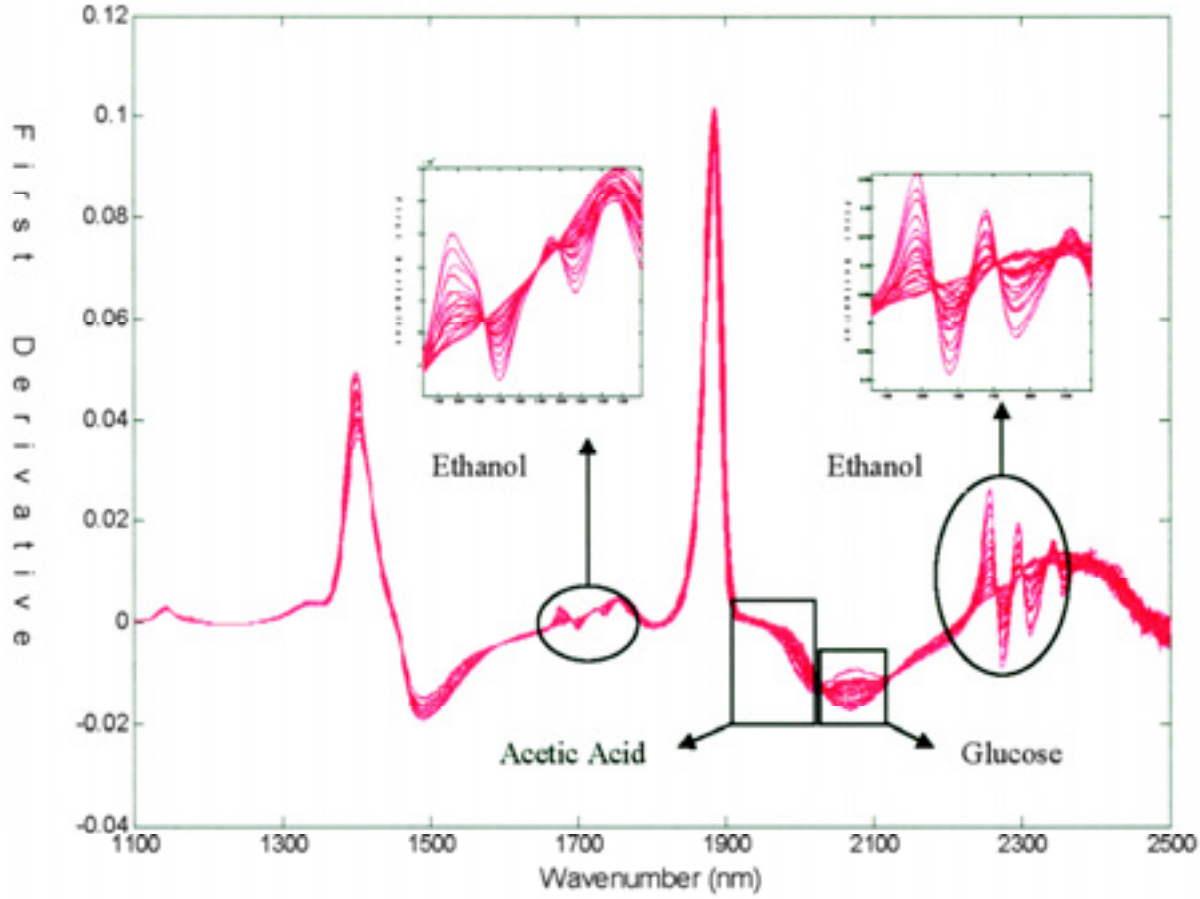


Figure 4. First-derivative spectra for laboratory samples. Detail of the ranges where spectra arranged as a function of the ethanol content.

The goodness of the results provided by the different PLS1 models was assessed in terms of the relative standard error,

$$\text{RES}\ (\%) = \sqrt{\frac{\sum_{i=1}^{m} (\hat{y}_{\text{NIR}} - y_{\text{REF}})^2}{\sum_{i=1}^{m} (y_{\text{REF}})^2}} \times 100$$

and designated RSEC (%) for calibration and RSEP (%) for external validation.

RESULTS AND DISCUSSION

Table I shows the number, origin, and concentration ranges spanned by the samples used for calibration and validation of the models. Both the calibration set and the validation set included a number of laboratory samples. This allowed collinearity in the concentrations of the different analytes involved in the fermentation process to be reduced and the concentration range encompassed by the calibration models to be expanded. For all analytes, the calibration range was broader than the validation range.

In order to ensure that variability in both fermentation samples and laboratory samples was considered in the calibration and validation sets, the samples to be included in each were chosen using principal component analysis (PCA). By way of example, Figure 3 shows a plot of the scores for the first principal component (PC) against those

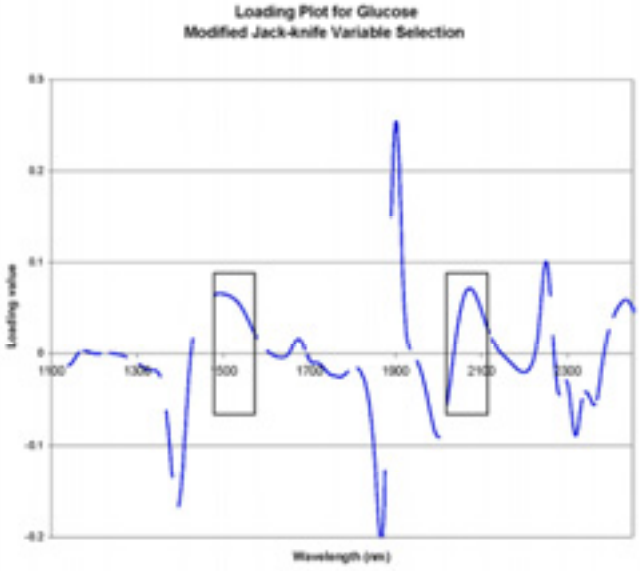


Figure 5. Loadings of the first component for glucose. Discontinuities correspond to variables not selected by the modified jackknife variable selection method. Rectangular areas correspond to spectral zones ascribed to glucose.

Table II. Descriptive statistics of the calibration and validation sets for the different analytes.

					Calibration		Validation	
Analyte	Spectra pretreatment	Wavelength range (nm)	Number of variables	PLS factors	R2	RSEC	R2	RSEP
Ethanol	First derivative	1,650–1,820 + 2,240–2,400	165	2	0.9997	1.60	0.9976	5.04
Glucose	First derivative	Jackknife	506	3	0.9959	4.07	0.9984	4.81
Glycerine	First derivative	Jackknife	168	9	0.9972	4.87	0.9955	6.20
Acetic acid	First derivative	Jackknife	312	8	0.9937	7.85	0.9842	7.98
Biomass	Second derivative	1,100–1,500 + 1,760–2,010 + 2,350–2,500	400	4	0.9905	4.41	0.9755	6.94

for the second; the samples included in the calibration and validation sets in order to construct the calibration model for glucose are highlighted. The first PC accounted for 79% of the overall spectral variance, and the second for 12%. Samples arranged in increasing glucose content in the direction of the first PC.

The spectra obtained during the fermentation process exhibited an increasing shift due to yeast growth and consisted mainly of the bands for water, which was the major component and that with the highest absorptivity. These spectra could not be used to assign specific spectral regions to the analytes with a view to selecting the most suitable spectral range for constructing the models. However, the first-derivative treatment corrected constant baseline shifts and exposed the spectral information concealed in the absorbance spectra. Figure 4 shows the first-derivative NIR spectra for 20 laboratory samples containing the analytes at concentrations spanning the ranges listed in Table I. The very strong band at ~1,900 nm and that at 1,420 nm are the combination band and first overtone, respectively, for the O—H bond in water.

The region between 2,200 and 2,360 nm contains a succession of alternate maxima and minima, and so does that from 1,660 to 1,780 nm, where peaks, however, are smaller than in the previous one. The peaks in these regions can be ascribed to a combination of tones and to the first overtone of C—H bonds in ethanol. Spectra were arranged in increasing concentration of ethanol.

Similarly, the 2,030–2,130 and 1,480–1,580 nm regions exhibit a sequence of spectra arranged in accordance with the glucose concentration. These regions are roughly those where the tone combinations and first overtone for O—H bonds in glucose occur.

In the spectral region from 1,950 to 2,030 nm, which corresponds to the spectral range where the COOH group absorbs, samples arranged as a function of their acetic acid content. The spectral region from 1,100 to 1,350 nm exhibited bands for the second overtones of C—H bonds. Samples arranged in terms of the glycerin and biomass contents in no spectral region, however. Models for the different analytes were constructed from whole spectra, using the absorbance, first-derivative, second-derivative, and SNV modes; various combinations and the jackknife variable selection test, as modified by Martens and Martens (2000), were used for each analyte.

The best results in all instances were provided by first-derivative data. By exception, the biomass was best determined from second-derivative data; this is rather surprising as the second-derivative treatment eliminated shifts in the absorption spectrum by effect of growth yeast. However, this result is consistent with those previously obtained by other authors using derivative spectral treatments in models for determining biomass (Sivakesava et al., 2001; Giavasis et al., 2003).

The jackknife variable selection test simplified the models and provided improved results for all analytes except ethanol and biomass, the best models for which were those based on manual selection of wavelengths. By way of example, Figure 5 shows the values of the first loading for the glucose model. Discontinuities in the variables were a result of the jackknife method excluding those variables significant at a level below 95%. As can be seen, every variable corresponding to the spectral regions that were assigned to glucose (namely, 2,030–2,130 and 1,480–1,580 nm, rectangular zones in the graph) was selected by the jackknife test.

Once the best models for the analytes were established, they were applied to their respective validation sets. Table II shows the values of the characteristic parameters for the models (namely, wavelength range, number of variables, and PLS factors), as well as the statistics defining the goodness of the best models for each analyte. The simplest models for ethanol and glucose were also those providing the best calibration and validation statistics; this was a result of the analytes being the compounds present at the highest concentrations during the fermentation process.

Table III. Figures of merit in the validation models for the different analytes.*

			Residuals		
Analyte	Slope	Offset	d.f.	t experimental	t critic
Ethanol	1.01 ± 0.08	0.12 ± 0.39	22	1.39	2.07
Glucose	1.04 ± 0.10	−3.71 ± 10.82	16	1.33	2.12
Acetic acid	1.01 ± 0.10	−0.11 ± 1.95	16	−0.28	2.12
Glycerine	0.97 ± 0.04	0.15 ± 0.21	16	−1.95	2.12
Biomass	0.97 ± 0.09	0.11 ± 0.7	10	1.81	2.23

*± indicate the confidence interval.

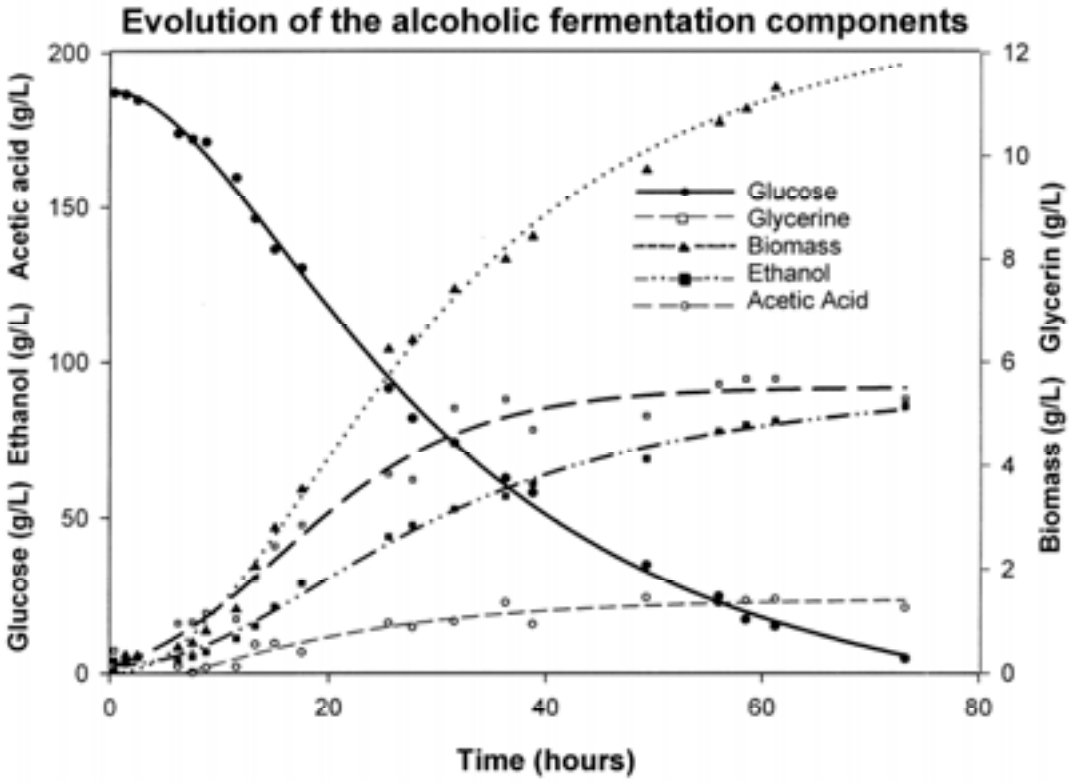


Figure 6. Temporal variation of the concentrations of the major analytes involved in alcoholic fermentation.

Table III lists the figures of merit obtained in the validation of the different models; as can be seen, the slope and intercept values included unity and zero, respectively, in the 95% confidence levels. Table III also shows the results of a significance test for the residuals. The experimental *t* values were all smaller than the critical *t* value at the 95% confidence level and the corresponding number of degrees of freedom (d.f.). There were thus no significant differences between the values predicted by the models for each analyte and those provided by the corresponding reference method.

The calibration models obtained for the different fermentation components were used to monitor other fermentation runs not used to construct or validate the models. Figure 6 shows the NIR predicted values for a fermentation process that was monitored for 74 hr; the graph shows the temporal variation of the different species studied. As can be seen, the glucose concentration decreased and the ethanol concentration increased throughout the process. The glycerin, acid, and biomass concentration initially increased and then virtually leveled off after a time that varied with the particular analyte.

CONCLUSIONS

The methods usually employed to monitor fermentation processes are slow and labor-intensive and use reagents with a potential environmental impact. In this work, we used NIR spectroscopy in combination with multivariate calibration techniques to accomplish the rapid, reliable, affordable, nondestructive determination of ethanol, glucose, biomass, glycerin, and acidity in samples from an alcoholic fermentation process under *Saccharomyces cerevisiae* yeasts. Although tests were conducted in the at-line mode, a probe will be used in future work to allow the process to be monitored in real time. The proposed models are thus effective alternatives to the analytical methods traditionally used to monitor fermentation processes.

Supported by a grant from Spain's Ministerio de Ciencia y Tecnologia (MCyT) (to A.C.P.).

References

Adamopoulos KG, Goula AM, Petropakis HJ. 2001. Quality control during processing of feta cheese NIR application. J Food Comp Anal 14:431–440.

Arikava Y, Kuroyanagi T, Shimosaka M, Muratsubaki H, Enomoto K, Kodaira R, Okazaki M. 1999. Effect of gene disruptions of the TCA cycle on production of succinic acid in *Saccharomyces cerevisiae*. J Biosci Bioeng 87:28–36.

Birol G, Doruker P, Kirdar B, Önsan Z, Ülgen K. 1998. Mathematical description of ethanol fermentation by inmobilised *Saccharomyces cerevisiae*. Process Biochem 33:763–771.

Blanco M, Coello J, Iturriaga H, Maspoch S, González R. 2000. On-line monitoring of starch enzymatic hydrolysis by near-infrared spectroscopy. Analyst 125:749–752.

Cimander C, Carlsson M, Mandenius C. 2002. Sensor fusion for on-line monitoring of yoghurt fermentation. J Biotech 99:237–248.

Cramer AC, Vlassides S, Block DE. 2002. Kinetic model for nitrogen-limited wine fermentations. Biotechnol Bioeng 77:49–60.

Geladi P, Kowalski BR. 1986. Partial least-squares regression: a tutorial. Ana Chim Acta 185:1–17.

Giavasis I, Robertson I, McNeil B, Harvey LM. 2003. Simultaneous and rapid monitoring of biomass and biopolymer production by Sphingomonas paucimobilis using Fourier transform-near infared spectroscopy. Biotechnol Lett 25:975–979.

Hansen TK, Tempel TV, Cantor MD, Jakobsen M. 2001. *Saccharomyces cerevisiae* as a starter culture in mycella. J Food Microbiol 69: 101–111.

Larsosson G, Enfors SO. 1991. Modelling of aerobic growth of *Saccharomyces cerevisiae* in a pH-auxostat. Bioprocess Eng 20:534–544.

Martens H, Martens M. 2000. Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Qual Prefer 11:5–16.

Overkamp KM, Bakker BM, Köteer P, Luttik MAH, Van Dijken JP, Pronk JT. 2002. Metabolic engineering of glycerol production in *Saccharomyces cerevisiae*. Appl Environ Microbiol 68:2814–2821.

Rhodes PM, Stanbury PF. 1997. Applied microbial physiology: a practical approach. Oxford: IRL Press.

Ribéreau-Gayon P, Dubourdieu D, Donèche B, Lonvaud A. 2000. Handbook of enology, vol. 1, the microbiology of wine and vinifications. New York: John Wiley and Sons.

Sainz J, Pizarro F, Pérez-Correa J, Agosin E. 2003. Modeling of yeast metabolism and process dynamics in batch fermentation. Biotechnol Bioeng 81:818–828.

Salmon JM. 1998. Determination of oxygen utilization pathways in an industrial strain of *Saccharomyces cerevisiae* during enological fermentation. J Ferment Bioeng 86:154–163.

Savchuk SA, Brodskii ES, Fomanovskii AA. 1999. Determination of glycols in potable water and alcoholic beverages by gas chromatography and chromatography-mass spectrometry. Anal Chem 54: 741–752.

Savitzky A, Golay MJE. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36:1627–1639.

Sivakesava S, Irudayaraj J, Ali D. 2001. Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques. Process Biochem 37: 371–378.

ELUCIDATING THE COMPOSITION PROFILES OF ALCOHOLIC FERMENTATION BY USE OF ALS METHODOLOGY

Blanco M., Peinado A. C., Mas J.

Analytica Chimica Acta, **2005**, 544(1-2), 199-205

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 544 (2005) 199–205
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Elucidating the composition profiles of alcoholic fermentations by use of ALS methodology

M. Blanco[a,*], A.C. Peinado[a], J. Mas[b]

[a] Department of Chemistry, Faculty of Sciences, Autonomous University of Barcelona, 08193 Bellaterra, Barcelona, Spain
[b] Department of Biology, Faculty of Sciences, Autonomous University of Barcelona, 08193 Bellaterra, Barcelona, Spain

Received 15 October 2004; received in revised form 14 January 2005; accepted 14 January 2005
Available online 11 February 2005

Abstract

Alcoholic fermentation runs under *Saccharomyces cerevisiae* yeasts were conducted on culture medium batches containing glucose as carbon source. Spectral changes during the process were monitored in-line with a near infrared (NIR) immersion probe. Data were analysed by using a multivariate curve resolution–alternating least-squares (MCR–ALS) method. Different regions of the NIR spectrum were examined in order to ensure optimum application of the ALS algorithm and elucidation of the chemical rank for the system. The ambiguity inherent in the ALS algorithm was resolved by using various combinations of inequality and equality constraints. Some combinations were found to perform quite well in terms of explained variance and lack of fit, even in the absence of information in the form of equality constraints. The resulting model exposed a highly significant relationship between the ALS response and the reference concentration. Application of the model to alcoholic fermentation runs performed under similar conditions resulted in also similar analytical figures of merit. This allows the MCR–ALS method to be used to obtain the analyte profiles as a function of time and the spectral profiles in evolving alcoholic fermentations.

Keywords: Alcoholic fermentation; Alternating least-squares; *Saccharomyces cerevisiae*; NIR; PLS; In-line monitoring

1. Introduction

The use of ethanol as a gasoline antioxidant has always been subject to the feasibility of obtaining it in inexpensively. Thus, 93% of all ethanol produced worldwide is obtained by alcoholic fermentation of an appropriate substrate [1], most often in the presence of *Saccharomyces cerevisiae* yeasts. Glucose uptake, biomass formation, the production of ethanol as end-product and glycerin as by-product, and the gradual acidification of the medium are the most prominent processes occurring during alcoholic fermentation in their presence.

There have been various attempts at increasing fermentation productivity including the search for highly efficient yeast strains [2], optimization of the fermentation conditions [3] and even the use of an extraction process in combination with fermentation [4].

One other point to be considered in accomplishing efficient alcohol fermentation is the implementation of an analysis and monitoring approach allowing the process to be maintained under optimum conditions throughout. This, however, is no easy task as the primary result of the complex nature of microbial metabolism and its non-linear kinetics [5].

Traditional analytical methods (HPLC for glucose and GC for ethanol) are expensive and time consuming for alcoholic fermentation monitoring. By contrast, infrared spectroscopy allows the expeditious determination of several analytes in a simultaneous manner without the need for several reagents; this has facilitated the monitoring of fermentation processes of diverse nature [6,7]. Near infrared spectroscopy (NIRS) provides major advantages for the analysis of products with highly absorbing and dispersive matrices such as cell culture media. The way analytical information is extracted from NIR signals has evolved in parallel to the development of new chemometric methods, one of the most widely checked and used is partial least-squares regression (PLSR) [8]. The tri-

* Corresponding author. Tel.: +34 935810983; fax: +34 935812345.
E-mail address: marcel.blanco@uab.es (M. Blanco).

0003-2670/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2005.01.033

linear version of this method [9], which is parsimonious but less markedly affected by noise in the original variables, has been successfully used to model the performance of industrial fed-batch fermentations [10]. Models are constructed from reference data obtained by using classical analytical techniques.

The chemometric methodology known as "multivariate self-modelling curve resolution" (MCR) does not require the knowledge of reference analytical information in order to provide information about composition changes in an evolving system. This is especially desirable with complex evolving chemical systems and processes where one or more influential variables (e.g. temperature, pH, analyte concentrations) change with time. These techniques have one feature in common: they allow the temporal information about the concentration of the system components (concentration or kinetic profiles) to be obtained, as well as the purely spectral information (spectral profiles or individual spectra). MCR methodology has been successfully used with process analysis systems [11]. It has not, however, been applied to biological processes—particularly alcoholic fermentation, where chemical transformations are effected by microorganisms.

MCR methods rely on a number of algorithms [12] of which alternating least-squares (ALS) is the most widely used for the simultaneous elucidation of spectral and concentration profiles. This algorithm requires no reference information but can profit from any available knowledge about the target system.

The aim of this work was to determine the kinetic and spectral profiles for the key species in alcoholic fermentation using the ALS algorithm with a view to comparing the results with the analytical information obtained by using a previously validated model and checking whether MCR methodology is an effective choice for studying alcoholic fermentations.

2. Theory

Curve resolution methods resolve the experimental data matrix, $\mathbf{A}(r \times c)$, into the following matrix product:

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

where $\mathbf{C}(r \times n)$ is the matrix containing the temporal variation of the analyte concentrations (i.e. the kinetic profiles); $\mathbf{S}^{\mathrm{T}}(n \times c)$ is that containing the variation of the response with respect to different variables (i.e. the spectral profiles); $\mathbf{E}(r \times c)$ is the random perturbation matrix, which contains the residual variation of the data due to no analyte and is assumed to be independent and exhibit a constant variance; r is the dimension related to the temporal variation of the system, which coincides with the number of measured samples; c is the dimension related to the variation of the recorded response, which coincides with the number of variables; n is the dimension related to the number of analytes the concentrations of which change with time and alter the signal recorded by the detector.

Resolving matrix **A** is no easy task. Curve resolution methods cannot deliver a single solution [13] as they are subject to both rotational and scale (or intensity) ambiguities [14]. In order to avoid such ambiguities, the system must be subjected to some constraint [15] on the concentration and/or spectral profiles. The constraints used on MCR methods are based on available information about the system concerned and can be of two different types: inequality and equality constraints. Inequality constraints include non-negativity, which can be applied to both the concentration domain and the spectral domain, provided profiles are positive; unimodality, which can be applied to the analyte concentration profile as long as it exhibits a single maximum. Equality constraints include closure or mass balance constraint, which are applicable when the combined concentration of the analytes remains constant and known throughout the process. A second type of possible equality constraints can be used when the kinetic and/or spectral profile for some species is known.

Fig. 1 depicts the ALS algorithm used. The key steps in its implementation are the determination of the chemical rank or number of significant components, and the preliminary estimation of the kinetic and/or spectral profiles by using an exploratory procedure such as evolving factor analysis (EFA) [16] or simple to use interactive self-modelling mixture analysis (SIMPLISIMA) [17]. The algorithm ends by repeating the loop a preset number of times—unless the tolerated convergence, δ, is reached before. In each iteration of the loop, the pseudo-inverse of the concentration and spectral profiles (i.e. the least-squares solution) is calculated, the matrices thus obtained being used to re-estimate the kinetic profiles from the spectra and the spectra from the kinetic profiles.

The performance of the model can be assessed in a generic manner by using only two parameters, namely: the variance explained by the model, EV, and the statistic LOF (lack of fit), where a_{ij} denotes the elements in experimental matrix **A** and $\hat{a}_{ij}$ the values calculated by using the model with Eq. (1):

$$\mathrm{EV} = \frac{\sum_i \sum_j \hat{a}_{ij}^2}{\sum_i \sum_j a_{ij}^2} \times 100,$$

$$LOF = \sqrt{\frac{\sum_i \sum_j (a_{ij} - \hat{a}_{ij})^2}{\sum_i \sum_j a_{ij}^2}} \times 100$$

The quality of the spectral profiles recovered can be assessed by comparing the ALS calculated spectra ($\hat{s}_i$) with the reference spectra (s_i) via the dissimilarity criterion:

$$\sin z = \sqrt{1 - \cos^2 \frac{s_i^{\mathrm{T}} \hat{s}_i}{\|s_i\| \|\hat{s}_i\|}}$$

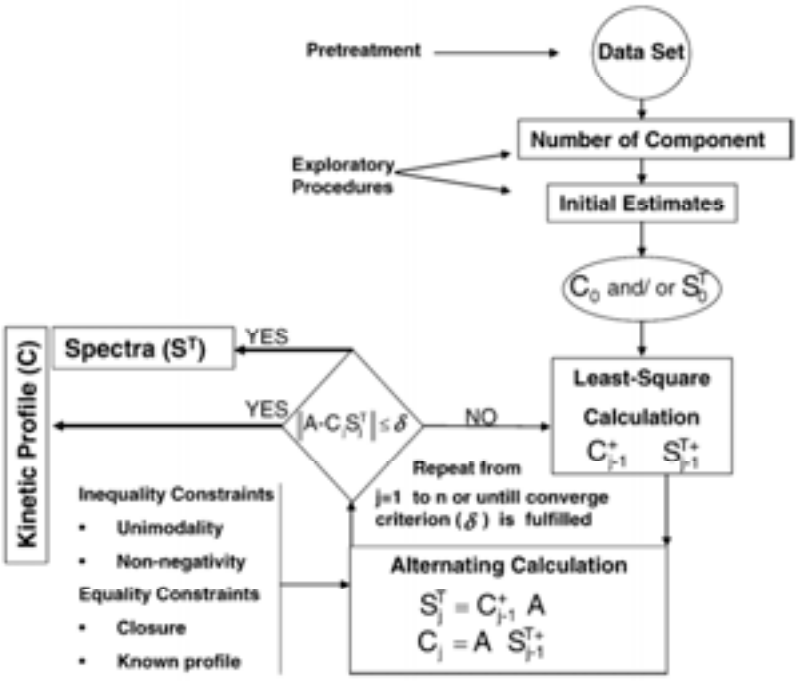


Fig. 1. Scheme of the alternating least-squares (ALS) algorithm.

A $\sin z$ value of zero means total agreement. Because the profiles are expected to be very similar, it is better to use a dissimilarity measure than a similarity measure as the cosine of the angle between vectors since sine values provide better discrimination than cosine values. A dissimilarity of 0.1 is equivalent to a correlation of 0.995 and one of 0.01 to a correlation of 0.9999.

If any chemical information obtained with a validated reference method is available, the quality of the results in the concentration domain can be checked via the coefficient of determination, R^2, which is a measure of closeness between the ALS and reference results.

3. Materials and methods

3.1. Microorganisms

S. cerevisiae ATCC 1326 strain was obtained from the American Type Culture Collection. The yeasts were maintained on YPD agar medium (1%, w/v yeast extract, 2%, w/v peptone and 2%, w/v glucose) at 4 °C.

3.2. Batch fermentation tests

Fermentation tests were performed in a 1 L bioreactor furnished with a double jacket through which water was circulated for temperature control and a silicone cap into which a sample collection device and an NIR probe were inserted. The bioreactor was placed on an electromagnetic stirrer and connected to a thermostated bath to maintain the working temperature at 25 °C (Fig. 2). Three different fermentation batches containing an initial glucose concentration of 200 g dm^{-3} but differing in pH were studied. The fermentation with a starting pH of 4 was used to construct the ALS model, and those starting at pH 3 and 5 were employed to validate the model. The culture medium used in the three contained 5 g dm^{-3} peptone, 5 g dm^{-3} yeast extract, and 3 g dm^{-3} malt extract. Yeasts were transferred from Petri dishes to the bioreactor with the aid of a Kohl handle.

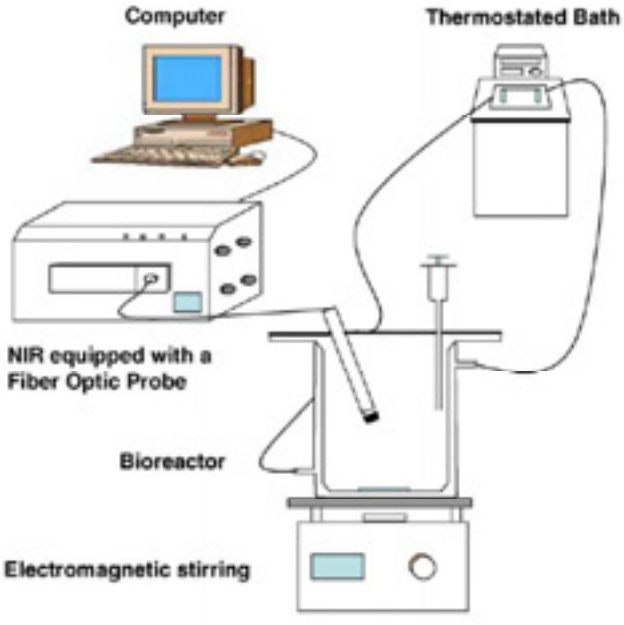


Fig. 2. Experimental set-up.

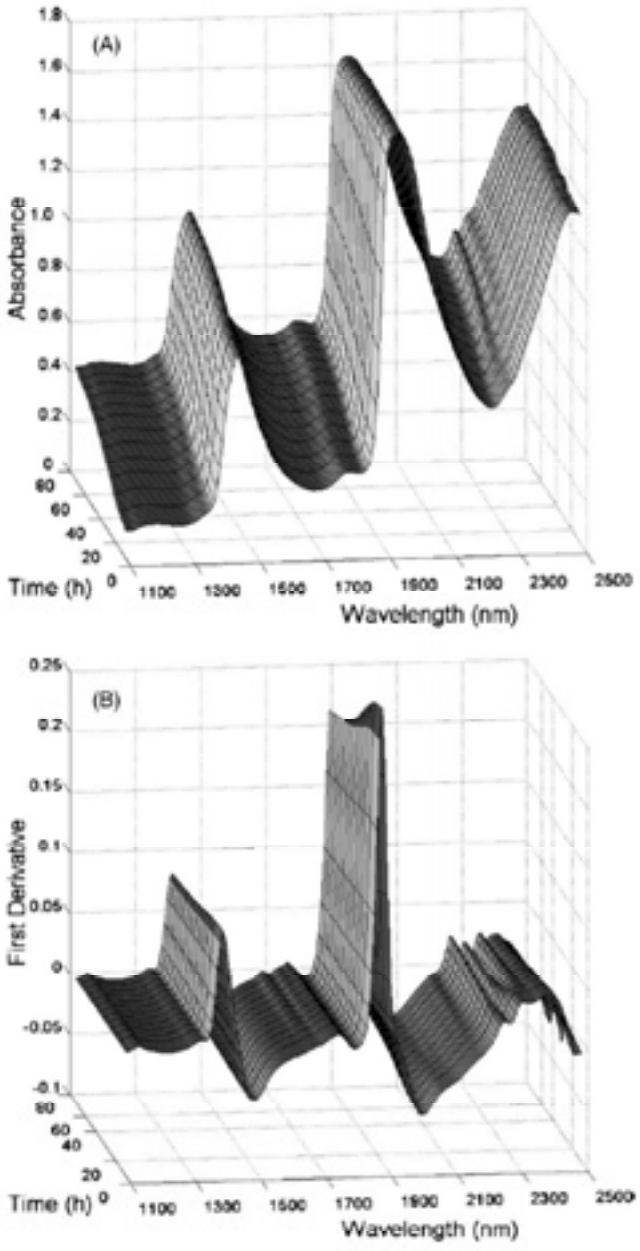


Fig. 3. Evolution of spectra during the fermentation process: (A) absorbance spectra; (B) first-derivative spectra.

Fermentation samples were withdrawn on a regular basis during the process. Glucose, ethanol and biomass were determined by using PLS models described elsewhere [18].

3.3. In-line recording or NIR spectra

In-line transflectance spectra were recorded on a Foss NIRSystems 5000 spectrophotometer equipped with a stainless steel optical probe. The instrument was controlled via the Vision v. 2.51 software package. Spectra were recorded at regular time intervals as the average of 32 scans performed at 2 nm intervals over the wavelength range 1100–2500 nm. Fig. 3 shows the variation of NIR spectra in the absorbance and first-derivative modes during a typical alcoholic fermentation.

3.4. Data processing

Experimental data were processed by using various functions included in Matlab v. 6.5. The ALS algorithm was applied to spectral data by using the software GUIPRO, from Gemperline [19], which runs under Matlab 6.0.

The chemical rank for each fermentation was determined from eigenvalues calculated by using singular value decomposition (SVD) and checking them for significance via an *F*-test based on the reduced error eigenvalue (REV) developed by Malinowski [20]. Decisions were also based on plots of the eigenvectors associated with each selected eigenvalue.

The initial estimated profiles were constructed by using the exploratory algorithm needle search also included in GUIPRO [21].

In calculating dissimilarity for glucose, the first spectrum of the fermentation batch, which was recorded immediately before the yeasts were added, was used as reference spectrum. For ethanol, the reference spectrum was the last spectrum for each batch. This was recorded once all the glucose had been depleted and the only studied analytes present in the medium were primarily ethanol, but also biomass which distorted the reference spectrum and made the result obtained were not as reliable as for glucose.

4. Results and discussion

4.1. Spectral regions

Changes during the fermentation process can be monitored by using an NIR probe to record spectra on a periodic basis (see Fig. 3). As can be seen from Fig. 3A, absorbance spectra shifted with time, mainly as a result of scattering due to biomass growth in the medium. The strong bands for water clearly observed at 1900 and 1420 nm conceal the spectral information for the other analytes. However, the first-derivative spectra (Fig. 3B) exhibit correct baseline shifts and expose such spectral information.

Based on spectral observations and the results of previous work [18], the study was focused on three different spectral regions, namely:

(a) 1100–2350 nm, which spanned the whole NIR spectrum studied. The final portion of the NIR region was excluded in order to avoid spectral noise from the recording probe.
(b) 1100–1800 nm, which included the first overtone (1420 nm) for the hydroxyl bond in water, glucose and ethanol, and also the first overtone (1660–1780 nm) for the C–H bond. First derivative treatment applied to this region corrected constant base line shift and revealed spectral changes due to the accumulation of ethanol

during the fermentation process; such changes were not evident in the absorbance spectral mode due to baseline shifts caused by biomass accumulation.

(c) 1800–2350 nm, which included the combinations for the hydroxyl and C–H bonds. This region exhibited the highest absorbances and hence the highest sensitivity (from 1950 to 2000 nm). First-derivative spectra exhibited a decrease in the signal at 2100 nm with time as a result of the gradual disappearance of glucose. Likewise, both first-derivative and absorbance spectra exhibited an increase in the analytical signal at 2200–2350 nm with time due to the production of ethanol.

Fibre-optic probes exhibit slight noise above 2200 nm that increases up to 2500 nm. This was the region where our system exhibited the strongest spectral changes, so it called for a compromise: the 2200–2350 zone was retained on the grounds of its high signal-to-noise ratio and the 2350–2500 nm zone excluded as it exhibited a considerably lower ratio.

4.2. Number of components

Table 1 lists the number of components (NCs) used in each spectral region studied, namely: two for all first-derivative spectra, five for absorbance spectra in the 1100–2350 and 1100–1800 nm regions, and three for absorbance spectra in the 1800–2350 region.

First-derivative spectra accounted for the analytes glucose and ethanol. When more than two components were used, these were associated with biomass and by-products of alcoholic fermentation resulting from yeast metabolism such as glycerine and acetic acid; however, the eigenvectors associated with the fourth and fifth components exhibited erratic profiles.

By way of example, Fig. 4 shows the right eigenvector plot associated with the first five components for the 1100–2350 nm region in the absorbance spectral mode. The right eigenvectors associated with components 1–3 echoed the variation of the concentrations of the analytes glucose, ethanol and biomass, respectively; on the other hand, those associated with components 4 and 5 exhibited no clear-cut trend and precluded a chemical interpretation. Therefore, the chemical rank for our system was 3.

The spectral range exhibiting the highest EV (99.97%) and lowest LOF (0.0645%) in the absence of equality constraints was that spanning the whole absorbance spectrum

Table 1
Number of components (NCs) and lack of fit (LOF) for the different spectral regions studied

Spectral region	Absorbance		First derivative	
	LOF	NC	LOF	NC
1100–2350	0.065	5	0.405	2
1100–1800	0.093	5	0.818	2
1800–2350	0.120	3	1.757	2

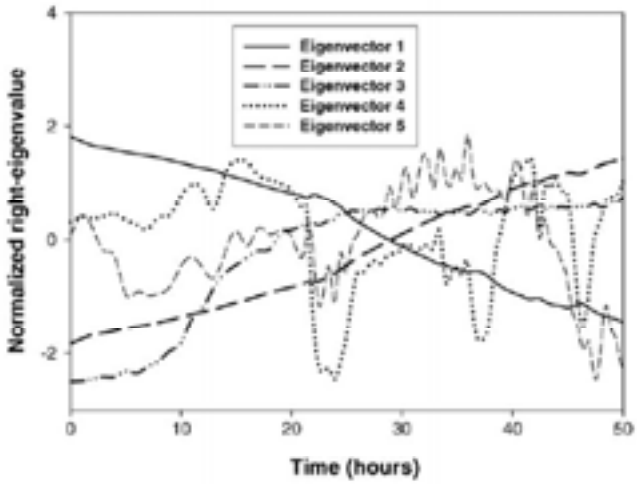


Fig. 4. Right eigenvector plot associated with the five most significant eigenvalues in the absorbance spectral mode.

(1100–2350 nm). This was the only region examined in all subsequent tests.

4.3. Constraints

The ALS algorithm was applied to absorbance spectra under the following inequality constraints: non-negativity and unimodality in the concentration domain, and non-negativity in the spectral domain.

Equality constraints were applied to the component glucose, which was the only analyte exhibiting a selective region at the beginning of the process before the yeast was added. The following four situations were examined:

(a) Using no spectral information.
(b) Assigning the first spectrum recorded, which corresponded to the culture medium immediately before inoculation of the yeasts, to the component glucose.
(c) Assigning the concentration profile provided by the EFA algorithm to the component glucose.
(d) Assigning the first spectrum of the batch and the glucose concentration profile provided by the EFA algorithm to the component glucose.

Table 2 shows the figures of merits $\sin z_{et}$, $\sin z_{glu}$ and LOF obtained in the previous four situations. The statistics were significantly greater in the two cases where information in the concentration domain was supplied as an equality constraint. This was a result of EFA estimates resemble true profiles in processes where intermediate products have profiles that exhibit an absolute maximum during the process (i.e. where they form from the starting analytes and are converted to other intermediate analytes or the end-products) in a higher degree than in fermentation systems. During alcoholic fermentation, however, both ethanol and biomass accumulate in the culture medium as they form, so their profiles are hyperboloidal or sigmoidal in shape.

Table 2
Lack of fit (LOF) and dissimilarity values for glucose ($\sin z_{glu}$) and ethanol ($\sin z_{et}$) as obtained by supplying different types of information as equality constraints to the MCR–ALS algorithm

Concentration equality constraint	Spectral equality constraint					
	None			First spectrum		
	LOF	$\sin z_{glu}$	$\sin z_{et}$	LOF	$\sin z_{glu}$	$\sin z_{et}$
None	0.0645	0.0014	0.0197	0.0646	0.0009	0.0201
Glucose	0.1688	0.1467	0.1800	0.1896	0.1654	0.2120

In the other two cases (viz. when no information or spectral information alone was supplied), the differences in the parameters were not significant, so the use of spectral information did not improve the resolution results. Because the statistics were not significantly improved by the introduction of an equality constraint, the ALS model involving no such constraint was adopted in order to avoid the burden of constraints.

4.4. Elucidation of the composition of the fermentation system

Fig. 5 shows the kinetic profiles for glucose, ethanol and biomass in a fermentation process conducted at an initial glucose concentration of 200 g dm^{-3}. The concentrations are on the real scale for the process as the ALS concentration profiles were corrected with the corresponding coefficients of linear regression (slope and offset) between the ALS values and those provided by the PLS reference model.

The coefficients of determination R^2 between the ALS and PLS values for glucose, ethanol and biomass were 99.73%, 99.55% and 99.15%. These values are quite acceptable and confirm that the ALS model provides good statistics despite its inherent rotational and scale ambiguities.

Fig. 6 shows the spectra for the three analytes. The spectrum associated with biomass formation, which is the origin of scattering in the culture medium, is highly similar to those

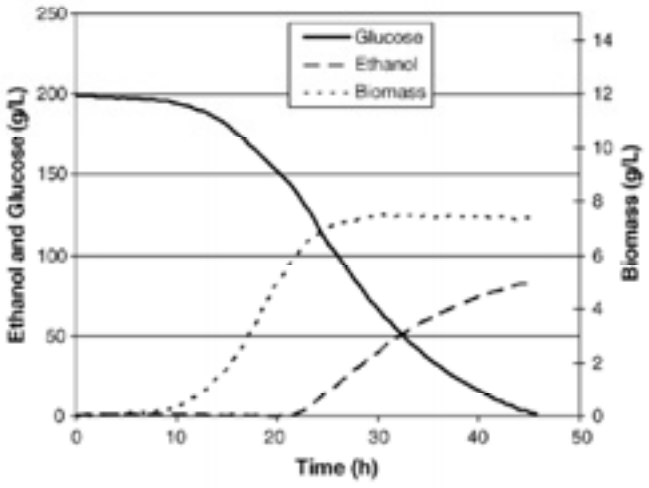


Fig. 5. Kinetic profiles for glucose, ethanol and biomass as determined by using the ALS model, all on a real scale.

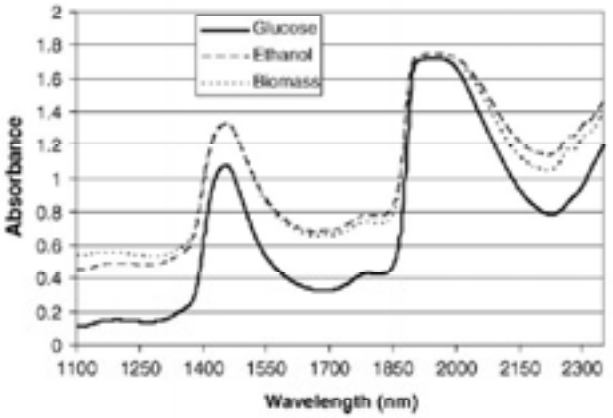


Fig. 6. Spectra associated with the glucose, ethanol and biomass profiles.

for glucose and ethanol. One should bear in mind that fermentation takes place in an aqueous medium and that its high absorbance dictates the spectral profiles for the three analytes.

The ALS model, obtained from the pH 4 experiment, was applied to two additional fermentation runs performed under identical conditions except for pH, which was 3 in one case and 5 in the other. The model was applied, by ordinary least-square, via the pseudo-inverse of its spectral data matrix. Concentration profiles predicted by the model for the two additional fermentations are shown in Fig. 7. As can be seen, the non-negativity and unimodality constraints imposed on the initial model were not strictly fulfilled by the predictions.

Table 3 shows intercept, slope and coefficient of determination between the analyte concentrations predicted by the model, for the two fermentations and those obtained by directly applying the ALS algorithm to the fermentation per-

Table 3
Figures of merit for the relationship between the concentration data provided by the ALS model and those obtained by applying the ALS method to fermentation runs performed at two different pH values

Analyte	Batch	Slope	Intercept	R^2
Glucose	pH 3	0.998 ± 0.001	−0.095 ± 0.001	0.999
	pH 5	0.998 ± 0.001	0.200 ± 0.001	0.999
Ethanol	pH 3	1.000 ± 0.005	−0.095 ± 0.006	0.998
	pH 5	0.949 ± 0.046	0.114 ± 0.033	0.994
Biomass	pH 3	0.959 ± 0.075	0.130 ± 0.048	0.959
	pH 5	0.949 ± 0.054	0.114 ± 0.031	0.962

Symbol ± denotes the 95% confidence interval.

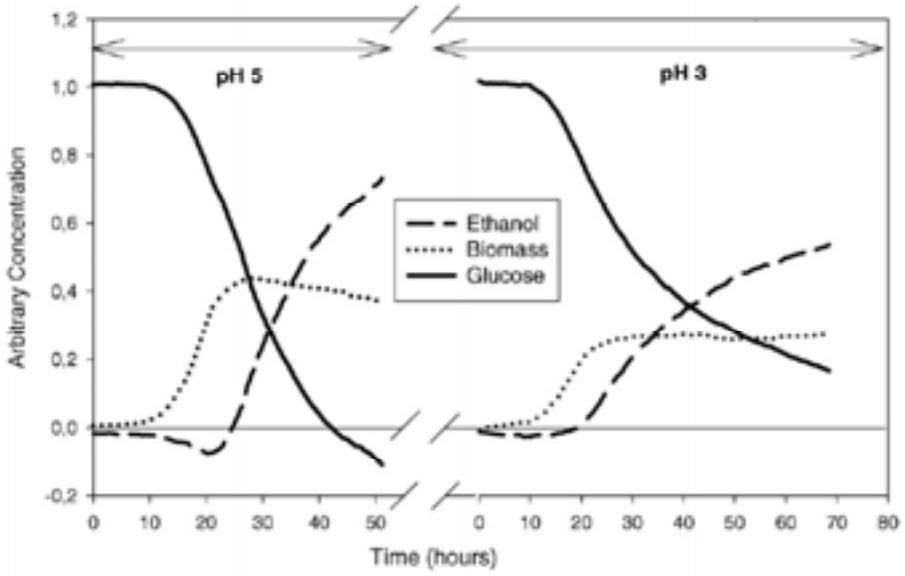


Fig. 7. Concentration profiles obtained by applying the ALS model to two fermentation runs performed at a different pH.

formed at pH 5 and 3 under the same conditions and constraints used to construct the model. Notwithstanding the uncertainties introduced by its ambiguities, the model provides highly significant correlations for the analytes glucose and ethanol, and significant correlations for biomass.

5. Conclusions

Despite the complexity of a fermentation medium, MCR–ALS was successfully used in this work to elucidate the composition profiles for glucose, ethanol, principal analytes involved in alcoholic fermentation. The results were also of good quality to model the evolution of biomass, non-absorbing specie, that promotes the scattering throughout the fermentation.

The results obtained for all the species were externally validated by comparison with a previously validated PLS reference method.

The proposed ALS model was applied to fermentation runs conducted under similar conditions and the results were found to be acceptable despite the ambiguities.

In future works we will use the combination of soft and hard modelling in order to minimize the ambiguities inherent to MCR–ALS algorithm.

Acknowledgements

The authors are grateful to Spain's DGICyT for funding this research within the framework of Project BQU2003–04247. Antonio C. Peinado wishes to acknowledge additional support from Spain's MCyT in the form of a grant, and the availability and assistance of Professor Paul Gemperline during his stay in East Carolina University.

References

[1] C. Berg, 2003. http://www.distill.com/berg/.
[2] S. Helle, A. Murria, J. Lam, D. Cameron, S. Duff, Biores. Technol. 92 (2004) 163.
[3] A. Costa, D. Atala, F. Maugeri, F. Maciel, Process. Biochem. 37 (2001) 125.
[4] F.L. Silva, M.I. Rodrigues, F. Maugeri, J. Chem. Technol. Biotechnol. 74 (1999) 176.
[5] A.C. Costa, L.A.C. Meleiro, R. Maciel Filho, Process. Biochem. 38 (2002) 743.
[6] A. Arnold, L. Harvey, B. McNeil, J. Hall, BioPharm. Int. 15 (2002) 26.
[7] D.A. Philips, D.L. Doak, Biotech. Prog. 15 (1999) 529.
[8] P. Fayolle, D. Picque, G. Corrieu, Food Contr. 11 (2000) 291.
[9] R. Bro, J. Chemom. 10 (1996) 47.
[10] J.A. Lopes, J.C. Menezes, Chemom. Intell. Lab. Syst. 68 (2003) 75.
[11] R. Tauler, B. Kowalski, S. Flemming, Anal. Chem. 65 (1993) 2040.
[12] J. Jiang, Y. Liang, U. Ozaki, Chemom. Intell. Lab. Syst. 71 (2004) 1.
[13] W.H. Lawton, E.A. Silvestre, Technometrics 13 (1971) 617.
[14] R. Tauler, A.K. Smilde, B.R. Kowalski, J. Chemom. 9 (1995) 31.
[15] A. de Juan, Y. Vander Heyden, R. Tauler, D.L. Massart, Anal. Chim. Acta 346 (1997) 307.
[16] M. Maeder, Anal. Chem. 59 (1987) 527.
[17] W. Windig, J. Guilment, Anal. Chem. 63 (1991) 1425.
[18] M. Blanco, A. Peinado, J. Mas, Biotech. Bioeng. 88 (2004) 536.
[19] P.J. Gemperline, E. Cash, Anal. Chem. 75 (2003) 4236.
[20] E.R. Malinowski, J. Chemom. 3 (1988) 49.
[21] P.J. Gemperline, J. Chem. Inf. Comput. Sci. 24 (1984) 206.

MONITORING ALCOHOLIC FERMENTATION BY JOINT USE OF SOFT- AND HARD-MODELLING METHODS

Blanco M., Peinado A. C., Mas J.

Analytica Chimica Acta. Accepted, September 29th, **2005**. *In press*

Available online at www.sciencedirect.com
SCIENCE DIRECT
Analytica Chimica Acta xxx (2005) xxx–xxx
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Monitoring alcoholic fermentation by joint use of soft and hard modelling methods

Marcelo Blanco [a,*], Antonio C. Peinado [a], Jordi Mas [b]

[a] Department of Chemistry, Faculty of Sciences, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[b] Department of Biology, Faculty of Sciences, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

Received 10 June 2005; received in revised form 28 September 2005; accepted 29 September 2005

Abstract

The strengths of hard and soft modelling were exploited by using both types of methods in combination to monitor alcoholic fermentations under *Saccharomyces cerevisiae* yeasts. Experimental work was performed in two steps. In the first, fermentation processes were conducted under identical conditions except for the initial glucose concentration in order to tests various previously reported empirical hard modelling methods. The product inhibition model of Hinshelwood was found to provide the best results in terms of goodness of fit and consistency in parameter values between runs under these conditions. In the second step, fermentation processes conducted at variable temperature and pH were monitored in-line by using an immersion NIR probe. The results were processed by using multivariate curve resolution–alternating least-squares (MCR–ALS) methodology in combination with the hard modelling information obtained in the first step as spectral equality constraints. Notwithstanding the complexity of the fermentation matrix introduced by variability in the species involved in yeast metabolism, the extracted profiles exhibited highly significant correlation with the reference values provided by a validated reference method for the determination of glucose, ethanol and biomass. The results testify to the efficiency of the joint use of soft and empirical hard modelling for studying evolving biological systems and opens up new avenues for application to other bioprocesses.

Keywords: NIR spectroscopy; Hard modelling; Soft modelling; ALS; PLS; In-line monitoring

1. Introduction

The dramatic rise in price and periodic shortages of petroleum, in addition to inevitable exhaustion of the existing fuel supply, have fostered research into the processes involved in the production of substances usable as alternative energy sources. Especially prominent in this respect is alcoholic fermentation as ethanol is being increasingly used as a substitute for gasoline [1]. Moreover, ethanol can be readily obtained from otherwise industrially useless waste such as plant residues and highly polluting pulping liquor [2]. This has led many countries to devise research programmes aimed at improving and optimizing alcoholic fermentation [3]. Despite the environmental advantages of ethanol as a fuel, it cannot be expected to be widely used as a substitute for traditional energy sources until a viable, economical, competitive production process becomes available [4]. In this respect, industrial alcoholic fermentation processes have traditionally been rather inefficient, as they have usually relied on the operator's long-life experience and knowledge to maintain the evolving system under control.

In his pioneering study of 1949 [5], Monod reported an empirical equation to describe cell growth as a function of a limiting factor. Since then, the mathematical modelling of fermentation processes has been a subject of much research in biotechnology [6] and a host of models describing alcoholic fermentation in kinetic terms have been reported. Such models can be formulated in various ways and range from structured models, [7,8] which consider a variety of phenomena influencing fermentation kinetics but require estimation of a number of parameters that are difficult to obtain, to black box-type models, which are based on readily measured parameters, but subject to severe restrictions, and fail when the constraints imposed are not met [9]. A wide range of empirical models derived from the mechanistic models used to study various types of enzymatic action also exists. Such empirical models, which have been widely used to model alcoholic fermentation, assume product formation to be related to cell growth and are in between structured models and black-

* Corresponding author.

0003-2670/$ – see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.aca.2005.09.066

Nomenclature

K_d	self-inhibition constant for growth (h^{-1})
K_{IP}	constant of substrate inhibition for ethanol production (g/L)
K_{IX}	constant of substrate inhibition for growth (g/L)
K_{PP}	constant of fermentation inhibition by ethanol (g/L)
K_{PX}	constant of growth inhibition by ethanol (g/L)
K_{SP}	constant in the product term for ethanol production (g/L)
K_{SX}	constant in the substrate term for growth (g/L)
P	ethanol concentration (g/L)
$P_{P,max}$	constant of maximum product formation in the fermentation term (g/L)
$P_{X,max}$	constant of maximum product formation in the growth term (g/L)
S	glucose concentration (g/L)
$S_{P,max}$	constant of maximum substrate consumption in the fermentation term (g/L)
$S_{X,max}$	constant of maximum substrate consumption in the growth term (g/L)
X	cell concentration (g (d.m.)/L)
$Y_{p/s}$	ratio of ethanol produced per substrate consumed for fermentation
$Y_{x/s}$	ratio of cell produced per substrate consumed for growth

Greek symbols

μ_{max}	maximum specific growth rate (h^{-1})
ν_{max}	maximum specific fermentation rate (h^{-1})

box models in terms of complexity and accuracy. Although the adoption of a modelling approach can provide a better understanding of a fermentation system, it does not ensure efficient alcoholic fermentation; in fact, maintaining the process under optimum conditions throughout entails using an analytical and monitoring approach.

Near infrared spectroscopy (NIRS) is a rapid, non-destructive technique of use to record changes in strongly absorbing, highly light-scattering matrices such as cell-culture media. Most research into fermentation systems has involved off-line work and only in a few cases has on-line monitoring been addressed [10].

The current widespread acceptance and success of NIRS would never have been possible without the simultaneous development of multivariate calibration methods for extracting analytically useful information from NIR signals. Partial least-squares' regression (PLSR) is no doubt the most thoroughly checked and widely used multivariate calibration method. NIRS has been used in combination with PLSR models to predict the concentrations of various analytes in the course of different types of fermentation; however, the ensuing models require the use of reference data provided obtained with classical analytical techniques.

There is, however, a family of chemometric methods known as self-modelling curve resolution (SMCR) [11], which factorizes a mixed instrumental signal into the pure contributions associated to each component in a system [12]. These methods provide a powerful approach with very modest demands; they require no reference analytical information, but are subject to inherent intensity and rotational ambiguities [13] that entail imposing non-negativity, unimodality or closure constraints [14] to shorten the range of feasible solutions [15]. Even with such constraints, however, a band of feasible profiles rather than a unique solution fitting equally well the experimental data and fulfilling the physical and chemical constraints on the system has to be considered [16]. The range of feasible solutions can be minimized and optimized by using various procedures [17]. One of the latest, most powerful and efficient choices in this respect is the use of a mixed approach introducing a hard modelling step based on a kinetic model as an additional constraint in the multivariate curve resolution–alternating least squares (MCR–ALS) algorithm [18]. This mixed approach has so far been used to solve kinetic problems [19], determine reaction rate constants [20], monitor batch chemical processes [21] and even improve analyte quantitation in the titration of complex mixtures [22], but never in combination with MCR–ALS to minimize ambiguities in fermentation systems.

The purpose of this work was to test various hard empirical methods widely used to model alcoholic fermentation with a view to identifying that best fitting the experimental results provided by a validated reference method for the determination of glucose, ethanol and biomass. The fitted results provided by the optimum model for the three analytes were incorporated as equality constraints into the MCR–ALS soft modelling algorithm, which was used to monitor alcoholic fermentation processes conducted under variable pH and temperature conditions.

2. Theory

2.1. Empirical models

A number of empirical models based on the mechanistic laws of kinetics have been proposed to describe alcoholic fermentation batches. Such models use the following system of differential equations to describe cell growth, fermentation and substrate uptake, respectively:

$$\frac{dx}{dt} = \mu X \quad (1)$$

$$\frac{dP}{dt} = \upsilon X \quad (2)$$

$$\frac{dS}{dt} = -\left(\frac{1}{Y_{x/s}}\frac{dX}{dt}\right) - \left(\frac{1}{Y_{p/s}}\frac{dP}{dt}\right) \quad (3)$$

The model assumes that the variation in substrate consumption as a function of time, $\frac{dS}{dt}$, is inversely related to the rate of change of biomass, $\frac{dX}{dt}$, and ethanol, $\frac{dP}{dt}$, being $Y_{x/s}$ the ratio of cell produced per substrate consumed for growth and $Y_{p/s}$ the ratio of ethanol produced per substrate consumed for fermentation.

Table 1
Coefficients for substrate uptake (μ) and product formation (ν) used by various empirical models

Number	μ	ν	References
1	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)$	Monod [5]
2	$\mu_{max}\left(1-\exp-\frac{S}{K_{SX}}\right)$	$\nu_{max}\left(1-\exp\left(-\frac{S}{K_{SP}}\right)\right)$	Teissier [36]
3	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)\exp\left(-\frac{S}{K_{IX}}\right)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)\exp\left(-\frac{S}{K_{IP}}\right)$	Edwards [37]
4	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)\left(1-\frac{S}{S_{X,max}}\right)^{n}$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)\left(1-\frac{S}{S_{P,max}}\right)^{n}$	Luong [38]
5	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)(1-K_{PX}P)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)(1-K_{PP}P)$	Hinshelwood [39]
6	$\mu_{max}\left(1-\frac{P}{P_{X\,max}}\right)$	$\nu_{max}\left(1-\frac{P}{P_{P\,max}}\right)$	Ghose and Tyagi [40]
7	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)\exp(-K_{PX}P)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)\exp(-K_{PP}P)$	Aiba et al. [41]

While the rates of change, as a function of time, for biomass (X), and product formation (P) are directly related to the biomass via appropriate forms of μ (the specific growth rate, in h^{-1}) and ν (the specific fermentation rate, also in h^{-1}), where h stands for hours. Models differ in the way the specific fermentation and growth rates are defined. Table 1 lists some of the most widely used forms. Eqs. (1) and (2) in the table represent inhibition-free kinetics; (3) and (4) include substrate inhibition effects; and (5)–(7) include product inhibition kinetics.

The way the parameters in these equations are determined has evolved hand-in-hand with the development of powerful, fast computers. The methods traditionally used for this purpose (e.g. linearization, polynomial approximations [23]) pose numerical problems as they involve the reciprocals or ratios of experimentally determined variables, which influences the precision with which the target parameters can be calculated. These problems can be avoided by direct simultaneous integration of the differential equations for growth, fermentation and substrate uptake [24]. The results thus obtained are compared with experimental data and an optimization procedure is used to change the values of the model parameter until the difference between the experimental and calculated data is minimized.

2.2. MCR–ALS

Curve resolution methods are based on the factor analysis technique, which factorizes the experimental data matrix, **A** ($r \times c$), into the following matrix product:

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \quad (4)$$

where **C** ($r \times n$) has column vectors containing the temporal variation of the analyte concentrations (i.e. the kinetic profiles); the row vectors of $\mathbf{S}^{\mathrm{T}}$ ($n \times c$) contains the variation of the response with respect to different wavelengths (i.e. the spectral profiles); **E** ($r \times c$) is the residual matrix, which contains the residual variation of the data due to the analyte and is assumed to be independent and exhibit a constant variance; r, number of row, is the dimension related to the temporal variation of the system, which coincides with the number of measured samples; c, number of columns, is equal to the number of wavelength and n is the dimension related to the number of analytes the concentrations of which change with time and alter the detector signal.

Before a SMCR method can be applied, the chemical rank or number of components of the target system must be determined. This can be done by using various methods [25] including MCR–ALS, which is based on an iterative optimization algorithm [13] that is applied in five steps. The procedure is as follows:

(1) Initial concentration ($\mathbf{C}_0$) or spectral profiles ($\mathbf{S}_0$) assumed to constitute good approximations to the pure components can be used as starting point. Furthermore, other resolution methods well established such as needle search [26], evolving factor analysis (EFA) [27] or a suitable alternative [28] can also be employed.
(2) Given some initial or intermediate estimate of **C**, the least-squares solution (i.e. $\mathbf{S}^{\mathrm{T}} = \mathrm{pinv}(\mathbf{C})^{*}\mathbf{A}$), is found by minimizing $||\mathbf{A} - \mathbf{C}\mathbf{S}^{\mathrm{T}}||$, being $||\ ||$ the Euclidean norm (i.e. squared root of the sum of the squared values).
(3) Given some initial or intermediate estimate of **S**, the least-squares' solution, (i.e. $\mathbf{C} = \mathbf{A}^{*}\mathrm{pinv}(\mathbf{S}^{\mathrm{T}})$), is found by minimizing $||\mathbf{A} - \mathbf{C}\mathbf{S}^{\mathrm{T}}||$.
(4) **A** is reproduced from **C** and $\mathbf{S}^{\mathrm{T}}$. The residual matrix and its norm is calculated.
(5) The alternating least-squares steps (2–4) are repeated until the convergence criterion (δ) is fulfilled (i.e. $||\mathbf{A} - \mathbf{C}\mathbf{S}^{\mathrm{T}}|| < \delta$) or the maximum number of iterations is reached.

The performance of the model can be assessed in generic terms by using only two parameters, namely: the explained variance, EV, and the statistic lack of fit, LOF, which are defined as follows:

$$\mathrm{EV} = \frac{\sum_i \sum_j \hat{a}_{ij}^2}{\sum_i \sum_j a_{ij}^2} \times 100 \quad (5)$$

$$\mathrm{LOF} = \sqrt{\frac{\sum_i \sum_j (a_{ij} - \hat{a}_{ij})^2}{\sum_i \sum_j a_{ij}^2}} \times 100 \quad (6)$$

a_{ij} denoting the elements in the experimental matrix **A** and $\hat{a}_{ij}$ the values calculated by the using the model with Eq. (4).

Provided the concentrations obtained with the validated references method are available, the quality of the results in the concentration domain can be checked via the coefficient of determination, R^2, which is a measure of closeness between the ALS and reference results.

Resolving matrix **A** is no easy task. Curve resolution methods cannot deliver a unique solution as they are subject to both rotational and scale (intensity) ambiguities. The rotational ambiguity,

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{C}\mathbf{T}\mathbf{T}^{-1}\mathbf{S}^{\mathrm{T}} = (\mathbf{C}\mathbf{T})(\mathbf{T}^{-1}\mathbf{S}^{\mathrm{T}}) = \mathbf{C}^{*}\mathbf{S}^{\mathrm{T}*} \quad (7)$$

reflects the fact that, for any factoring scheme, an infinite number of non-singular (i.e. invertible) transformation rotation matrices **T** ($n \times n$) exist that make the solution non-unique. On the other hand, the intensity ambiguity,

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} = (b\mathbf{C}^{*})\left(\frac{1}{b}\mathbf{S}^{\mathrm{T}*}\right) = \left(b \times \frac{1}{b}\right)(\mathbf{C}^{*}\mathbf{S}^{\mathrm{T}*}) = \mathbf{C}^{*}\mathbf{S}^{\mathrm{T}*} \quad (8)$$

allows one to extract a scalar b from matrix **C** and its inverse from matrix $\mathbf{S}^{\mathrm{T}}$ with any factoring scheme; however, $\mathbf{C}^{*}$ is b times **C** and $\mathbf{S}^{\mathrm{T}*}$ is $1/b$ times $\mathbf{S}^{\mathrm{T}}$, but $\mathbf{C}^{*}$ and $\mathbf{S}^{\mathrm{T}*}$ are identical in shape with **C** and $\mathbf{S}^{\mathrm{T}}$, respectively.

In order to avoid the previous ambiguities, the target system must be subjected to some constraints on the concentration and/or spectral profiles. The constraints used in MCR methods are based on available information about the target system and can be of two different types, namely: natural and equality constraints. Natural constraints include (a) non-negativity, which can be applied to both the concentration domain and the spectral domain – provided both are positive –; (b) unimodality, which can be applied to the analyte concentration profile as long as it exhibits a single maximum; (c) the closure or mass balance constraint, which is applicable when the combined concentration of the analytes remains constant and known throughout the process. Equality constraints are used when the kinetic and/or spectral profile for some species is known. Recently, Gemperline developed the alternating least-squares with penalty functions (p-ALS) algorithm by introducing the penalty function into the ALS algorithm in order to soften or harden the constraints [29]. "Hard" constraints are constraints that are strictly enforced, whereas "soft" constraints allow small deviations from constrained values.

3. Materials and methods

3.1. Microorganisms

Saccharomyces cerevisiae ATCC 1326 strain yeasts were obtained from the American Type Culture Collection. The yeasts were maintained on YPD agar medium (1%, w/v yeast extract, 2%, w/v peptone and 2%, w/v glucose) at 4 °C.

3.2. Batch fermentation runs

Fermentation runs were conducted in a 1 L bioreactor furnished with a double jacket through which water was circulated for thermostating and a silicone cap into which a sample collection device and an NIR probe were inserted. The bioreactor was placed on an electromagnetic stirrer and connected to a thermostated bath to maintain the operating temperature. Processes were run without aeration. A total of nine fermentation runs were performed as follows:

(a) Three runs were conducted at 30 °C, an initial pH of 4 and three different levels of initial glucose concentration (200, 100 and 50 g/L). These runs were used to identify the empirical model best fitting the experimental system. Also, the fermentation performed at a 200 g/L initial glucose concentration was used to establish the equality constraints for the three analytes.
(b) The other six runs were conducted at a 200 g/L initial glucose concentration, two different temperatures (25 and 35 °C) and three levels of initial pH (3, 4 and 5). These runs were monitored by incorporating the equality constraints into the MCR–ALS algorithm.

In addition to glucose, the culture medium used in the fermentation tests contained 5 g/L peptone, 5 g/L yeast extract and 3 g/L malt extract. The initial volume of medium was 800 mL and 1% w/v inoculum grown overnight was used.

3.3. In-line recording of NIR spectra

Transflectance spectra were recorded in-line by using a Foss NIRSystems 5000 spectrophotometer equipped with a stainless steel optical probe the light path of which was adjusted in order to obtain appropriate absorbance measurements and maintained throughout the tests. The spectrophotometer was governed via the Vision v. 2.51 software package. Spectra were recorded at 30 min intervals as the average of 32 scans performed at 2 nm intervals over the wavelength range 1100–2500 nm. Fig. 1 shows the variation of the NIR spectra in the absorbance and first-derivative modes during a typical alcoholic fermentation run.

3.4. Reference methods

Concentration of glucose, ethanol and biomass were determined using previously validated PLS models. These models were built by combining NIR spectra of lab samples along with fermentation samples in order to span the variability inherent to the fermentation process. The model construction is described elsewhere [30].

3.5. Data processing

Empirical models were constructed by using the software Berkeley–Madonna [31], which computes the parameter values for each model by simultaneous numerical integration of the differential equation system. The integration process is cou-

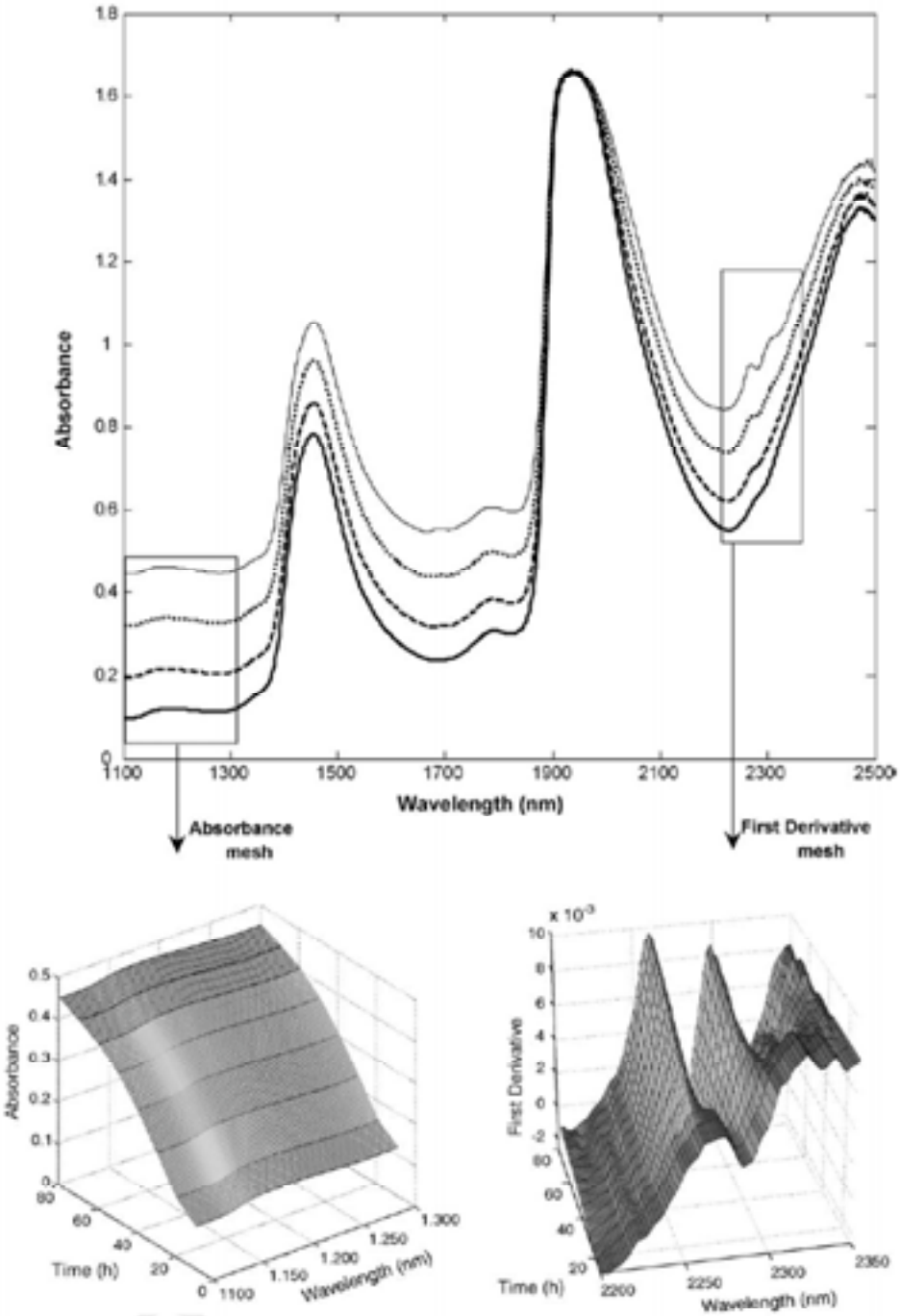


Fig. 1. Main figure show absorbance spectra taken just at the beginning of the fermentation (—), at the first third part (– –), at the second third part (· · ·), and at the end (—). Absorbance mesh detail in the 1100–1300 nm region. First derivative mesh detail in the 2200–2350 nm region. Run performed at 30 °C and pH 4.

pled to a simplex-optimization routine that compares the results obtained from the integration with their experimental counterparts and feedbacks the integration step until the difference between experimental and integrated data is minimized. We chose to use the Runge–Kutta method for integration on the grounds of its widely-documented efficiency with stiff systems [32].

The ALS algorithm was applied to spectral data by using the software GUIPRO, from Gemperline [29], which runs under Matlab v. 6.0 and higher.

The chemical rank of the fermentations was determined by using the eigenvalue F-test and plots of the eigenvectors associated to each eigenvalue as described in a previous paper [33].

The initial estimated profiles were constructed by using the needle search exploratory algorithm included in GUIPRO [34].

3.6. Combining soft- and hard-modelling

The present study involved two steps. In the first, various empirical models were tested to identify the best fitting for the

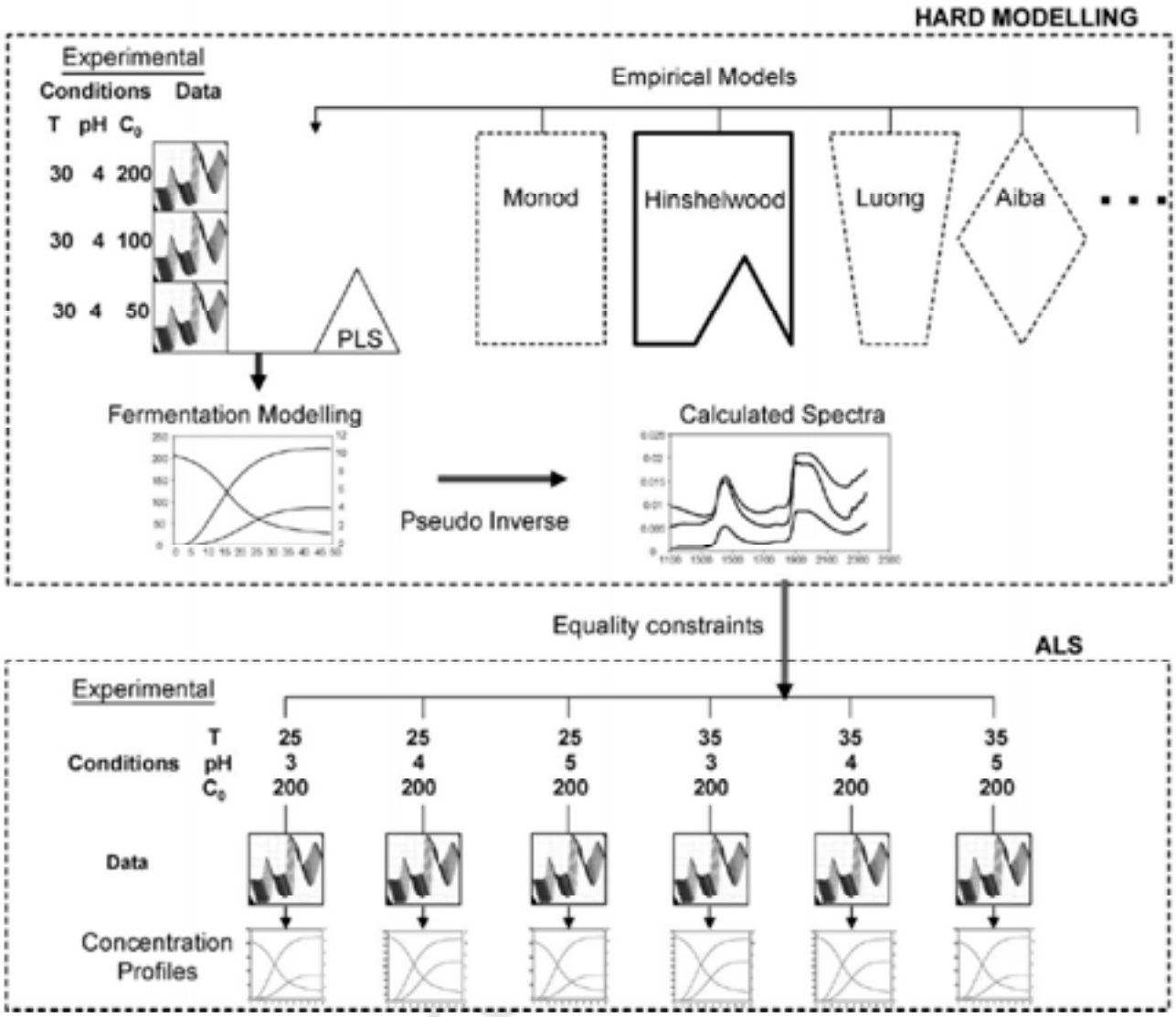


Fig. 2. Scheme of the process combining soft and hard modelling via equality constraints.

experimental results. Concentration profiles were obtained by applying the previously selected model to fermentation runs conducted at variable glucose concentrations. In the second step, the concentration profiles previously obtained were used to calculate the corresponding spectra via the pseudo-inverse, pinv

$$\mathbf{S}^{\mathrm{T}} = \mathrm{pinv}(\mathbf{C})^{*}\mathbf{A} \quad (9)$$

The calculated spectra were utilized as equality constraints in the ALS algorithm to monitor fermentation runs performed at different temperature and pH values. Fig. 2 depicts the procedure used to combine soft and hard modelling.

4. Results and discussion

4.1. Spectra

Fig. 1 shows the temporal variation of the NIR spectra in the absorbance mode during the fermentation process. The following regions are worth special note:

(a) 1100–1300 nm, where the spectral baseline is shifted as the primary result of dispersion through biomass growth and accumulation in the culture medium.
(b) Those around 1420 and 1950 nm, which exhibit the absorption maxima corresponding to the first overtone and tone combination for the O–H bond in water. These strong, broad bands conceal useful spectral information for functional groups of other species present in the reactor.
(c) 2200–2350 nm, which exhibits slight spectral changes with time that can be ascribed to ethanol. The changes are more apparent in the expanded region of the first-derivative spectrum.

4.2. Selection of an empirical model

The results obtained in the fermentation runs conducted at pH 4, 30 °C and three different initial glucose concentrations (200, 100 and 50 g/L) were used to model biomass formation, ethanol production and glucose. Performance curves were fitted

Table 2
Parameter values for the Hinshelwood model at initial glucose concentrations of 200, 100 and 50 g/L

Model parameters	Initial glucose concentration (g/L)		
	200	100	50
μ	0.153	0.132	0.137
K_{SX}	36.121	17.748	10.792
K_{PX}	0.070	0.098	0.128
ν	0.310	0.297	0.354
K_{SP}	22.851	26.211	16.514
K_{PP}	0.010	0.010	0.011
Y_{XS}	0.544	0.215	0.244
Y_{PP}	0.411	0.453	0.476

to the experimental data by using all the models tested. Models were validated according to two qualitative criteria, namely: (a) consistency between the parameter values at different initial substrate concentrations in the fermentation medium; (b) the degree of fitting in the transition from the exponential to the steady-state phase—where the growth rate decreased very strongly over a short interval.

All models in Table 1 were checked and transcribed to the language used by the software Berkeley–Madonna. Hinshelwood model was the one exhibiting the lowest standard deviation of the residuals for the three analytes and also that best meeting the above-described criteria. All other models were discarded as they resulted in inconsistencies between values for the fitting parameters (e.g. negative values) or fermentation runs.

Only the results obtained with the selected model are reported here to summarize the computations done in this work. Table 2 shows the parameter values for the fermentation runs performed at an initial glucose concentration of 200, 100 or 50 g/L. As can be seen, the values were consistent between fermentation runs and with previously reported data [24,35].

Fig. 3A–C show the analytical data obtained for glucose, ethanol and biomass (symbol lines), and the fitted results provided by the Hinshelwood model (solid lines), at an initial glucose concentration of 200, 100 and 50 g/L, respectively. As can be seen, fitting was close in the exponential growth stage, and very good in the latency and steady-state stages; also, transitions occurred in a gradual, continuous manner.

The fermentation run conducted at an initial glucose concentration of 200 g/L was additionally used to obtain the spectral and concentration profiles. Empirical spectra used as constraints were calculated from the pseudo-inverse of the concentration profiles obtained with the Hinshelwood model and the experimental spectral data matrix. Fig. 4 shows the calculated spectra for the fermentation run performed at a 200 g/L glucose concentration. These spectra are very similar to that for water, which conceals the spectral information for glucose and ethanol; however, the weak bands for ethanol at 2200 and 2350 nm distinguish it from the spectrum for glucose. Also worth noting is the difference in the biomass spectrum – which is not an actual spectrum as it contains some scattered light – over the range 1100–1300 nm, where none of the species present absorbs.

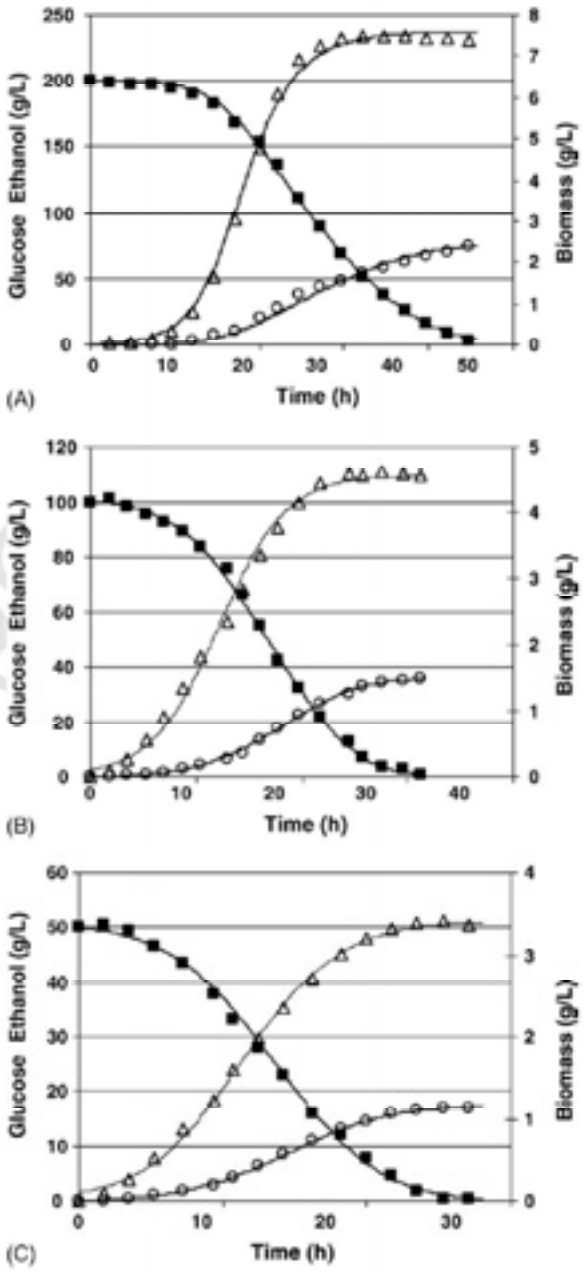


Fig. 3. Fitting of the Hinshelwood model to the reference values for the fermentation runs conducted at an initial glucose concentration of (A) 200, (B) 100, and (C) 50 g/L. Reference values are represents as ■ (for glucose) ○ (for ethanol) and △ (for biomass). Solid lines represent the model response.

4.3. MCR determination based on hard modelled information

The rotational and intensity ambiguities in the ALS algorithm preclude the obtainment of a unique solution. In order to minimize both, the empirical spectra were incorporated as equality constraints into the algorithm. The concentration profiles could not be used as equality constraints as they were time-dependent and each fermentation run lasted a different time.

Both single spectra and combinations of two or the three were tested. Selective application of the constraints to one or two ana-

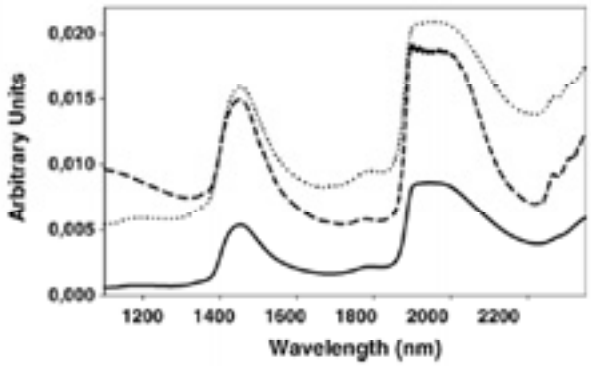


Fig. 4. Spectra for glucose (—), ethanol (···) and biomass (– –), as calculated using the proposed Hinshelwood model.

lytes resulted in profiles for the constrained species that spanned a scale 10^2 to 10^3 larger than that for the unconstrained species, i.e. the equality constraints modulated the spectral profiles for the constrained species relative to those under no equality constraint. In order to avoid such massive differences in scale, equality constraints were applied simultaneously to the three analytes.

Fig. 5 shows the concentration profiles for glucose, ethanol and biomass as obtained from the analysis of the six fermentation batches following individual application of the ALS algorithm to each. Using equality constraints on the spectra afforded the analysis of fermentation runs involving different durations' as the time dimension was unconstrained. Table 3 shows the correlation, as a coefficient of determination, between the concentration profiles estimated by the ALS model (Fig. 5) and those obtained with the validated PLS reference method; the goodness of this statistic for the three analytes—biomass included—testifies to the robustness gained by incorporating equality constraints into

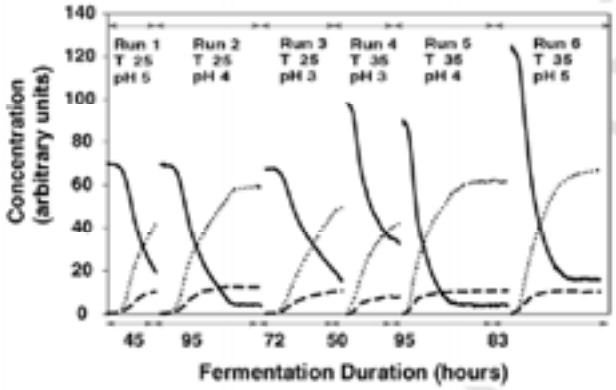


Fig. 5. Concentration profiles obtained by applying the ALS model with equality constraints to six fermentation runs conducted at two different temperatures (25 and 35 °C) and three pH values (3–5). Spectra for glucose (—), ethanol (···) and biomass (– –).

Table 3
Coefficients of determination between the reference PLS values and those provided by the ALS model

Conditions		Determination coefficient		
Temperature	pH	Glucose	Ethanol	Biomass
25	5	0.9999	0.9983	0.9937
25	4	0.9945	0.9943	0.9808
25	3	0.9960	0.9949	0.9853
35	3	0.9970	0.9992	0.9764
35	4	0.9930	0.9984	0.9705
35	5	0.9994	0.9997	0.9882

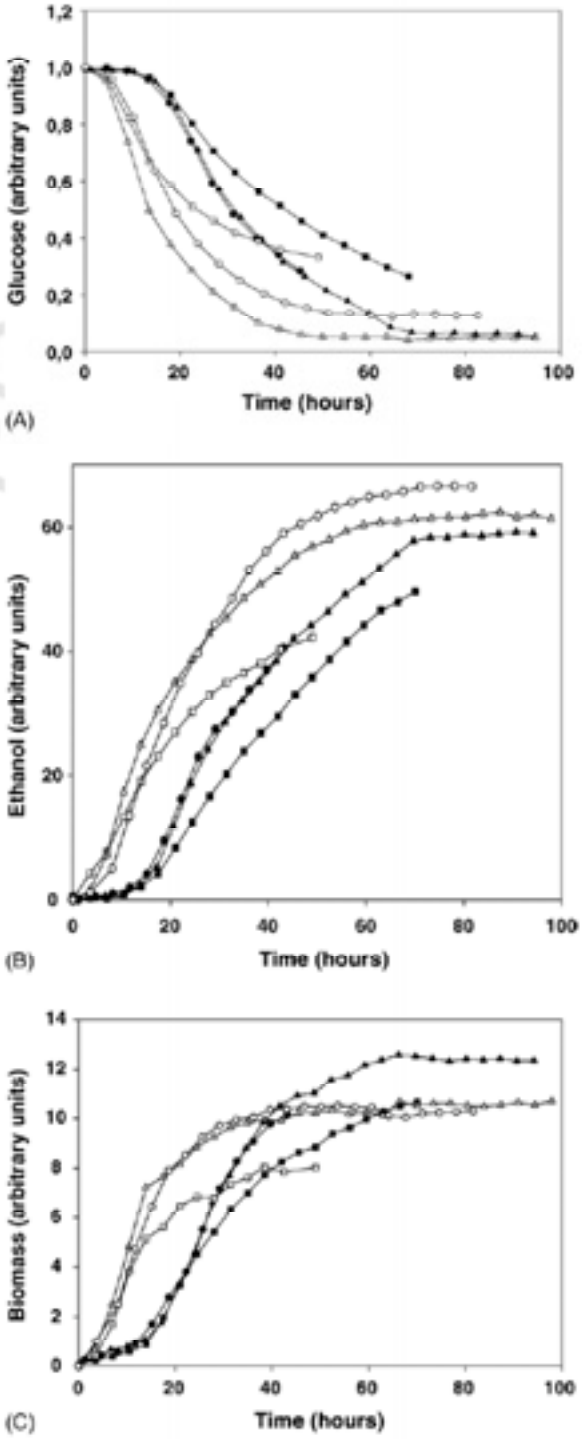


Fig. 6. Concentration profiles for: (A) glucose, (B) ethanol and (C) biomass in fermentation runs performed at two different temperatures, 25 °C (black symbols) and 35 °C (white symbols) and three pH values 3 (□), 4 (△) and 5 (○).

the algorithm. Also, the variance explained by the model (EV) was always greater than 99.9% and the lack of fit (LOF) smaller than 0.1%.

Fig. 6A–C show the concentration profiles for glucose, ethanol and biomass obtained with the ALS model. In order to facilitate comparison, the profile for glucose (Fig. 6A) was scaled between 1 and 0 by dividing the calculated profile for each fermentation run into the initial value as the initial glucose concentration was the same (200 g/L) in all runs. Fig. 6B and C show the profiles estimated by the ALS model.

An overall analysis of the figures reveals a strong influence of temperature: Glucose was taken up at an early stage at 35 °C while at 25 °C exhibited a latency period of nearly 10 h with virtually no uptake; Biomass evolved similarly where initial slope of the curve (Fig. 6C) was greater at 35 °C than 25 °C, where the slope was close to zero; for ethanol this behavior was even more apparent (Fig. 6B).

The pH of the fermentation medium also influenced the results, albeit not so strongly as the temperature. As can be seen from Fig. 6A, the variation of glucose uptake at 25 °C was very similar at pH 4 and 5; in fact, the two curves overlap. On the other hand, the curve obtained at the same temperature but pH 3 clearly departs from the previous two, with a much smaller slope. This was also the case at 35 °C: the curves obtained at pH 4 and 5 are mutually similar but clearly different from that obtained at pH 3. The effect is also clearly apparent in the ethanol production and biomass profiles. Thus, ethanol production at a given temperature was lower at pH 3 than at pH 4 and 5. It should be noted that pH 3 is an extreme, near-lethal value for *Saccharomyces* yeasts.

5. Conclusions

A comparative study of alcoholic fermentation under similar pH and temperature conditions but different initial concentrations of glucose allowed various empirical models to be tested with a view to identifying that best describing this evolving system. The linear product inhibition model of Hinshelwood was found to be that providing best fitting and thus selected for subsequent work.

The profiles obtained with the Hinshelwood model were incorporated as equality constrains in the spectral domain into the ALS algorithm. This combination of soft and hard modelling was successfully applied to fermentation runs conducted under different pH and temperature conditions. The goodness of the results for the three analytes studied was quantified in terms of the coefficient of determination between the profiles provided by the model and the reference values obtained with previously validated analytical methods.

The profiles thus obtained and a comparative study of the results for the different fermentation batches allowed us to establish the effects of pH and temperature on glucose uptake, and ethanol and biomass production.

As shown in this work, the combined use of soft modelling and hard empirical models facilitates the study of bioprocesses and biological systems.

Acknowledgements

The authors are grateful to Spain's DGICyT for funding this research within the framework of Project BQU2003-04247. Antonio C. Peinado wishes to acknowledge additional funding from Spain's MCyT in the form of a grant, and the availability and assistance of Professor Paul Gemperline during his stay at East Carolina University.

References

[1] S. Ueda, C.T. Zenin, D.A. Monteiro, Y.K. Parker, Biotech. Bioeng. 23 (1981) 291.
[2] S. Helle, A. Murray, J. Lam, D. Cameron, S. Duff, Bioresource Technol. 92 (2004) 163.
[3] A.D. Wheals, L.C. Basso, D.M.G. Alves, H.V. Amorim, Tibtech. 14 (1999) 482.
[4] A. Costa, D. Atala, F. Maugeri, R. Maciel, Process Biochem. 37 (2001) 125.
[5] J. Monod, Ann. Rev. Microbiol. 3 (1949) 371.
[6] M.R. Marín, Am. J. Enol. Viticult. 50 (1999) 166.
[7] U. Veeramallu, P. Agrawal, Biotech. Bioeng. 36 (1990) 694.
[8] F. Lei, M. Rotboll, S.B. Jorgensen, J. Biotech. 88 (2001) 205.
[9] I.C. Trelea, M. Titica, S. Landaud, E. Latrille, G. Corrieu, A. Cheruy, Math Comput. Simulat. 56 (2001) 405.
[10] A.G. Cavinato, D.M. Mayes, Z. Ge, J. Callis, Anal. Chem. 62 (1990) 1977.
[11] W.H. Lawton, E.A. Silvestre, Technometrics 13 (1971) 617.
[12] N.K.M. Faber, R. Bro, P.K. Hopke, Chemom. Intell. Lab. Syst. 65 (2003) 119.
[13] R. Tauler, A. Smilde, B. Kowalski, J. Chemom. 9 (1995) 31.
[14] A. De Juan, Y. Vander Heyden, R. Tauler, D.L. Massart, Anal. Chim. Acta 346 (1997) 307.
[15] P.J. Gemperline, Anal. Chem. 71 (1999) 5398.
[16] R. Tauler, Chemometrics 15 (2001) 627.
[17] B.M. Kim, R.C. Henry, Chemom. Intell. Lab. Syst. 49 (1999) 67.
[18] A. De Juan, M. Maeder, M. Martínez, R. Tauler, Chemom. Intell. Lab. Syst. 54 (2000) 123.
[19] H. Haario, V. Taavitsainen, Chemom. Intell. Lab. Syst. 44 (1998) 77.
[20] S. Bijlsma, H.F.M. Boelens, H.C.J. Hoefsloot, A.K. Smilde, J. Chemom. 16 (2002) 28.
[21] J.P. Gemperline, G. Puxty, M. Maeder, D. Walker, F. Tarczynski, M. Bosserman, Anal. Chem. 76 (2004) 2575.
[22] J. Diewok, A. De Juan, M. Maeder, R. Tauler, B. Lendl, Anal. Chem. 75 (2003) 641.
[23] J.P. Bovee, P. Strehaiano, G. Goma, Y. Sevely, Biotech. Bioeng. 26 (1984) 328.
[24] G. Birol, P. Doruker, B. Kirdar, Z.I. Onsan, K. Ulgen, Process Biochem. 33 (1998) 763.
[25] A. De Juan, R. Tauler, Anal. Chim. Acta 500 (2003) 195.
[26] P.J. Gemperline, Anal. Chem. 58 (1986) 2656.
[27] M. Maeder, Anal. Chem. 59 (1987) 527.
[28] J. Jiang, Y. Liang, U. Ozaki, Chemom. Intell. Lab. Syst. 71 (2004) 1.
[29] P.J. Gemperline, E. Cash, Anal. Chem. 75 (2003) 4236.
[30] M. Blanco, A.C. Peinado, J. Mas, Biotech. Bioeng. 88 (2004) 536.
[31] W.T. Vetterling, S.A. Teukolsky, W.H. Press, B.P. Flannery, Numerical Recipes. Example Book (C), 2nd ed., Cambridge University Press, 1988.
[32] R.L. Borrelli, C.S. Coleman, Differential Equations: A Modelling Perspective, John Wiley & Sons, New York, 1998.

[33] M. Blanco, A.C. Peinado, J. Mas, Anal. Chim. Acta 544 (2005) 199.
[34] P.J. Gemperline, Chem. Inf. Comput. Sci. 24 (1984) 206.
[35] F. Godia, C. Casas, C.J. Sola, Chem. Tech. Biotechnol. 41 (1988) 155.
[36] G. Teissier, Rev. Sci. 209 (1942) 3208.
[37] V.H. Edwards, Biotech. Bioeng. 27 (1970) 679.
[38] J.H.T. Luong, Biotech. Bioeng. 29 (1987) 242.
[39] C.N. Hinshelwood, The Chemical Kinetics of the Bacterial Cells, Oxford University Press, London, 1946.
[40] T.K. Ghose, R.D. Tyagi, Biotech. Bioeng. 21 (1979) 1401.
[41] S. Aiba, M. Shoda, M. Nagatani, Biotech. Bioeng. 10 (1968) 845.

IV

TEMPERATURE-INDUCED VARIATION FOR NIR TENSOR BASED CALIBRATION

Peinado A. C., Van der Berg F., Blanco M., Bro R.

Sent to.....

TEMPERATURE-INDUCED VARIATION FOR NIR TENSOR-BASED CALIBRATION

Peinado A.C.[a], Van der Berg F.[b], Blanco M.[a], Bro R.[b]

[a] Dept. of Chemistry, Faculty of Science, Universidad Autónoma de Barcelona, 08193 Bellaterra, Barcelona. Spain.
[b] Dept. of Food Science, Chemometrics Group. The Royal Veterinary and Agricultural University,1958 Frederiksberg C. Denmark

Keywords: PARAFAC, MLR, NIR, Temperature, Tensor, Modelling

ABSTRACT

The calibration strategy described in this work takes advantage of the synergic combination of temperature-induced spectral variation in Near Infrared (NIR) spectroscopy and the properties of tensor models. Rather than seeing spectroscopic temperature effects as artefacts that have to be circumvented or eliminated, a Parallel Factor (PARAFAC) model is used to extract and separate the relevant sources of information about the physical and chemical changes in a system. This information is highly related to the sources that provoke changes in the system as a function of temperature, but can not be ascribed directly to them, mainly due to the nonlinearities induced in the spectra. For quantification purposes Multiple Linear Regression (MLR) is used to build a least squares calibration model from the PARAFAC sample scores.

Temperature plays a key role in our strategy by providing an additional, meaningful dimension to a standard two dimensional spectroscopic data structure, thereby turning the quantification and qualification task into a tensor problem. That way the temperature effect on the spectral data can be modelled and predicted in a straightforward and highly effective way using this novel approach.

The combining strategy has been successfully applied to two NIR data sets. The first one is a laboratory off-line data set well-known, described and utilised by different research groups for testing new methods to cope with the temperature effect, treating it as an undesirable artefact. The other set is a batch process in-line data set with simultaneous changes in temperature and chemical composition. In this paper we will introduce a novel way of generating tensor data, show the advantages from an interpretational and predictive point of view, and present a comparison with traditional chemometric tools.

INTRODUCTION

Near Infrared (NIR) radiation and temperature are an inseparable pair. In fact, Fredrick William Herschel discovered NIR energy in 1800 just by using a thermometer. During the last 50 years where this part of the electromagnetic spectrum was recalled to the attention of the scientific community and has been studied more thoroughly. Amongst the characteristics of NIR [1], one of the most remarkable is its high sensitivity to temperature changes. This property makes NIR radiation a suitable, valuable and useful tool in e.g. the agricultural and environmental fields for measuring and monitoring the evolution of temperature in ecosystems [2]. Yet, when NIR spectrometers, along with multivariate calibration techniques, are applied to solve analytical chemistry problems in scenarios where the measurement conditions are not well controlled and/or known, the influence of external process variables such as temperature, pressure and moisture can dramatically affect the performance of calibration models in a negative way. The difficulty of keeping those parameters constant and/or the need to change them during the evolution of a process has been an important stimulus for the development of data treatment and pre-processing methods focused on minimizing their influence on the predictive performance in NIR calibration models [3-6].

In liquid-sample NIR spectra, temperature changes have a particular marked effect on the absorption bands for functional groups forming inter- or intra-molecular hydrogen bonds, provoking (nonlinear) shifts and a narrowing of the spectral bands [7]. For instance in aqueous systems, undoubtedly the simplest and most studied system in chemistry, the effects of temperature on NIR spectra are well-known, described and documented [8]. However, there is not a consensus model that describe the water peaks evolution with temperature. Even more, it is still an open question whether water can be considered as a continuous system [8-9] in which hydrogen bonds weaken with increasing temperature or, on the contrary, its behaviour can be described as a discrete system of two [10-11] or more components [12-13]. The situation gets even

worse when more complicated systems are handled and that is one reason why a large number of alternative correction and calibration methods have been proposed for modelling the temperature and to get rid of its distorting effects in the prediction of unknown samples. The most widespread solution involves implicitly inclusion of temperature in inverse multivariate calibration models such as Principal Component Regression (PCR) and Partial Least Squares (PLS), in order to span all the variability provoked and treat it as if it were additional interferents [3]. The disadvantage of this approach is that the final models are more complex, in terms of the number of components/factors, and temperature effects are not isolated or removed. Another approach is robust variable selection methods [4], but the results obtained are not always better than the implicit inclusion of temperature although the models are typically less complex in terms of number of factors.

Different linear [5] and nonlinear [6], [14] methods for explicit inclusion have also been tested but in the former no improvements were reported in relationship to implicit inclusion of the temperature information, and in the latter, although improvements were achieved in terms of prediction ability, time consuming calculations, difficulties in parameter interpretation and in the implementation of the algorithms make these methods not to be widespread techniques. In general terms, temperature has always been considered as an external nuisance factor that necessarily must be corrected for or modelled in order to get reliable and robust calibration models. In this work we will confront the problem from another perspective. We will consider temperature not as a nuisance factor but as a constructive parameter that can provide detailed chemical information when systematically changed during a measurement. As a modelling method we will use three-way tensor algorithms that have already shown a good performance when applied to a variety of different problems [15]. To our knowledge, it has never been reported what contributions and/or performance improvements can be achieved from such a temperature-induced tensor expansion applied to multivariate measurements.

THEORY

PARAFAC is a decomposition method that can be considered as a generalization of the bilinear Principal Component Analysis (PCA) for two and higher order-tensors or multi-way arrays [16]. The principle behind the PARAFAC decomposition is to minimize the sum of squares of the residuals e_{ijk} as indicated in equation (1) for a three-way *F*-component PARAFAC model:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad (1)$$

The element x_{ijk} represents the data for sample *i* at variables *j* and *k*. The rank of the PARAFAC model is given by the number of factor or components, F, needed to describe the variation in the data array. Each component, f, consists of one score or loading vector in the first mode, a_{if}, and two loading vectors, b_{jf}, c_{kf} for, respectively, loading vectors in the second and third mode. The important difference between PCA and PARAFAC is that in the latter there is no need for requiring orthogonality to identify the model and loading vectors are directly related to the main sources of variation in each mode. In this way, when PARAFAC is applied to approximately trilinear data sets, the solution enjoys the property of chemically meaningful uniqueness, and e.g. score values can be used to create univariate models and chemical-physical interpretation can be draw from the trilinear decomposition [15], [17]. In the case of systems affected by sources of nonlinearity, the PARAFAC models fails in providing factors directly related with the main sources of variation in the data, although they hold allied information about the changes in the system. This general, *PCA-like* rank approximation property of tensor models is what we will employ in this paper. The information can be extracted by using indirect calibration method such as MLR with the retrieved sample scores as independent variables.

In Figure 1 a schematized version of the calibration strategy proposed in this paper is presented in which PARAFAC and MLR are combined to create a joined model. First, PARAFAC is applied to the temperature-induced, tensorized samples in

the calibration set. An MLR modelling is performed with the scores (**A**) and the external reference values (**y**) to obtain a least squares regression vector $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \qquad (2)$$

Using the PARAFAC loadings in the second (**B**) and third mode (**C**) plus an unknown data sample (**X**; e.g. a validation or test-set sample), the estimated scores ($\hat{\mathbf{a}}$) for each sample is obtained by solving a least square problem [18]:

$$\arg\min_{\mathbf{a}} \left\| \mathbf{X} - \mathbf{B}\,\mathrm{diag}(\hat{\mathbf{a}})\mathbf{C}^T \right\|_F^2 \qquad (3)$$

Finally, estimated values ($\hat{\mathbf{y}}$) for the validation samples are obtained from the regression vector and the matrix of predicted scores ($\hat{\mathbf{A}}$):

$$\hat{\mathbf{y}} = \hat{\mathbf{A}}\,\hat{\boldsymbol{\beta}} \qquad (4)$$

This strategy can be used for modelling data and predict the principle species in systems with shifting profiles induced by nonlinear agents such as temperature. When PCA is applied to such systems, the first components determined by the main source of variation which usually represents both chemistry and shifts, can be modelled by including additional components, whose loadings will look like derivatives of the original shape [19]. This observation can be generalized to PARAFAC but with the difference that orthogonality is not required to identify the model and components turn out to be highly related with the main phenomena and sources of variation - temperature and chemical composition.

In this work we will show the possibilities of this combined strategy when applied to systems influenced by nonlinear effects due to the temperature-induced variation in NIR spectra. In a future publication we will describe and elaborate on the theoretical and practical implications of this strategy.

EXPERIMENTAL SECTION

Samples

Two data sets have been used in this work. The first one contains the pure spectra of water, ethanol and iso-propanol plus binary and ternary combination of the pure analytes according to a mixture design. Spectra were recorded at the temperatures 30, 40, 50, 60, 70ºC in the short-wave NIR spectral region 580 to 1090 nm. This data set is well described in literature [3] and has been used and studied in different papers [4-6], [14]. The other set is an in-line data set composed of six batches, three corresponding to the pure species water, glycerine and ethanol, while the other three correspond to the binary mixtures of glycerine diluted with ethanol, glycerine diluted with water, and ethanol diluted with water over time. A common changing temperature profile was used for all batches as shown in Figure 2. The profile always started and finished at 25ºC and is formed by three concatenated cycles of heating up and cooling down with respective gradients of 1.0, 1.5 and 2.0ºC per minute. The relative maximum and minimum temperatures for each cycle were 50, 65 and 75ºC and 20, 15 and 10ºC respectively; when a relative extreme temperature was reached it was maintained for a period of five minutes. For binary samples, a common dilution pattern was also set (Figure 2). The pattern was formed by alternating dilution and non-dilution intervals spanning the three temperature cycles

Apparatus and Software

The in-line/batch spectra were recorded in a LabMax laboratory reactor from Mettler Toledo equipped with automatic temperature and dosage control. An immersion Pt-100 sensor was used for the in-line register of temperature (±0.2ºC). Transflectance spectra were recorded using a Foss NIRSystems 5000 spectrophotometer equipped with a stainless steel optical probe. The light path was adjusted in order to obtain appropriate absorbance measurements. The spectrophotometer was accessed via the Vision v. 2.51 software package. Spectra were recorded at one minute intervals as the

average of 32 scans performed at 2 nm intervals in the NIR wavelength range 1100–2500 nm. Spectra and temperature registration were synchronized.

Data Processing

For both data sets, temperature was used as the key factor to build the three-way data structures from the set of spectra (two-way data). In this way, each dimension of the cube structure is related to and contains organised information about either samples/chemical changes (expressed as the analyte concentration) or spectroscopic variation (measured as absorbance for each wavelength) or energetic state (measured as temperature). The in-line three-way data structures were built by piling up three batches, two of them corresponding to pure species and the third being the evolving batch resulting from diluting one of the analytes with the other analyte. The number of samples was 258 for the ethanol-water and glycerine-water mixtures and 233 for the glycerine-ethanol mixture. Figure 3 gives a graphical representation of how the at-line/laboratory and in-line/batch two-way data set were arranged into three-way data structures.

The at-line laboratory data set was de-trended by fitting a straight line for each spectrum at the wavelength range 749 to 849 nm, where no absorbance bands are present, and subtracting it from the entire spectrum. Final data analysis was performed for the region 850-1049 nm [3].

The in-line spectra were used in the region between 1100-2350nm. Fibre-optic probes exhibit slight noise above 2200 nm that increases up to 2500 nm. However this region has useful analytical information related to the spectral changes in the system, so it called for a compromise: the 2200–2350 zone was retained on the grounds of its high signal-to-noise ratio and the 2350–2500 nm zone excluded as it exhibited a considerably lower quality.

Matlab (ver. 7.0, The Mathworks, Inc.) and the PLS toolbox (ver. 3.0, Eigenvector Research, Inc.) were used for data pre-processing and modelling.

RESULTS AND DISCUSSION

Figure 4 shows, by way of example, the spectral variation as a function of the temperature profile, applied to one of the in-line data sets of pure water (a) and glycerine (b) and the batch glycerine diluted with water (c). The most significant changes due to temperature can be seen around 1440 nm, the region in which the combination of anti-symmetric and symmetric O-H stretching modes are found (first overtone). In this band, an increase in temperature causes a hypsochromic shift in the absorption maxima. In addition to the effect of temperature, Figure 4c also shows the effect of dilution on the glycerine spectrum. The spectrum steadily evolves into the water spectrum, thus, the regions around 2100 nm and 1500 nm, corresponding respectively to the combination band and first overtone of the R-OH bonds, loose intensity and are embedded into the emerging bands around 1440 nm and 1930 nm, corresponding respectively to the first overtone of the O-H bond and to the combination band of the stretching and bending water O-H bonds. The changes induced by temperature and the evolving chemical composition in the system plus the combination of both effects can be considered sources of nonlinear variation in the spectra.

Determining the true underlying sources of variation is one of the main reasons why PARAFAC has traditionally been applied to multi-linear systems. However, the aim of applying PARAFAC to three-way systems, formed by using temperature as a building parameter will be to extract factors that will be highly related to the (two) main causes of variation. They will however certainly not be the pure effects by themselves due to the nonlinear effects. Hence, the result of the PARAFAC model should be seen as a reasonable low rank approximation much in line with the way principal component analysis is often used. The PARAFAC factors (sample scores) will be used as independent variables to model temperature and chemical composition by using Multiple Linear Regression (MLR) as was depicted in Figure 1.

In-line/batch data set

Before applying PARAFAC to the three-way data, these were centred across the concentration mode in order to remove small inter-batch differences due to uncontrolled sources of variation. By way of example, Figure 5 shows the score plot (a), loading plot in the second mode (b) and loading plot in the third mode (c) obtained by fitting PARAFAC with three components to the water-ethanol batch. The sample score values for the first component are highly related to the temperature profile, with a correlation coefficient equal to 0.9966. In the second mode, the spectral loading vector associated to the first factor shows a characteristic shape. It can be seen that the highest variance in intensity is found around the two stronger bands of the O-H water bond in the region of 1450 nm and 1940 nm where maxima and minima are closely linked. This loading vector shows a very obvious relationship with the loading vector found by Libnau et.al. [10] when studying the effect of temperature on NIR spectra of pure water. Finally, the sample-stacking loading in the third mode of the first component shows a high value in the batches where there exists water; and its value is at a minimum for the pure-ethanol batch.

Second and third factors in the first mode are related to the dilution pattern as can be seen when scores are compared with the applied dilution pattern in Figure 2. In the spectral mode it is shown that the loading vectors of the second and third component capture the main absorption bands characteristics, not only of water but also of ethanol. These structures can be assigned to the first overtone of the CH_2 and CH_3 around 1700 nm, the combination band of R-OH around 2000 nm and different combination bands of the CH_2 group around 2280 nm. Comparisons can be made with the pure spectra collected in Figure 5d. In the third tensor mode, components two and three follow a similar behaviour, and give a maximum value in the mixture batch.

Similar appraisals about the behaviour of the loading vectors in the different PARAFAC modes were also observed when the water-glycerine and ethanol-glycerine data structures/batches were analyzed (results not shown). From these assessments it

can be inferred that score vectors that mainly ascribe temperature have, in the spectral mode, associated loading vectors that ascribe variability due to the molecular structure mostly affected by temperature, i.e. the OH bond. In mixtures were the water is present such as ethanol-water and glycerine-water, the loading assigned to temperature follows a similar shape in the region where the OH bonds are manifest, having a correlation coefficient of 0.9975 between them. In the same sense, when pure water is not present as in the glycerine-ethanol batch the loading assigned to temperature is the R-OH structure as the next most temperature-sensitive one. In Figure 6 the loading vectors associated to temperature for the three data sets studied are shown. It can be seen that the differences between the two principal minima, with water versus without water, are approximately 50 nm in the region around 1400 nm and 2000 nm in the region between 1850 nm - 2050 nm. Those values are related to the different absorption bands between the first overtone and the combination band of the OH and R-OH structures respectively [20]. In the same way, for the different data structures analyzed the score vectors related to the chemical changes, i.e. the dilution pattern, have associated loading vectors in the spectral mode that describe the absorption bands of the main structure involved, such as is presented in Figure 5.

For each data set MLR models were built by using as independent variables the three loading vectors in the sample mode (i.e. score vectors) obtained from the PARAFAC models. Every fifth sample was used to compose the calibration set, the rest were used for external validation. In Table 1 the figures of merit for temperature and concentration predictions of the external test data are shown. As reference concentration values the known weight percentage for each sample are used. Since the system is closed, only the figures of merit for the concentration of the analyte initially present in the reactor are presented. In the two batches where glycerine was used, samples belonging to the first dilution step were not included, neither in calibration nor validation, due to the anomalous spectral behaviour in the initial stage of the experiment due to a slow homogenization during the initial stage of the experiment.

It is emphasized that only three PARAFAC factors have been used in all the MLR calibration models as independent variables. These are well interpretable main sources of variation due to chemical structures that characterize the different analytes and the changes in temperature. On the other hand, we do not consider the constructed data tensor to follow an exact trilinear structure. Three PARAFAC components/factors is a good approximation of the data, sufficient for quantification as shown in Table 1. One might argue that two factors should suffice (one for temperature and one for the closure system with two analytes), but we view the third factor is required to achieve proper modelling of the nonlinear behaviour due to temperature. Experiments have shown that the combination of four PARAFAC factors and PLS regression give similar predictive performance.

At-line/Laboratory data set

This data set has become a de facto standard and used by different groups to test the performance of their algorithms to cope with the influence of temperature effect on vibrational spectra, see for example [3-6], [14]. The calibration and validation sets employed were the same as those cited in literature with the intention of being able to establish and evaluate the usefulness of the strategy proposed in this paper. In the calibration set thirteen samples, located in the periphery of a ternary mixture design plus the central point were used. The external validation set was made up from six samples positioned inside the ternary mixture design for all cases. Following the same nomenclature used in the precedent papers, one global and two local models for ethanol, water and iso-propanol were built.

The global model for each analyte was built with all temperatures and was used to select the number of components and to test different spectral pre-treatments such as first and second derivatives, SNV (Standard Normal Variate) scaling and also combination of SNV and derivatives. The best results in terms of RMSEP were achieved by using four components and first derivative (with a moving window of seven points and a second degree polynomial order). RMSEP values obtained were 0.32 for

water, 0.67 for ethanol and 0.71 for iso-propanol. Based on these findings, local models were built. Two local model where tested: for temperature interpolation 30 plus 70ºC samples were used for calibration while 50 plus 60ºC samples were used for validation; for temperature extrapolation, calibration was performed with 50 plus 60ºC samples while validation was done with 30 plus 70ºC samples. In Table 2 results are shown for the PARAFAC-MLR and the PLS models (number in brackets is model complexity), the values of RMSEP for external validation with four different validation setups: two equal-temperature models, one interpolation model and one extrapolation model. As can be seen, in all cases the best RMSEP results were achieved for water prediction. This is not surprising for two reasons: (a) the higher absorptivity (implying higher sensitivity) of the water OH bond in relationship to the other bond structures, (b) PARAFAC scores used in the calibration are directly related to the loading set. These loadings mainly compile the information of the structures that change the most with temperature, i.e. the water OH bond.

Considering the PARAFAC-MLR results, when validation temperatures were the same as those used to build the local models, the performance obtained was similar and of the same order of magnitude as achieved with the global model and differences between results were not significant. In the same way, when local models were validated with samples recorded at different temperatures, the results obtained for water were not significantly different from the above results. However, the results for ethanol and iso-propanol were slightly worse than those obtained when validation was performed at the same temperature. This observation is foreseeable due to the extrapolation performed but, notwithstanding the higher RMSEP values, the performance is similar to that obtained with a recently proposed algorithm [21] in which the influence of temperature on the predictive ability of the model was concluded to be eliminated.

Comparing the RMSEP values for the PARAFAC-MLR and PLS models, it can be seen that, in general terms, values are one order of magnitude higher in the case of

PLS models. Besides, PLS models are also complex in terms of numbers of components; never less than five. For the PARAFAC-MLR models the number of factors was four in all the cases.

CONCLUSIONS

In this work temperature is introduced as a key agent to generate three-way structure from multivariate spectroscopic data. Via this second-order tensors can be made just by measuring samples at different (controlled) temperatures. This simple procedure widens the field of application of multiway calibration methods, well-known for its favourable mathematical properties, but limited in its present-day use by the lack of suitable three-dimensional structures, especially in the analytical chemistry field.

A new calibration strategy based on the combined use of PARAFAC and MLR was tested and applied to two different three-way data sets. The external validation results obtained were very good for all the chemical species, even under temperature interpolation and extrapolation. This means that, although temperature is a known source of non-linearities on vibrational spectra, when it was utilised as a building agent, predictions were liberated of its distorted effect. PARAFAC is used to extract the relevant information about the physical-chemical changes happening in evolving systems with a minimum number of factors. MLR is used to build a least squares calibration method from the factors extracted. Based on the different figures of merit obtained in both data sets, it seems that the suggested approach is an interesting calibration tool for reliable, effective, and accurate results.

The strategy proposed can be interesting and easy to implement in different industrial applications to exploit the effect of temperature in calibration models. In view of future spectroscopic instrument, it would be beneficial to develop rapid modules that were able to efficiently register spectra of single samples at different temperatures in short time intervals.

BIBLIOGRAPHY

[1] J.Workman, Applied Spectroscopy: a compact reference for practitioners. Ed. Academic Press, 1998.

[2] B. Duchemin, Remote Sens. Environ. 67 (1999) 51-67.

[3] F. Wulfert, W. T. Kok, A.K. Smilde, Anal. Chem. 70 (1998) 1761-1767.

[4] H. Swierenga, F. Wulfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, ACA 411 (2000) 121-135.

[5] F. Wulfert, W. T. Kok, O. E. de Noord, A. K. Smilde, Chemom. Intell. Lab. Syst. 51 (2000) 189-200.

[6] P.H.C. Eilers, B. D. Marx, Chemom. Intell. Lab. Syst. 66 (2003) 159-174.

[7] B. Osborne, T. Fearn, Ed. John Wiley & Sons,1993.

[8] M. Starzaka, M. Mathlouthi, Food Chemistry 82 (2003) 3–22.

[9] D. Eisemberg, W. Kauzmann, The structure and properties of water. Ed. Oxford University Press, 1969.

[10] F. O. Libnau, O. M. Kvalheim, A. A. Christy, J. Toft, Vib. Spectrosc. 7 (1994) 243-254.

[11] S. Navea, A. de Juan, R. Tauler, Anal. Chem. 75 (2003) 5592-5601

[12] V.H. Segtnan, S.Sasic, T. Isaksson, Y. Ozaki, Anal. Chem. 73 (2001) 3153-3161

[13] H. Büning-Pfaue, Food Chemistry, 82 (2003) 107-115

[14] U. Thissen, B. Üstün, W. J. Melssen, L.M.C.Buydens, Anal. Chem. 76 (2004) 3099-3105

[15] A. K. Smilde, R. Bro, P. Geladi, Multi-way Analysis. Application in the Chemical Sciences. Ed. John Wiley & Sons, 2004

[16] R.A. Harshman, UCLA working papers in phonetics,16 (1970) 1-84.

[17] A. S. Field, D. Graupe, Brain Topography, 3 (1991) 407.

[18] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149-171.

[19] J. Möcks, Psychophysiology 23 (1986) 480-484.

[20] D. Burns, E. Ciurczak, "Handbook of Near-Infrared Analysis 2nd Edition" Ed. Marcel-Dekker, Inc. New York, 2001.

[21] Z. Chen, J. Morris, E. Martin, Anal. Chem. 77 (2005) 1376-1384

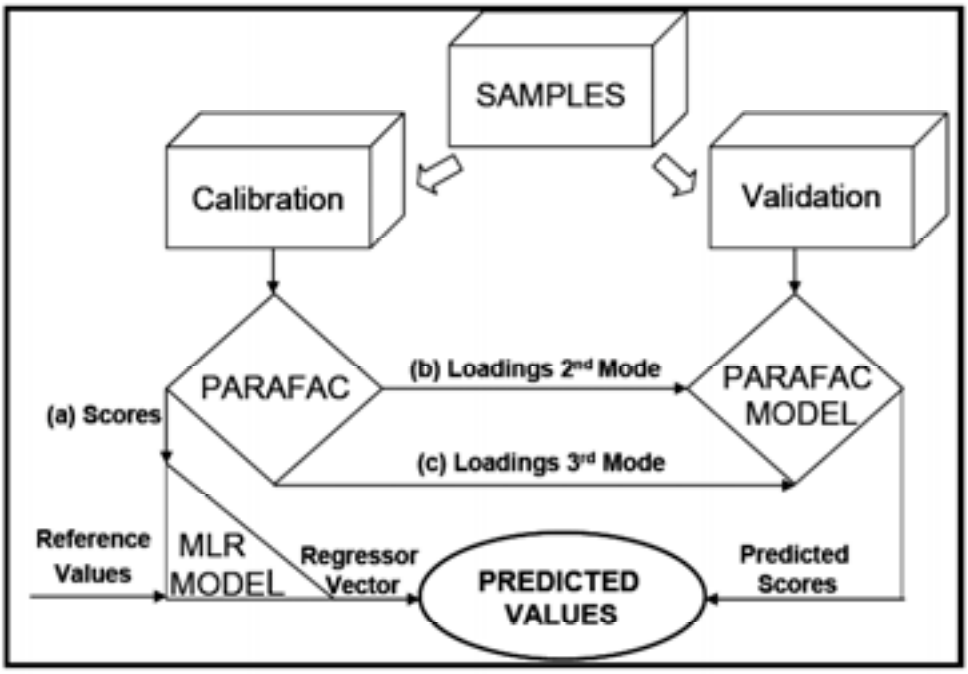


Figure 1. Scheme of the calibration strategy followed by combining PARAFAC and MLR models

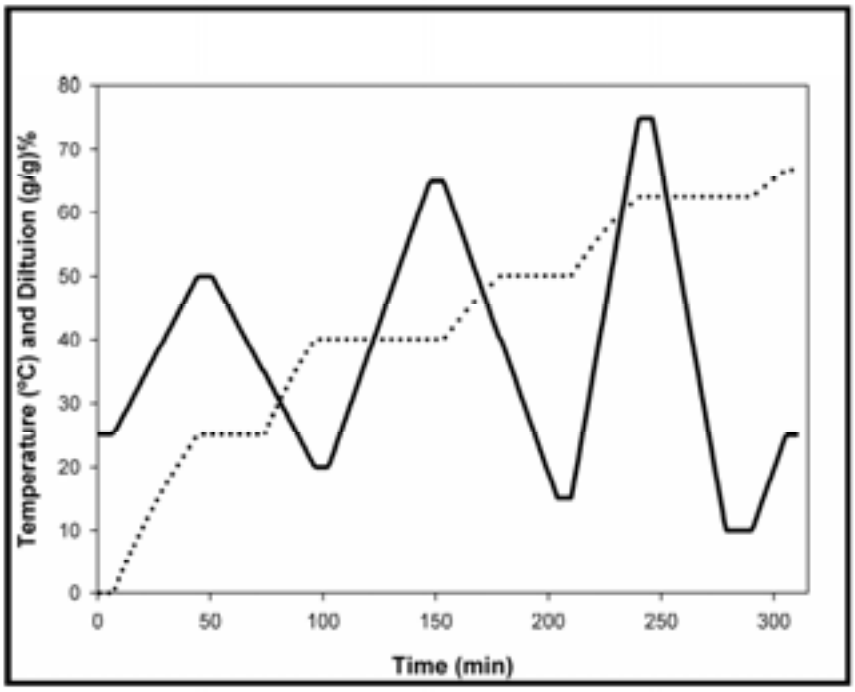


Figure 2. Temperature (solid line) and dilution profiles (dotted line) as applied to the batches

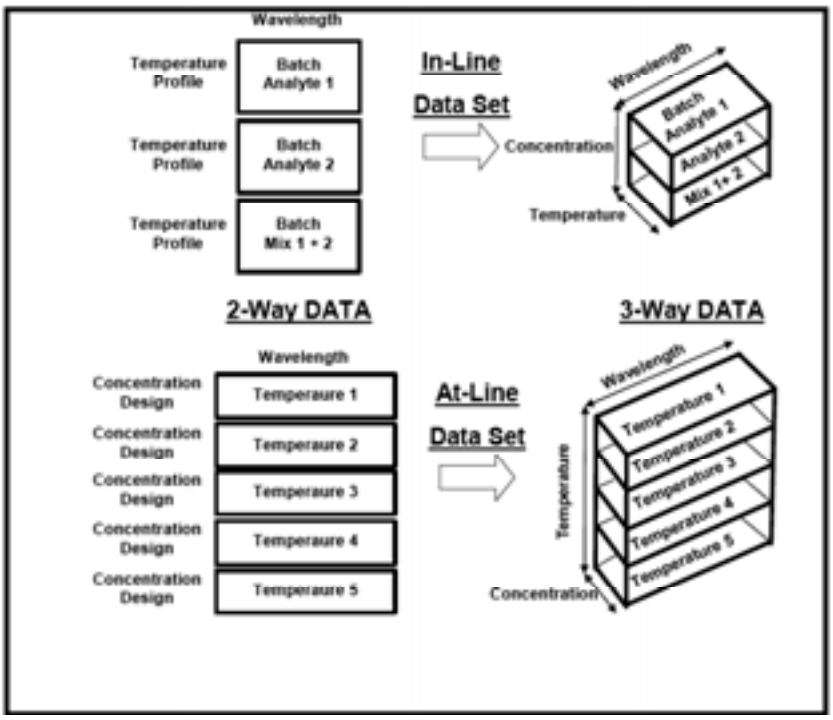


Figure 3. Three-way data structure construction for the in-line/batch and at-line/laboratory data sets

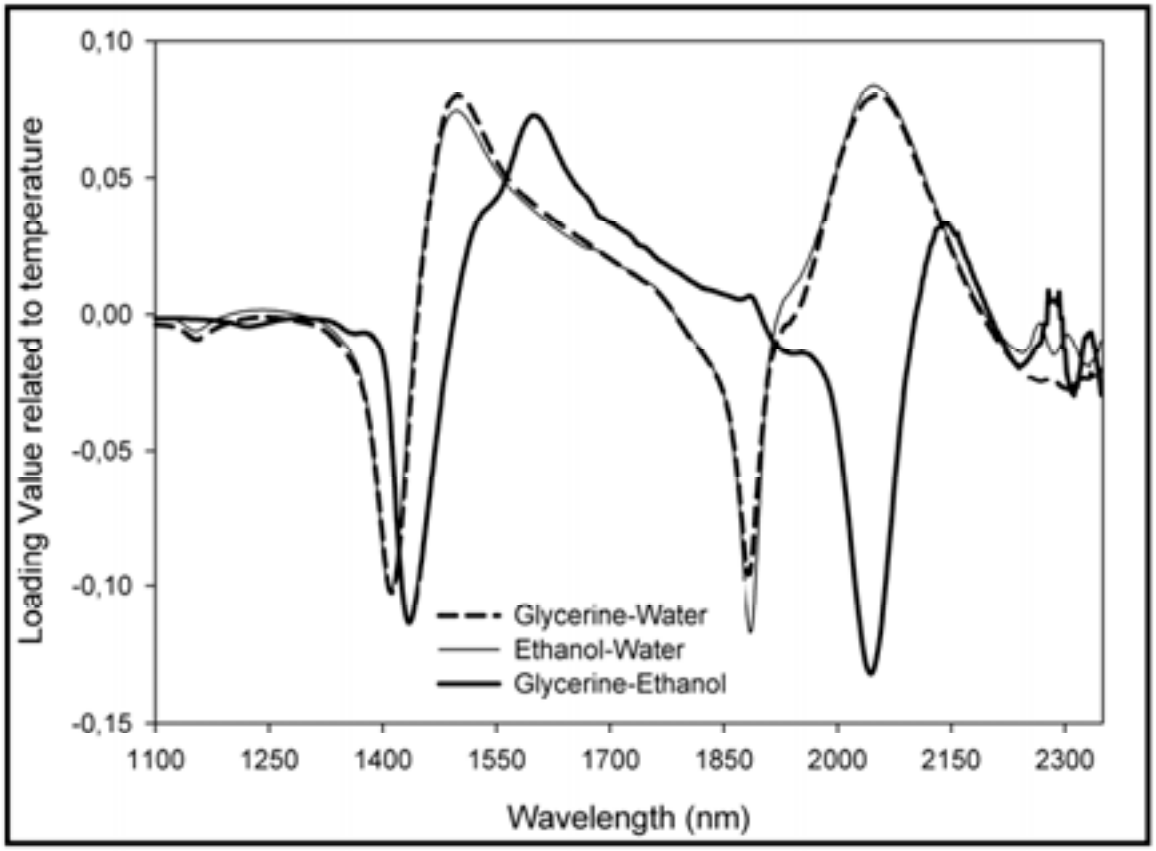


Figure 6. PARAFAC spectral loadings assigned to temperature for glycerine-water, ethanol-water and glycerine-ethanol systems.

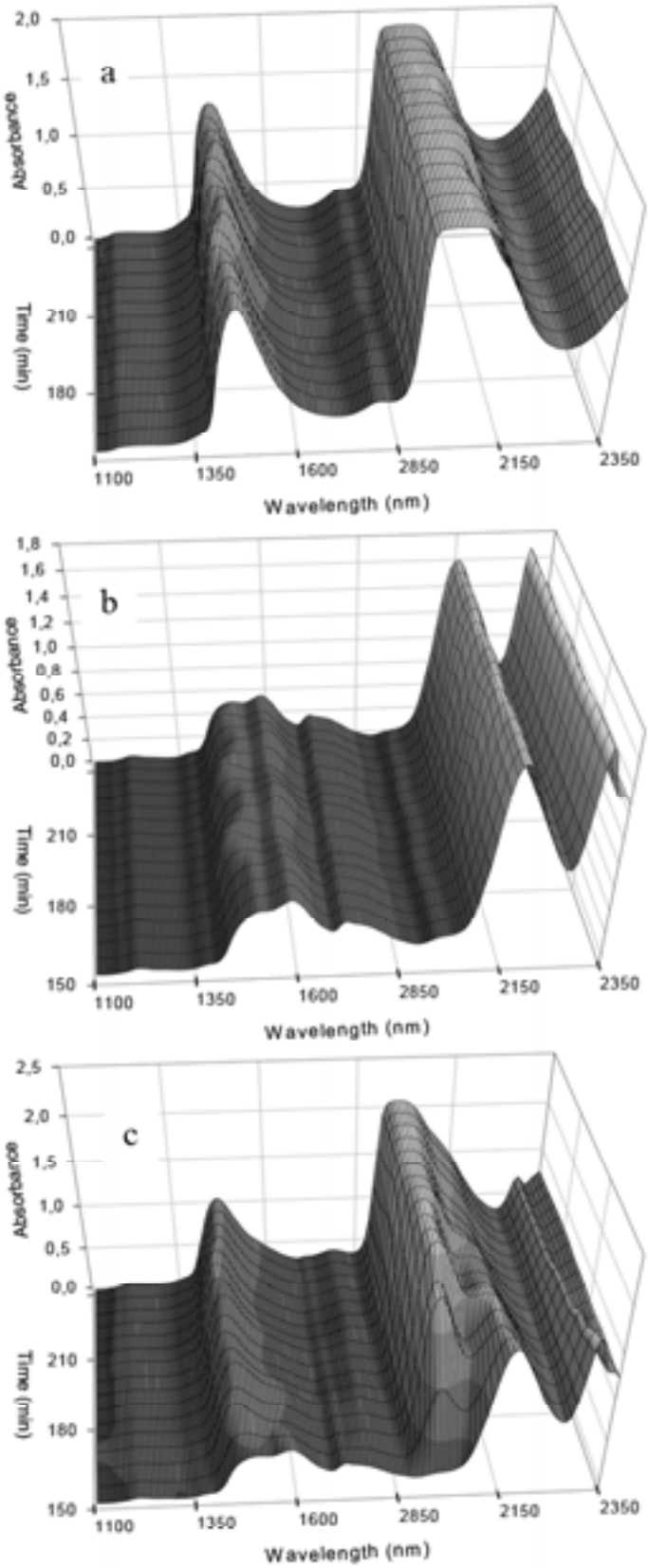


Figure 4. Spectral time-evolution of (a) water, (b) glycerine and (c) mixture profile according to the patterns presented in Figure 1.

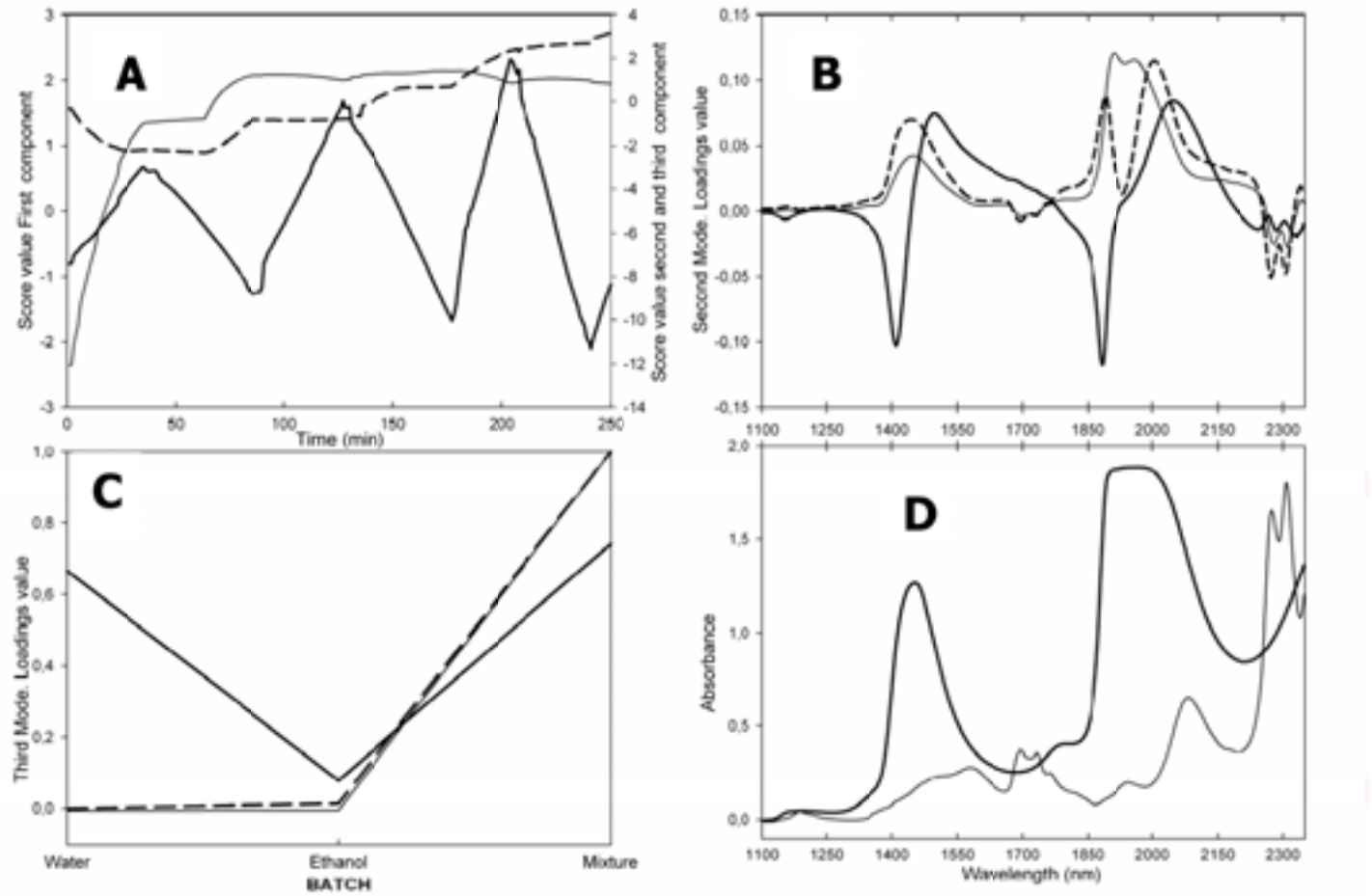


Figure 5. PARAFAC results for water-ethanol system. (Full line = factor 1, broken line = factor 2, thin line = factor 3): (a) Sample score plot; (b). Spectral loadings (2nd mode); Sample-stacking loadings (3rd mode). (d) Spectra of water (full line) and ethanol (thin line);

Table 1. External validation figures of merit for temperature and analyte concentration for the different data structures studied in-line. Nc/Np is the ratio between the number of samples used in the calibration and in the validation set. RMSEP is the Root Mean Square Error of Prediction for the validation set. R^2 is the coefficient of determination between the reference and predicted value.

	Ethanol-Water		Glycerine-Water		Glycerine-Ethanol	
	Temp	Ethanol conc	Temp	Glycerine conc	Temp	Glycerine conc
Nc/Np	51/207	51/207	51/207	43/165	47/186	40/160
RMSEP	0.0079	0.0039	0.0095	0.0064	0.0066	0.0070
R^2	0.9984	0.9997	0.9982	0.9993	0.9991	0.9984
Slope	0.9962 ± 0.0077	1.0004 ± 0.0032	1.0083 ± 0.0085	1.0026 ± 0.0057	0.9994 ± 0.0062	0.9935 ± 0.0088
Offset	0.1468 ± 0.3221	0.0032 ± 0.1634	-0.2895 ± 0.3566	0.2846 ± 0.3188	0.0475 ± 0.2589	0.3476 ± 0.5138

Tabla 2. RMSEP values for external validation in PARAFAC-MLR and PLS models. For PARFAFAC_MLR models the number of factors was always four. The number between brackets indicates the number of PLS factors used.

T VAL (ºC)	Chemical Specie	T Calibration (ºC)			
		30 & 70		50 & 60	
		PARAFAC_MLR	PLS	PARAFAC_MLR	PLS
30 & 70	Water	0.0044	0.0047 (8)	0.0024	0.0176 (6)
	Ethanol	0.0082	0.0149 (9)	0.0147	0.0311 (11)
	Isopropanol	0.0086	0.0129 (9)	0.0143	0.0194 (9)
50 & 60	Water	0.0035	0.0072 (5)	0.0045	0.0059 (6)
	Ethanol	0.0106	0.0257 (6)	0.0099	0.0142 (9)
	Isopropanol	0.0091	0.0177 (5)	0.0096	0.0162 (9)