On-line concentration measurements in wastewater using nonlinear deconvolution and partial least squares of spectrophotometric data

A. Vargas* and G. Buitrón*

* Environmental Bioprocesses Dept., Institute of Engineering, UNAM, Circuito Exterior s/n, Ciudad Universitaria, Coyoacán D.F., 04510 Mexico (E-mail: avargasc@iingen.unam.mx, *gbm@pumas.iingen.unam.mx*)

Abstract
A procedure for the on-line measurement of concentrations of toxic in wastewater using spectrophotometric data is proposed. The complete absorbance spectrum of a wastewater sample is used to predict the concentrations of all possible substances within. Two techniques are examined. In nonlinear spectral deconvolution the spectrum is decomposed as a linear combination of base spectra and the coefficients of this deconvolution are used to nonlinearly estimate the concentrations. Under partial least squares analysis the concentrations are directly estimated as a linear combination of the measured spectrum data. Both techniques show good results for estimating the kinetics of samples taken during the reaction phase in a laboratory anaerobic/aerobic SBR used for *p*-nitrophenol biodegradation.

Keywords
Concentration measurement, deconvolution, nitrophenol, PLS, SBR, spectrophotometry, wastewater treatment

INTRODUCTION

Wastewater treatment processes require monitoring of certain parameters to ensure correct operation. In discontinuous processes such as sequencing batch reactors (SBRs), the operating conditions are dynamic, so determining the usual water quality parameters (total organic and inorganic carbon, chemical and biochemical oxygen demand, etc.) may not be enough to supervise the process. Furthermore, the techniques to determine these parameters may be time consuming and expensive (APHA, 1992). It may therefore be of interest to directly measure some other parameters using an inexpensive and fast technique. In particular, the aim of this work was to propose a methodology for on-line measurement of the concentration of several compounds present in industrial type wastewaters using UV-Visible spectrophotometric data.

Biological wastewater treatment processes have demonstrated their ability to treat toxic wastewaters (Grady Jr. *et al.*, 1999) and sequencing batch reactors (SBRs) are a viable alternative (Wilderer *et al.*, 2001). A SBR is typically a tank that operates under time-controlled predefined phases: fill, react, settle, draw, and idle. At the beginning of the reaction phase the concentration of the toxic to be treated is high; it then decreases during this phase, until at the end it is ideally close to zero. Some intermediates may also be formed, increasing and decreasing their concentration during the reaction. To study the process, or merely to monitor that complete biodegradation takes place, as well as to establish the operating regime, it is convenient to determine the kinetics of these concentrations with respect to time. However, sensors for measuring these concentrations are either too expensive or unavailable, and one is forced to take samples and process them off-line using time consuming techniques, such as HPLC, which requires lengthy pre-treatment of the samples and is furthermore destructive of these samples.

Spectrophotometry is widely used in chemometrics to determine the concentration of certain species in samples. The technique is based on the well known Beer-Lambert law that states that the absorbance at a certain wavelength is directly proportional to the concentration if the path length

traversed by the light remains constant. Commercial apparatus and standard methods using this technique are based on measuring absorbance at a certain wavelength and using this law to estimate the concentration, assuming a linear calibration curve was determined beforehand (Jeffrey *et al.*, 1989).

Spectrophotometers operating in the ultraviolet and visible (UV/Vis) range are capable of reliably scanning the absorbance spectrum of a sample for a whole range of wavelengths. This multiwavelength data can be used to further enhance the concentration measurements. In this sense, a widely used technique is *partial least squares* (PLS) regression, in which the concentration(s) in a sample are estimated from a vector of absorbance data at different wavelengths by extracting the principal components and linearly combining the measured absorbances (Dahlén *et al.*, 2000). This methodology has also recently been used for monitoring wastewaters (Langergraber *et al.*, 2004).

Another recently used technique is *spectral deconvolution* (SD) (Thomas and Constant, 2004), which is based on expressing the measured absorbance spectrum as a linear combination of a few base spectra. The concentrations (or other water quality parameters) are then assumed to be another known linear combination of the coefficients of the deconvolution (Thomas *et al.*, 1996). Both spectral deconvolution and PLS lead to similar mathematical expressions for estimating the concentrations. The difference lies in the way the methods are calibrated.

This paper presents a modification of the SD technique, incorporating quadratic nonlinearities in the function relating the deconvolution coefficients and the concentrations, thereby increasing its flexibility. It also proposes a new calibration method that does not require external measuring equipment, only the preparation of artificial samples. Furthermore, a comparison of the method proposed with usual deconvolution and PLS for a specific application is also made, namely the biodegradation of *p*-nitrophenol in an anaerobic/aerobic SBR.

MATERIALS AND METHODS

A 7 litre jacketed laboratory biorreactor with an exchange volume of 4 litres was operated as an anaerobic/aerobic SBR. It was used to treat synthetic wastewater composed of *p*-nitrophenol (PNP) as model substrate and propionic acid in a molar relationship of 1:20. Nutrients were added according to AFNOR (1985). The anaerobic phase was used to reduce PNP to *p*-aminophenol (PAP), while the aerobic phase was used to finally mineralize PAP. The aerobic phase takes place in the same tank immediately after the anaerobic phase. The reactor was inoculated with activated sludge from a municipal wastewater treatment plant and acclimated according to the procedure by Melgoza and Buitrón (2001). Two peristaltic pumps (MasterFlex, Cole-Parmer) were used to feed and draw, while a mixer (65 to 85 rpm) was used to homogenize the reactor during the reaction phase. A solenoid valve was used to control passage of air through a diffuser at the bottom of the reactor at approximately 2.5 L/min during aeration. Temperature was controlled around 27ºC using a heater and a pump that recycles water through the reactor jacket. The operating strategy was controlled through a programmable timer (Chrontrol XT), allowing variable times for anaerobic and aerobic reaction, depending on the observed kinetics, 45 min for sedimentation, and 5 min of idle time. The concentrations of PNP at the beginning of the reaction phase varied from 15 to 40 mg/L. This was done to test the estimation procedure for different initial conditions.

To obtain the spectrophotometric data, a UV/Visible spectrophotometer with appropriate software was used (Lambda 25 and Win-Lab 2.85, Perkin-Elmer). Samples of 10 mL were taken at different times during both the anaerobic and aerobic phases of the reaction, centrifuged at 3000 rpm (Sol-Bat J600) for 3 min and filtered through a 1.6 μm microfiber filter (Millex). The filtered samples were

then measured by manually filling the corresponding quartz cell in the spectrophotometer; distilled water was used as the reference sample. After measurement, the samples and the centrifuged sludge were returned to the reactor. For each kinetics the spectrophotometer was calibrated and baseline corrected only once.

The data for calibration was prepared as follows: since the concentration of PNP in the synthetic wastewater feed was known, this was diluted with mixed liquor taken at the idle phase (PNP and PAP concentrations are assumed minimal). Additionally a base solution with known concentration of PAP was prepared and combined with this dilution to prepare "fake" samples with known concentrations of PNP and PAP. The measurement procedure was followed and the reference spectra were obtained, *i.e.* centrifuging and filtering as explained before. This was done individually and as fast as possible to avoid a reaction in these so constructed mini-reactors. PNP and PAP concentrations were also estimated by HPLC to corroborate the calculated concentrations of the calibration samples (Melgoza and Buitrón, 2001).

The data obtained was exported to MATLAB (The Mathworks) to perform the analysis, with a toolbox (ConSpect) developed by our group for this purpose.

ON-LINE CONCENTRATION ESTIMATION

The data used in spectrophotometry is conceptually a curve of wavelength *vs*. absorbance of a sample. A monochromatic ray is passed through a solution sample and the light intensity at the other end is compared to that of a blank reference sample. The wavelength of light is swept from a maximum λ_{max} to a minimum λ_{min} and the absorbance curve $s(\lambda)$ is recorded. For UV/Visible spectrophotometry this range goes usually from 190 to 900 nm. In practice, however, absorbance is measured only at discrete wavelengths and the data are given as two N-dimensional vectors $\boldsymbol{\lambda}$ and $\boldsymbol{s}$, i.e. as a set of N ordered pairs (λ_i,s_i) for $i=1,\ldots,N$, where $\boldsymbol{s} = [s_1,\ldots,s_N]^T$.

Nonlinear spectral deconvolution

Given a solution with a certain compound and assuming that the reference blank does not contain this compound, the Beer-Lambert law states that the absorbance measured at *any* wavelength is directly proportional to its concentration if the cell path length ℓ (the distance light travels through the sample) remains constant. That is: $s=\varepsilon\ell x$, where x is concentration and ε is a constant called *molar absorptivity*. If both concentration x and the cell path length ℓ do not change during the measurement of the absorbance spectrum $s(\lambda)$, this means that the only variable depending on the wavelength is $\varepsilon(\lambda)$, and thus one may define $\alpha(\lambda)=\ell\varepsilon(\lambda)$ as a *base spectrum* upon which

$$s(\lambda) = x \cdot \alpha(\lambda)\,. \tag{1}$$

In the case of a stable solution with n compounds, assuming that the Beer-Lambert law holds, it is possible to state that the measured spectrum will be a sum of the hypothetic individual spectra of each of the n constituents of the sample. In this case, one may propose that the measured spectrum is a linear combination of the base spectra of the compounds (Escalas *et al.*, 2003), *i.e.*

$$s(\lambda) = \sum_{i=1}^{n} \beta_i \cdot \alpha_i(\lambda)\,, \tag{2}$$

which expressed in vector notation becomes

$$s = \sum_{i=1}^{n} \alpha_i \beta_i = \begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \mathbf{A} \cdot \boldsymbol{\beta} \,. \tag{3}$$

Ideally, if calibration was performed correctly, the coefficient vector $\boldsymbol{\beta}$ will equal the concentration vector. It is here proposed, however, to add some level of flexibility to account for deviations from the Beer-Lambert law and consider a quadratic nonlinear relationship defined by the matrix **M**:

$$\begin{aligned} \boldsymbol{\beta} &= \varphi(\mathbf{x}) = \mathbf{x} + \mathbf{M} \cdot \boldsymbol{\xi}(\mathbf{x}) \\ \boldsymbol{\xi}(\mathbf{x}) &= \begin{bmatrix} x_1 x_2 & \cdots & x_1 x_n & x_2 x_3 & \cdots & x_{n-1} x_n \end{bmatrix}^T . \end{aligned} \tag{4}$$

As an extension of the SD method, this one is called *nonlinear spectral deconvolution* (NSD).

If the matrix of base spectra **A** and the nonlinearity matrix **M** are known, then, given a measured spectrum ***s***, to obtain an estimate $\hat{\mathbf{x}}$ of the concentrations in the sample a simple least squares procedure could be performed. However, if the blank reference is something other than wastewater, *e.g.* distilled water, then there always exists a nonzero spectrum of what is here called the *effluent, i.e.* water plus all other non-considered substances, even some suspended solids. Calibration of the matrix **A** is done performing a regression for each row of **A**. Furthermore, after calibration, it is assumed that a vector $\boldsymbol{\sigma} = [\sigma_1^2, \ldots, \sigma_N^2]^{\mathrm{T}}$ of estimated variances for each wavelength is obtained, i.e. σ_i^2 is a measure of fit for the linear regression at wavelength λ_i.

The concentration estimation procedure consists on two steps:
1. *Determination of* **β.** The following linear unconstrained optimization problem is solved: minimize

$$J_1(\boldsymbol{\beta}) = \left((\mathbf{s} - \mathbf{f}) - \mathbf{A} \cdot \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\right)^T \cdot \left((\mathbf{s} - \mathbf{f}) - \mathbf{A} \cdot \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\right) \tag{5}$$

with $\boldsymbol{f}$ the effluent spectrum and $\Sigma = \mathrm{diag}(\boldsymbol{\sigma})$. Then the estimate $\hat{\boldsymbol{\beta}}$ is the vector that minimizes the cost function $J_1(\boldsymbol{\beta})$. This is performed by usual least squares with known covariance.

2. *Determination of* **x**. The following nonlinear constrained optimization problem is solved: minimize

$$J_2(\mathbf{x}) = \left(\boldsymbol{\beta} - \varphi(\mathbf{x})\right)^T \cdot \left(\boldsymbol{\beta} - \varphi(\mathbf{x})\right) \tag{6}$$

subject to $x_i \geq 0$ for all $i=1,\ldots,n$. Then the estimated concentration $\hat{\mathbf{x}}$ is the vector that minimizes the cost function $J_2(\boldsymbol{\beta})$. Numerical optimization algorithms, *e.g.* Newton methods, must be used for this purpose and the structure of $\boldsymbol{\varphi}$ can be used to simplify such an algorithm.

Calibration

To calibrate the above methodology, a set of ℓ samples with known concentrations grouped in vectors $\boldsymbol{x}_j$ is required with $j=1,\ldots,\ell$. For each reference sample its absorbance spectrum $\boldsymbol{r}_j$ is measured and the corresponding $\boldsymbol{\xi}_j$ can also be calculated. The following matrices can therefore be constructed:

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_1 & \cdots & \mathbf{r}_\ell \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_\ell \end{bmatrix}, \quad \Xi = \begin{bmatrix} \boldsymbol{\xi}_1 & \cdots & \boldsymbol{\xi}_\ell \end{bmatrix}. \tag{7}$$

Calibration implies looking for estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{M}}$ such that

$$\mathbf{R} = \hat{\mathbf{A}}\left(\mathbf{X} + \hat{\mathbf{M}} \cdot \Xi\right) = \underbrace{\begin{bmatrix} \hat{\mathbf{A}} & \hat{\mathbf{A}} \cdot \hat{\mathbf{M}} \end{bmatrix}}_{\mathbf{Q}} \cdot \begin{bmatrix} \mathbf{X} \\ \Xi \end{bmatrix} . \tag{8}$$

For every row of the above equation, provided each row of **R** is assumed to be $\mathbf{r}_k^T$ for $k=1,\ldots,N$, the

corresponding row $\mathbf{q}_k^T$ of $\mathbf{Q}$ can be obtained by constrained linear least squares, *i.e.* minimizing

$$J(\boldsymbol{\beta}) = \left\| \mathbf{q}_k - \left[\mathbf{X}^T \quad \Xi^T \right] \cdot \bar{\mathbf{r}}_k \right\|^2 \tag{9}$$

subject to $q_{k,i} \geq 0$ for all i=1,...,ℓ. Furthermore if the unexplained part of each regression is assumed to have a certain probability distribution, *e.g.* normal, its variance σ_k^2 can be estimated by maximum likelihood optimization. Knowing $\hat{\mathbf{Q}}$, the first n columns are an estimate $\hat{\mathbf{A}}$, while the rest are an estimate of $\mathbf{A}$ $\mathbf{M}$. To obtain $\hat{\mathbf{M}}$, again usual least squares can be used and thus denoting as $\hat{\mathbf{Q}}_2$ this sub-matrix,

$$\hat{\mathbf{M}} = \left(\hat{\mathbf{A}}^T \hat{\mathbf{A}} \right)^{-1} \hat{\mathbf{A}}^T \hat{\mathbf{Q}}_2 . \tag{10}$$

Partial least squares

A widely used method to estimate concentrations or other parameters from spectrophotometric data is partial least squares regression (PLS) (Geladi and Kowalski, 1986). In this case it is assumed directly that

$$\boldsymbol{x} = \boldsymbol{B} \cdot \boldsymbol{s} + \boldsymbol{b}. \tag{11}$$

Consider that several outputs $y_1,...,y_p$ are linear combinations of several inputs $z_1,...,z_m$ with $m>p$, and these are grouped in row vectors $\boldsymbol{y}^T$ and $\boldsymbol{z}^T$ respectively. Assume also that n data pairs $(\boldsymbol{y}_i^T, \boldsymbol{z}_i^T)$ are available for i=1,...,n, such that the matrices $\boldsymbol{Y} \in \Re^{n \times p}$ and $\boldsymbol{Z} \in \Re^{n \times m}$ can be formed by stacking these row vectors. The linear model $\boldsymbol{F}=\boldsymbol{E} \cdot \boldsymbol{G}$ is proposed, where $\boldsymbol{E}$ and $\boldsymbol{Y}$ are $\boldsymbol{Z}$ and $\boldsymbol{Y}$ centred and normalized with respect to their columns, respectively.

It is not possible to use usual least squares to find an estimate when $n<m$, because $\boldsymbol{E}^T\boldsymbol{E}$ is likely to be singular. However, the m columns of $\boldsymbol{E}$ are assumed to be highly correlated with the p columns of $\boldsymbol{F}$. PLS therefore finds an orthogonal matrix $\boldsymbol{T}$ with only $r \leq n < m$ columns such that the independent variables are decomposed as $\boldsymbol{E}=\boldsymbol{T} \cdot \boldsymbol{P}^T$, while the dependent variables are *estimated* with $\boldsymbol{F}=\boldsymbol{T} \cdot \boldsymbol{C}^T$, *i.e.* both $\boldsymbol{E}$ and $\boldsymbol{F}$ are explained as linear combinations of the same basis. In particular, to propose this basis of so-called latent vectors $\boldsymbol{t}_i$, PLS finds two linear transformations $\boldsymbol{W}$ and $\boldsymbol{V}$ iteratively, such that with $\boldsymbol{T}=\boldsymbol{E} \cdot \boldsymbol{W}$ and $\boldsymbol{U}=\boldsymbol{F} \cdot \boldsymbol{V}$ the products $\boldsymbol{t}_i^T \boldsymbol{u}_i$ for i=1,...,r are maximized. In fact, the elements $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ are the left and right singular vectors of the spectral decomposition of $\boldsymbol{E}^T\boldsymbol{F}$ corresponding to the largest singular value, to which $\boldsymbol{t}_i^T \boldsymbol{u}_i$ is equal. This is equivalent to maximizing the covariance between $\boldsymbol{E}$ and $\boldsymbol{F}$. With the matrices $\boldsymbol{C}$ and $\boldsymbol{P}$ then estimate $\boldsymbol{G}= \boldsymbol{C} \cdot \boldsymbol{P}^+ = \boldsymbol{C} \cdot (\boldsymbol{P}^T\boldsymbol{P})^{-1} \boldsymbol{P}^T$.

Determining the size r of $\boldsymbol{T}$ is critical for obtaining a useful model, and techniques to avoid overparameterization, such as bootstrapping, are necessary. Assume the matrices $\boldsymbol{R}$ of reference spectra and $\boldsymbol{X}$ of known concentrations are given. By centring and normalizing $\boldsymbol{R}^T$ and $\boldsymbol{X}^T$ to obtain respectively $\boldsymbol{E}$ and $\boldsymbol{F}$ above, PLS is used to obtain $\boldsymbol{G}$, which is then denormalized and decentred to finally get $\mathbf{B}$ and $\mathbf{b}$.

APPLICATION RESULTS

A bioreactor operating as an anaerobic/aerobic SBR was implemented to biodegrade *p*-nitrophenol (PNP) with propionic acid as cosubstrate. During the anaerobic phase PNP is transformed to *p*-aminphenol (PAP), which is mineralized later during the aerobic phase. It is therefore important to determine the concentrations of PNP and PAP during the whole reaction. It has also been observed that propionic acid is quickly used at the beginning of the anaerobic phase in the reaction, as proton

donor for the anaerobic reactions, so this component is of no interest to monitor. Calibration was therefore performed only to determine the concentrations of PNP and PAP. In Figure 1 the concentrations of the samples used for this calibration are shown. The base spectra obtained with the NSD calibration method are also shown. It is clear from the figure that PNP and PAP have distinct base spectra and that indeed the effluent has a spectrum which must be considered in the deconvolution. Calibration of the PLS method was carried out using r=4, which was determined by stopping the iterations in the procedure when reconstruction of the output data was good enough.

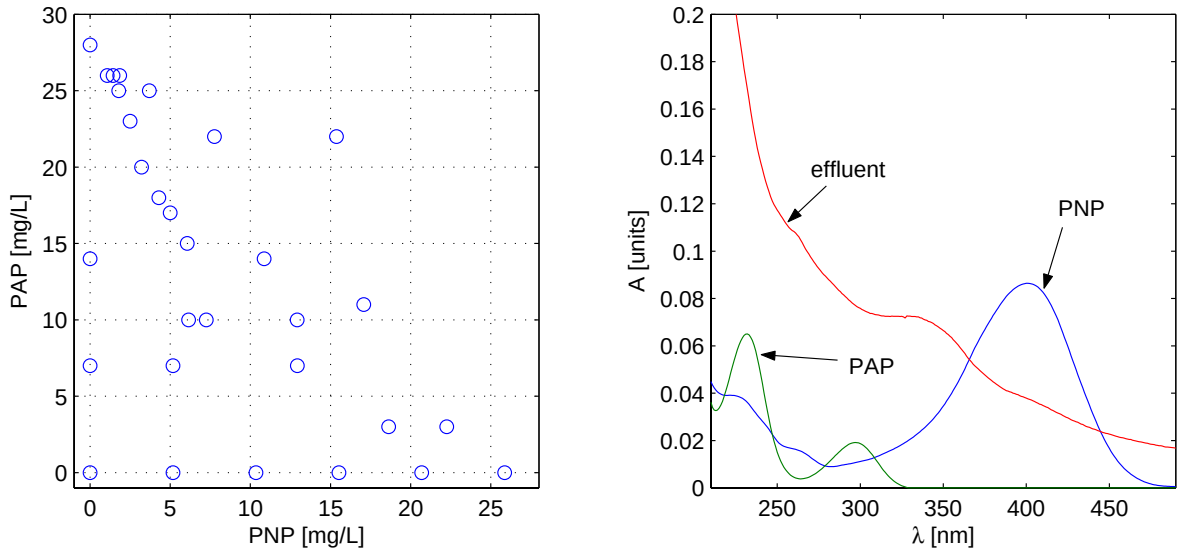

Figure 1: Data points used for calibration (left) and the resulting base spectra (right).

In Figure 2 the results obtained for the estimated kinetics of a reaction phase (both aerobic and anaerobic) are shown. Notice how at the beginning there exists a high concentration of PNP, which is reduced to PAP during the anaerobic phase. The PAP concentration then decreases to zero during the aerobic phase because of mineralization. This is well reproduced by the estimated kinetics. Results for other cycles are very similar and the method has been shown to be very robust experimentally.

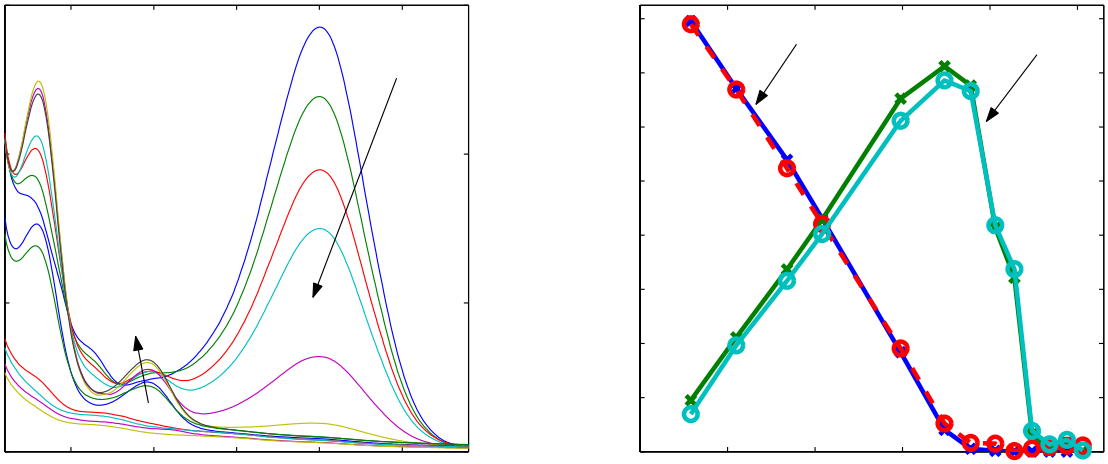

Figure 2: Spectra obtained for a reaction phase with initial PNP concentration of 18 mg/L (left) and the resulting estimated kinetics (right), both for NSD (-x-) and PLS (-o-).

CONCLUSIONS

Two methods for the estimation of concentrations in wastewater were developed and applied to obtain the kinetics of an anaerobic/aerobic SBR used to biodegrade *p*-nitrophenol. Both NSD and PLS are a viable alternative, although NSD seems to be more intuitive, since it can be checked how well the spectra are reconstructed with the deconvolution Additionally, it is well known that in PLS the choice of the parameter *r* (the number of columns in $\boldsymbol{T}$) is critical for obtaining good estimates and a wrong choice of *r* may deteriorate the estimation. This parameter is usually not known *a priori*. Under way of investigation is the setup of a completely automated sampling and measurement system in order to obtain data frequently enough to do mathematical modelling of dynamics of the process. In this sense, preliminary investigations show that it is possible to obtain concentration measurements as frequently as every 6 min if an on-line centrifuge for solid-liquid separation is used.

ACKNOWLEDGEMENTS

The authors wish to thank Leticia García and Eduardo Maciel for their eager participation in conducting the experiments and data collection. Furthermore, this research was partly supported by DGAPA-UNAM (IX104204 and IN106002). This paper also includes results of the EOLI project that is supported by the INCO program of the European Community (ICA4-CT-2002-10012).

REFERENCES

AFNOR, (1995). *Evaluation en milieu aqueux de la biodégradabilité aérobie "ultime" des produits organiques solubles. Méthode par analyse du carbone organique dissous (COD)*, Norme Française NF T 90–312, Paris.

APHA, AWWA, and WPCF (1992). *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association, New York.

Dahlén, J., Karlsson, S., Bäckström, M., Hagberg, J., and Pettersson, H. (2000). Determination of nitrate and other water quality parameters in groundwater from UV/Vis spectra employing partial least squares regression, *Chemosphere,* **40**, 71–77.

Escalas, A., Droguet, M., Guadayol, J., and Caixach, J. (2003). Estimating DOC regime in a wastewater treatment plant by UV deconvolution, *Wat. Res.*, **37**, 2627–2635.

Geladi, P. and Kowalski, B. (1986).Partial least-squares regression: a tutorial, *Anal. Chim. Acta*, **185**, 1–7.

Grady Jr., C., Daigger, G., and Lim, H. (1999). *Biological Wastewater Treatment*, 2nd ed, Marcel Dekker, New York.

Jeffrey, G., Basset, J., Mendham, J., and Denney, R., (eds) (1989). *Vogel's Textbook of Qualitative Chemical Analysis*, 5th ed., Longman Scientific and technical, Essex.

Langergraber, G., Fleischmann N., Hofstaedter, F., and Weingartner, A. (2004). Monitoring of a paper mill wastewater treatment plant using UV/Vis spectroscopy, *Wat. Sci. Tech.*, **49**(1), 9–14.

Melgoza, R. and Buitrón, G. (2001). Degradation of p-nitrophenol in a batch biofilter under sequential anaerobic/aerobic environments, *Wat. Sci. Tech.,* **44**(4), 151–157.

Thomas, O. and Constant, D. (2004). Trends in optical monitoring, *Wat. Sci. Tech.*, **49**(1), 1–8.

Thomas, O., Theraulaz, F., Agnel, C., and Suryani, S. (1996). Advanced UV examination of wastewater, *Environ. Technol.*, **17**, 251–261.

Wilderer, P., Irvine, R., and Goronszy, M. (2001). Sequencing Batch Reactor Technology, Vol. 10 of *Scientific and Technical Reports*, IWA Publishing, London.