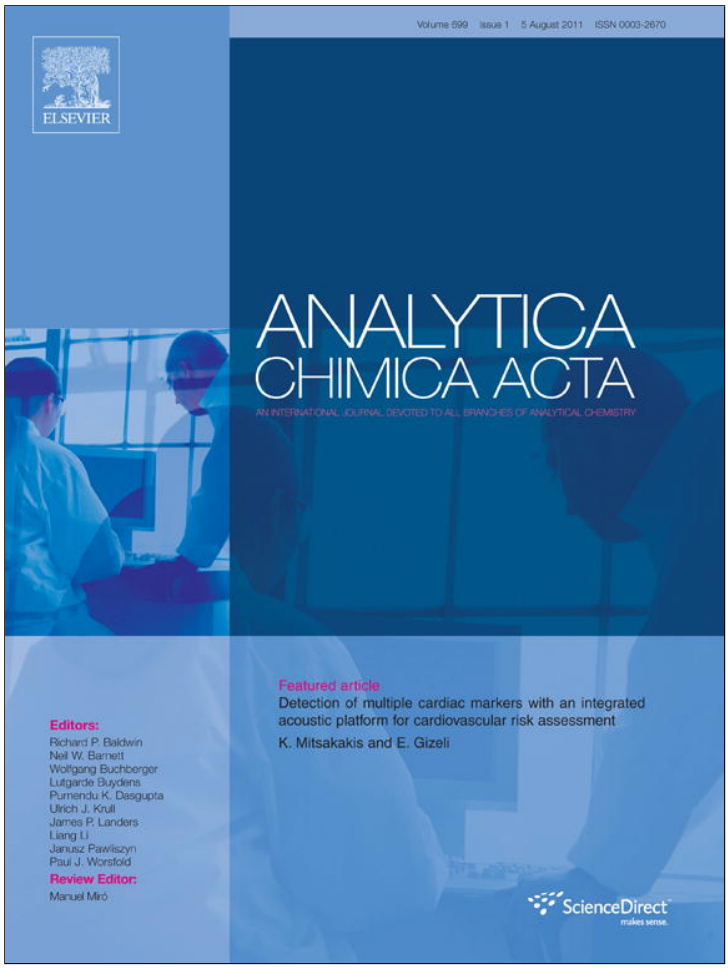

Analytica Chimica Acta 699 (2011) 18–25

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis

Franco Allegrini, Alejandro C. Olivieri*

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina

ARTICLE INFO

Article history:
Received 26 January 2011
Received in revised form 6 April 2011
Accepted 28 April 2011
Available online 11 May 2011

Keywords:
Ant colony optimization
Variable selection
Near infrared spectroscopy
Partial least-squares regression

ABSTRACT

A new variable selection algorithm is described, based on ant colony optimization (ACO). The algorithm aim is to choose, from a large number of available spectral wavelengths, those relevant to the estimation of analyte concentrations or sample properties when spectroscopic analysis is combined with multivariate calibration techniques such as partial least-squares (PLS) regression. The new algorithm employs the concept of cooperative pheromone accumulation, which is typical of ACO selection methods, and optimizes PLS models using a pre-defined number of variables, employing a Monte Carlo approach to discard irrelevant sensors. The performance has been tested on a simulated system, where it shows a significant superiority over other commonly employed selection methods, such as genetic algorithms. Several near infrared spectroscopic experimental data sets have been subjected to the present ACO algorithm, with PLS leading to improved analytical figures of merit upon wavelength selection. The method could be helpful in other chemometric activities such as classification or quantitative structure-activity relationship (QSAR) problems.

1. Introduction

Multivariate spectroscopic analysis intends to predict analyte concentrations or material properties from the spectrum of a given sample. Pertinent examples are the determinations of octane number in gasolines, glucose content in blood, oil concentration or moisture in seeds, and Brix degrees in sugar cane from near infrared (NIR) spectra [1]. For this purpose, a multivariate model is built which mathematically relates the spectra for a group of reference samples with their known property values. The multivariate model is usually of the inverse type, indicating that it considers the reference property values as a function of the matrix of collected spectra. The relationship between properties and spectra is expressed through the so-called vector of regression coefficients. This latter vector can be estimated in various ways, one of the most popular being partial least-squares regression (PLS, see below) [2]. Once estimated, this vector can be employed to predict the property of a new sample from its spectrum.

From the complete spectral data, which can be recorded for a given sample, it is likely that some of the signals may not be selective as regards the property of interest, while some others may be only partially selective. Hence, variables are usually subjected to a careful selection process before submitting them to PLS regression. This means that the multivariate model is built with only a limited number of signals. The purpose of variable selection is the obtainment of models based on spectral data carrying a higher information content as regards the analyte or property of interest. Additionally, less spectral overlapping with interferences is sought [3]. Improved PLS analytical performance has been reported upon variable selection, which supports the continuing interest in this chemometric activity [4–9]. The subject has been recently reviewed, with particular emphasis on NIR spectroscopic applications [10].

Two general types of variable selection methods are available: (1) inspecting the full spectral PLS regression coefficients or latent variables, and (2) searching for sensor ranges for which the prediction error is minimum. The simplest one, still advocated by many researchers, is the visual inspection of the spectrum of regression coefficients [3]. Variables for which the regression vector is significant are included in the PLS model, whereas those for which the regression vector is of low-intensity or noisy are removed. This simple strategy has been modified in various ways with a similar objective in mind [11–13]. However, it may be noticed that the intuitive power of regression coefficients to aid in variable selection has been challenged on a theoretical basis [14–17].

* Corresponding author. Tel.: +54 341 4372704; fax: +54 341 4372704.
E-mail addresses: olivieri@iquir-conicet.gov.ar, aolivier@fbioyf.unr.edu.ar (A.C. Olivieri).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.04.061

The search for sensor ranges where the predictive indicators are optimum constitutes a valid alternative for variable selection. Sensor ranges with improved analytical performance are assumed to correspond to spectral windows with higher information content regarding the analyte of interest. One of these methods is called interval-PLS (i-PLS). It builds a multivariate model in each of the spectral windows given by a fixed-size moving-window strategy [18]. The best spectral region for regression corresponds to the window providing the minimum prediction error. A more elaborate alternative employs variable-size windows, with the error indicator depending on both the first window sensor and the sensor width [19,20]. This latter method allows one to find regions with a width, which can be larger than the minimum window, but cannot locate regions, which combine separate sub-regions. An interesting derivation of i-PLS and window search has been recently described [21].

Since a fully comprehensive search may be prohibitively time consuming when the full spectral range includes a large number of sensors, such as those employed in visible/near infrared (Vis–NIR) spectroscopy, alternative strategies have been proposed, based on algorithms for global searches inspired in natural processes. Genetic algorithms (GA) are popular tools in this regard; they are based on concepts related to natural selection [22–26]. They proceed to select variables by assigning binary digits to selected and unselected features (i.e., 1 s and 0 s respectively), and to construct vectors ("chromosomes") of binary digits ("genes"). These vectors are sorted according to a certain objective function to be minimized, typically the average prediction error over a pre-determined set of samples. The best individuals are allowed to survive, breed and randomly mutate from one generation to the next one. The new offspring continues with this process until a certain number of generations elapse. The final best chromosome is assumed to encode the sought solution, in terms of selected features to be included in the multivariate model under scrutiny.

Recently, ant colony optimization (ACO) has been introduced for variable selection in PLS regression problems [27]. ACO resembles the behavior of ant colonies in the search for the best path to food sources [28]. Variables are identified with space dimensions defining the available paths followed by ants, with allowed coordinates of 1 or 0 (selected and unselected features respectively, as in GA). In this way, a given path is connected to a number of selected variables, which in turns corresponds to a given prediction error. In each generation, ants deposit a certain amount of pheromone, which increases with decreasing values of the objective function defined by each path. They find new paths based on the following information: (1) the pheromone amount accumulated in each of the dimension coordinates, (2) a heuristic measure of path goodness, and (3) a random search across all available paths. Ant search is then based on a probabilistic combination of these factors, which allow deviations from the best looking paths.

One potential problem with GA is rooted in its own fundamentals: randomness allows the algorithm to find new solution candidates and to avoid local minima. However, the solutions are different in different algorithm runs. GA have a tendency to include irrelevant variables in the final solutions together with those which are relevant to the problem under study. One alternative to avoid these problems is to run the GA several times, registering a statistics of the selected variables. The premise underlying this Monte Carlo-type methodology is the assumption that irrelevant variables are randomly selected, and repeated runs will tend to average out their appearance in the final solution. The relevant variables, on the other hand, will be persistently included. If the Monte Carlo results are presented in the form of a histogram, then selectable variables will appear as more intense peaks than irrelevant variables in this histogram.

Unfortunately, however, these expectations are not completely realized, and additional activities have been proposed to reach chemically reasonable solutions to the problem of feature selection. One of them involves the re-initialization of the GA with elitist chromosomes, i.e., those having 1 s for the variables corresponding to histogram peaks in a first GA run [25,29]. This leads to a certain improvement in variable selection in subsequent runs. The process is repeated again until the histogram stabilizes. This is the basis of the iteratively reinitialized genetic algorithm (IRGA) [25].

Another possibility is the introduction of chemically reasonable variable selection tools after the GA is run, avoiding the time consuming IRGA. For example, the combination of GA and i-PLS for weighting the histogram led to ranked regions genetic algorithm (RRGA) [26]. Another resource is the removal of irrelevant selection by testing their relative significance, using backward interval-PLS (bi-PLS) [30].

Variable selection based on ACO does also present, in principle, a similar problem. In previously reported papers, ACO-inspired algorithms were applied to the selection of variables aimed at the quantitative structure-activity relationship (QSAR) modeling of the inhibiting action of diarylimidazole derivatives on the enzime cyclooxygenase [31], the rate constants of *o*-methylation of phenol derivatives and activities of antifilarial antimycin compounds [32], the anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)-uracil derivatives [33], and the activity of glycogen synthase kinase-3β inhibitors [34]. ACO variable selection was also employed for improving a PLS regression analysis [27], with irrelevant variables being selected along with relevant ones. In none of these previous works Monte Carlo repeated calculations were attempted.

In this report we have applied both GA and ACO to PLS modeling of simulated and experimental data sets. Monte Carlo calculations show that GA and the already published ACO versions display an analogous behavior towards less relevant variables. However, a new and highly simplified ACO version which keeps most of the original ACO features produces stimulating results under Monte Carlo philosophy, different than those of the remaining algorithms. The selection was applied to several experimental sets of NIR data with improved analytical results.

2. Algorithms

2.1. Genetic algorithms

The GA applied in the present work has already been described [29]. In this case, however, we did not employ the final i-PLS weighting scheme, in order to compare all algorithms on the same basis, i.e., whit no post-processing procedures. The parameters for running the GA were similar to those employed for ACO (see below) in terms of number of blocks (and sensors per block), time steps, Monte Carlo cycles and maximum number of latent PLS variables. The number of chromosomes was equal to the number of ants in ACO algorithms.

2.2. ACO algorithms

Two ACO versions already described in the literature for variable selection have been employed, which will be called ACO-1 [31] and ACO-2 [27]. The basic MATLAB code for ACO-1 has been generously provided by Prof. Wu (Hunan University), and has only been modified in order to adapt it to the Monte Carlo-type calculations described in this paper. See Ref. [31] for details. The ACO-2 version was programmed in MATLAB according to the description given in Ref. [27], and then modified to incorporate Monte Carlo calculations.

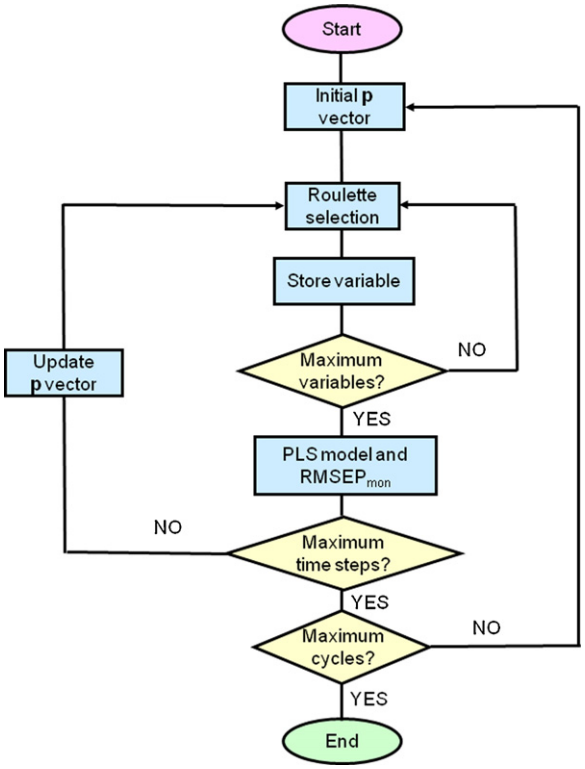

Fig. 1. Flow chart of the new ACO-3 algorithm based on ant colony optimization for the selection of relevant variables in PLS regression.

In the present work, a new algorithm will be described and called ACO-3. As in other wavelength selection strategies, in ACO-3 a variable to be selected is defined as a sensor range (or block of sensors) with a pre-defined spectral width. Additionally, the present ACO-3 methodology selects variables (or sensors blocks) one by one until a certain pre-defined maximum number of variables are chosen.

The purpose of introducing this new ACO-3 methodology is not to follow exactly the original ACO formulation (see Ref. [28]), but to develop a very simple algorithm inspired in the general philosophy of ant colony optimization, which could render acceptable selection results when coupled to Monte Carlo repeated calculations. In any case, the basic algorithm described in this report bears some resemblance with the so-called ant system discussed by Dorigo [35].

The ACO-3 flow chart shown in Fig. 1 compactly illustrates the proposed algorithm steps. A vector $\mathbf{p}$ of size $N \times 1$ is initially created, where N is the full number of available variables (i.e., blocks of individual spectral sensors). A generic vector element $p(n)$ collects the amount of pheromone at each time step which is associated to the nth variable. Initially, all elements of $\mathbf{p}$ are equal to 1, meaning that all variables have the same probability of being selected.

A certain number of variables (s) are then selected from the available N variables according to the pheromone content at the corresponding element of vector $\mathbf{p}$, using the roulette-wheel selection mode (Fig. 1). In this selection method, a fitness value is assigned to all possible variables, which is associated to a probability of selection. If $p(n)$ (the nth element of vector $\mathbf{p}$) is ascribed to the fitness of the nth variable, its probability of being selected is $prob(n)$:

$$prob(n) = \frac{p(n)}{\sum_{n=1}^{N} p(n)} \quad (1)$$

This could be imagined similar to a roulette wheel in a casino: a proportion of the wheel is assigned to each of the possible candidates based on their fitness values. This is achieved by dividing the fitness of a selection by the total fitness of all the selections, thereby normalizing them to 1. Then a random selection is made similar to how the roulette wheel is rotated. In our case, the fitness of each variable is given by the elements of the vector $\mathbf{p}$, which provides a probability of selecting a given variable. After selection, the $p(n)$ value for the latter variable is set to zero (to avoid duplication), and the selection starts again following the same roulette scheme, until all s variables have been selected. This provides the vector $\mathbf{v}$ (size $s \times 1$) of selected variables. Notice that in the first time step, all variables have the same probability of being selected, but as $\mathbf{p}$ is updated in successive time steps, these probabilities will differ.

With the selected variables, the $RMSE_{mon}$ (root mean square error) is estimated for the prediction of the property of interest in an independent monitoring set of samples (Fig. 1). This parameter is computed by building a PLS model with the calibration properties and signals, the latter being taken at the variables selected by the information carried by each ant. It should be noticed that a certain maximum number of latent PLS variables should be defined before program operation, and the optimum number of factors is estimated as the one leading to an $RMSE_{mon}$ value which is not statistically different than the minimum $RMSE_{mon}$, in order to avoid overfitting.

At successive time steps, the vector $\mathbf{p}$ is updated (see Fig. 1) according to:

$$\mathbf{p}(t) = (1 - \rho)\,\mathbf{p}(t-1) + \Delta\mathbf{p} \quad (2)$$

where t implies the current time step, ρ is the rate of pheromone evaporation ($\rho < 1$) and $\Delta\mathbf{p}$ is the vector of pheromone changes. The latter changes take place at certain variables, because each ant deposits pheromone at the vector element corresponding to its selected variable. Specifically, if $\mathbf{v}$ is the vector of selected variables, the contribution to $\Delta\mathbf{p}$ by a given ant occurs at the vector element with index $v(i)$ in such a way that:

$$\Delta p_a[v(i)] = -\log(RMSE_{mon})_a \quad (3)$$

where a identifies a particular ant.

Once the s variables are selected by each ant, Δp_a values for all variables and all ants are summed at the appropriate vector positions in order to obtain the $\Delta\mathbf{p}$ vector required in Eq. (2).

The above scheme shows that various ants may contribute to the same element of vector $\Delta\mathbf{p}$, demonstrating a cooperative behavior which is absent in GA. Notice in Eq. (3) that the contribution to $\Delta\mathbf{p}$ increases with decreasing monitoring error: the lower the error, the higher the pheromone content deposited by a specific ant in the corresponding variable. Hence the ant colony will give increasing importance (translated to larger pheromone deposits) to specific variables, which are expected to be the most significant ones for PLS property prediction.

Once completed the calculations for the required number of time steps, the whole cycle is repeated in order to register a histogram of the selected variables (Fig. 1). In this histogram, the variables selected at each calculation cycle are weighted inversely proportional to the final value of $RMSE_{mon}$, in order to give comparatively more importance to better solutions. After the histogram is obtained, it is scaled to a maximum value of 1, and the sensor blocks having histogram values larger than a pre-defined tolerance (for example, 0.3) are employed to build a final PLS regression model, which allows to compute a single $RMSE_{mon}$ for the monitoring set of samples, as well as the $RMSE_{test}$, computed for the independent set of test samples.

The required parameters for running this new ACO-3 version for variable selection are (suggestions in parenthesis): (1) ρ parameter (0.65), (2) number of ants (at least half the number of variables or

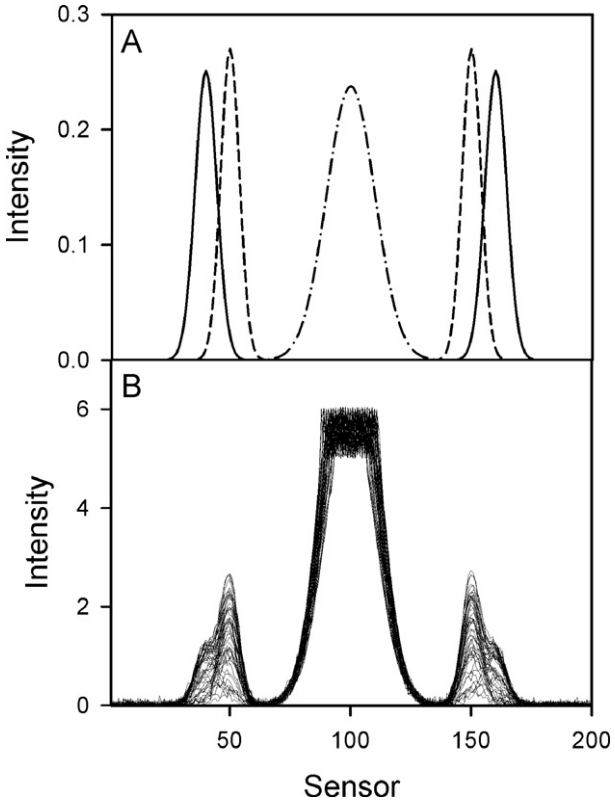

Fig. 2. (A) Pure component spectra at unit concentration, employed for building the simulated sets of spectra: solid line, analyte 1, dashed line, component 2, dashed-dotted line, component 3. (B) Spectra for the 50 calibration samples.

sensor blocks N), (3) sensor width of each range (which should be smaller than a typical spectral width of an absorption band), (4) maximum number of PLS latent variables (estimated by leave-one-out cross validations applied to the complete spectral range), (5) number of sensor ranges to be selected from each ant (selected by gradually increasing s until no significant $RMSE_{mon}$ changes are observed), (6) maximum time steps (50), (7) number of cycles or repeated calculations for histogram building (20).

The algorithm is given in Fig. 1 as a flow chart, and the complete code is provided as Supporting Information.

2.3. Software

All programs were written in MATLAB 7.10 [36], and are available from the authors on request.

3. Data sets

3.1. Simulated data

A synthetic data set was created by mimicking the spectra of three components, with component 1 being the analyte of interest. All constituents are present in 50 calibration samples, 100 monitoring samples and 100 test samples, at randomly chosen concentrations ranging from 0 to 5 units for analyte 1, from 0 to 10 for component 2, and from 25 to 50 units for component 3 (in the latter case to ensure high relative concentrations of component 3). Fig. 2A shows the pure component spectra, all at concentrations of 1 unit. From these noiseless profiles, calibration and test spectra were built. Specifically, each spectrum $\mathbf{x}$, whether belonging to the calibration, monitoring or test set, was created using the following expression:

$$\mathbf{x} = y_1\mathbf{s}_1 + y_2\mathbf{s}_2 + y_3\mathbf{s}_3 \quad (4)$$

where $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are the pure component spectra at unit concentration, and y_1, y_2 and y_3 are the component concentrations in a specific sample. Gaussian noise with a standard deviation of 0.01 units was added to all concentrations, before inserting them in Eq. (4). A vector of signal noise (standard deviation = 0.05 units) was then added to each $\mathbf{x}$ vector after applying Eq. (4). Signals higher than 5 units were cut at this latter value, and noise was added to them with 1 unit of standard deviation (this mimics the saturation of the detector at high absorbances in a real experiment). Fig. 2B shows the resulting matrix of calibration signals.

3.2. Experimental BRIX data

For building this first experimental data set, NIR spectra were measured for a series of sugar cane juices with a NIRSystems 6500 spectrometer, equipped with a cell with 1.0 mm optical path. Spectra were acquired using the spectrometer software ISISCAN, and then converted to ASCII files for further data processing. Reference Brix data were measured with a Leica AR600 refractometer. Sugar cane juices were analyzed at the quality control laboratory of the Estación Experimental Obispo Colombres, Tucumán, Argentina. The laboratory receives samples from several different cane processing units of the sugar-producing province of Tucumán. Cane samples are first processed in the sugar mills, where juice (65% of the cane) is extracted, and are then sent to the laboratory. For the calibration set, 59 samples were randomly selected, having Brix values in the range 11.76–23.15, as measured with the refractometer. The monitoring set was composed of 23 samples, and the test set of additional 23 samples with Brix values different than those employed for calibration. NIR spectra were measured in the wavelength range 400–2498 nm each 2 nm (i.e., 1050 data points).

3.3. Experimental OCTANE data

The second set consists of NIR spectra of gasoline samples collected in a local distillery, in the range 4020–9996 cm^{-1} each 12 cm^{-1} (499 data points) using a Bran + Luebbe Infraprover II FT NIR spectrophotometer. The corresponding octane numbers were determined by the reference method for research octane number of spark-ignition engine fuel [37]. The set was randomly divided into a 91-sample calibration set with octane numbers ranging from 91.0 to 97.6, and two different 45-sample sets employed as monitoring and test respectively.

3.4. Experimental CORN data

This is a data set which is freely available on the internet at http://www.eigenvector.com/data/Corn/. It consists of 80 samples of corn measured on three different NIR spectrometers. The wavelength range is 1100–2498 nm at 2 nm intervals (700 channels). The moisture, oil, protein and starch values for each of the samples are included. The data was originally taken at Cargill. A data set measured in one of the instruments was randomly divided into calibration (40 samples), monitoring (20 samples) and test (20 samples) sub-sets. For calibration, the parameter moisture was selected, with calibration values ranging from 9.41 to 10.88.

3.5. Experimental VISC data

This data set is also available on the internet, at http://www.eigenvector.com/data/SWRI/. It consists of NIR spectra of diesel fuels along with various properties including the

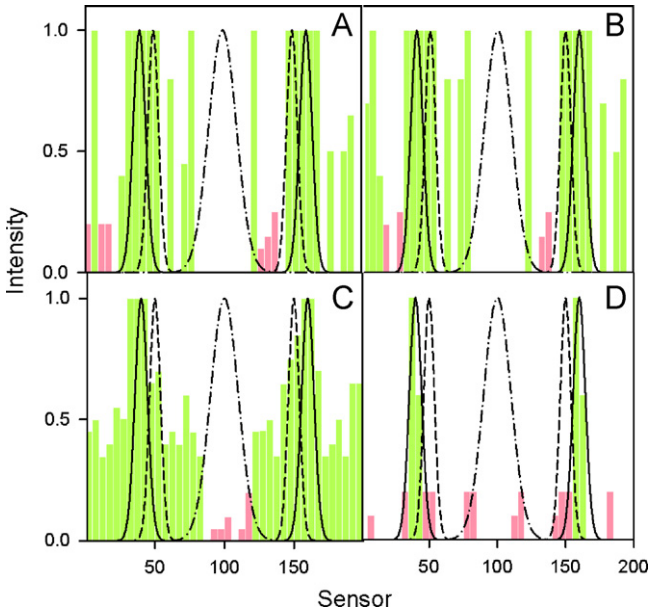

Fig. 3. Variable selection results for the simulated data set. (A) GA. (B) ACO-1. (C) ACO-2. (D) ACO-3. In all cases, the green bars show the histogram values larger than 0.3, and the red bars those with values smaller than 0.3. Superimposed are the spectra for the pure sample components: solid line, analyte 1, dashed line, component 2, dashed-dotted line, component 3, using an arbitrary vertical scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

viscosity measured at 40 °C. They were obtained at Southwest Research Institute (SWRI) on a project sponsored by the US Army. For calibration, 116 samples were selected at random (having viscosity values from 1.33 to 3.38), and two sub-sets of 58 samples each were chosen for monitoring and test respectively. Spectra were measured from 750 to 1550 nm each 2 nm.

4. Results and discussion

4.1. Simulated data

As explained above, the simulated data consist of signals for the analyte of interest, centered at sensors 40 and 160 (analyte 1) and for two additional components, one at sensors 50 and 150 (component 2) and one in the region 80-120 (component 3). Component 3 is included in large relative concentrations in the calibration, monitoring and test samples (Fig. 2A). These features are apparent in Fig. 2B. The application of variable selection algorithms is thus expected to extract, from these simulated data, the regions where analyte 1 responds, discarding at the same time the high interfering signal in the middle of the spectra, the overlapping signals from component 2, and the regions dominated by noise.

When GA, ACO-1 and ACO-2 were applied to this simulated system, the results showed that this expected outcome was not realized. In fact, Fig. 3A–C shows that GA, ACO-1 and ACO-2 render very similar results, in terms of inclusion of irrelevant regions. One favorable feature of these two algorithms, however, is the fact

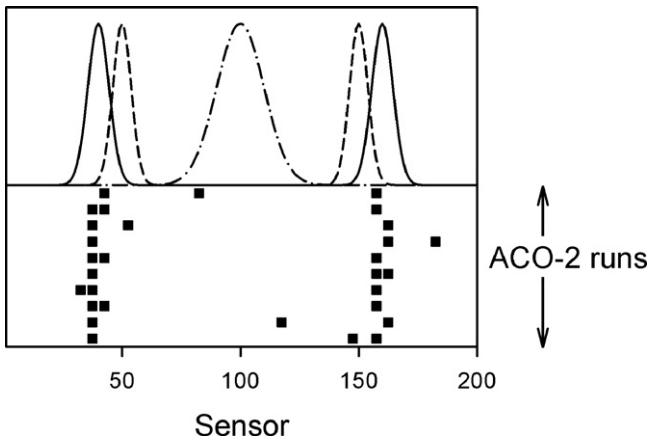

Fig. 4. Variables selected in the first 10 individual ACO-3 runs (solid squares) in the simulated system. The top plot shows the pure component spectra for reference.

that component 3 is avoided (Fig. 3A–C). However, as was observed before for GA in several simulated and experimental systems, a complement would be required in the form of post-processing the histogram of Fig. 3A with either i-PLS or bi-PLS procedures in order to get the expected selection output [26,29].

The goal of variable selection inspired in natural mechanisms, however, is to reach the expected answer without the help of intuition or complementary selection methods. Before application of ACO-3 to this simulated system, a parameter to be tuned is the number of variables s. In order to estimate it, ACO-3 was run for trial values of s ranging from 2 to 10, and for each of these values, the $RMSE_{mon}$ was computed by selecting the sensor blocks for which the histograms showed values larger than 0.3 (on a scale where the maximum histogram value was set to 1). The results showed that after an initial decrease of $RMSE_{mon}$ from 0.095 units to 0.070 units in going from $s = 2$ to $s = 3$, the final monitoring error stabilized in 0.070 units for larger values of s. Hence $s = 3$ seemed to be a reasonable choice for this ACO-3 parameter.

Application of ACO-3 with the parameters shown in Table 1 to the simulated data set provided the gratifying results displayed in Fig. 3D, where they can be compared with the remaining algorithms. The fact that ACO-3 provides the expected answer is also related to the procedure herein adopted of repeated histogram cycles. This activity is supported by Fig. 4, which shows the specific selected variables in the first 10 calculation cycles of ACO-3. As can be seen, while individual cycles select significant variables together with irrelevant ones, the latter ones are scattered throughout the cycles, in such a way that the averaging the histogram would provide comparatively higher importance to the variables, which are consistently being selected at each cycle. Precisely this outcome is the expected one on repeating the calculation cycles. Setting a reasonable cut-off for the histogram values shown in Fig. 3D at 0.3, the expected sensor regions where the analyte of interest responds are selected without the help of additional post-processing steps, corresponding to sensors 36–45 and 156–165.

After the selection process, PLS regression was applied to the independent test data set, in order to compare the performance

Table 1
Parameters for running ACO-3 in the different data sets.

Data set	Variables (s)	Total sensors/sensor width	Blocks	Time steps	Cycles	Maximum latent variables	Ants
Simulated	3	200/5	40	50	20	4	20
BRIX	5	1050/25	42	50	20	11	20
OCTANE	5	500/15	33	50	20	17	16
CORN	5	700/20	35	50	20	20	18
VISC	5	401/10	40	50	20	15	20

Table 2
Statistical predictive results for the independent test samples in the different data sets after applying various variable selection methods.[a]

Data set	Parameter	Selection method				
		None	GA	ACO-1	ACO-2	ACO-3
Simulated	Factors	4	3	3	3	2
	$RMSE_{test}$	0.076	0.059	0.056	0.062	0.054
	REP%	3.0	2.2	2.2	2.5	2.2
	R^2	0.9967	0.9980	0.9982	0.9978	0.9983
BRIX	Factors	11	10	9	9	7
	$RMSE_{test}$	0.79	0.27	0.30	0.31	0.24
	REP%	4.5	1.5	1.7	1.81	1.4
	R^2	0.9090	0.9894	0.9869	0.9860	0.9916
OCTANE	Factors	17	12	14	14	5
	$RMSE_{test}$	0.52	0.39	0.44	0.34	0.30
	REP%	0.55	0.42	0.47	0.36	0.32
	R^2	0.8136	0.8952	0.8666	0.9203	0.9380
CORN	Factors	18	14	17	14	13
	$RMSE_{test}$	0.0072	0.070	0.005	0.013	0.001
	REP%	0.073	0.690	0.050	0.13	0.01
	R^2	0.9997	0.9563	0.9696	0.9210	0.9939
VISC	Factors	12	8	4	12	12
	$RMSE_{test}$	0.11	0.11	0.24	0.087	0.087
	REP%	4.7	4.4	10.2	3.7	3.7
	R^2	0.8940	0.8912	0.4830	0.9337	0.9337

[a] $RMSE_{test}$ = root mean square error in the test sample set; REP% = relative error of prediction; R^2 = correlation coefficient. Units are as follows, for BRIX data, BD (Brix degrees), for OCTANE data, octane units, for CORN data, %, and for VISC data, viscosity units.

on samples not employed for either calibration or monitoring. The results are provided in Table 2. Two facts should be noticed on inspecting this table: (1) the $RMSE_{test}$ value for the test set decreases in going from the full spectral data to the selected regions, accompanied by a decrease in the relative error of prediction or ($REP\% = 100\,RMSE_{test}/\bar{y}_{cal}$, where $\bar{y}_{cal}$ is the mean calibration value of the target property) and a slight improvement in the correlation coefficient R^2 between predicted and nominal property values,

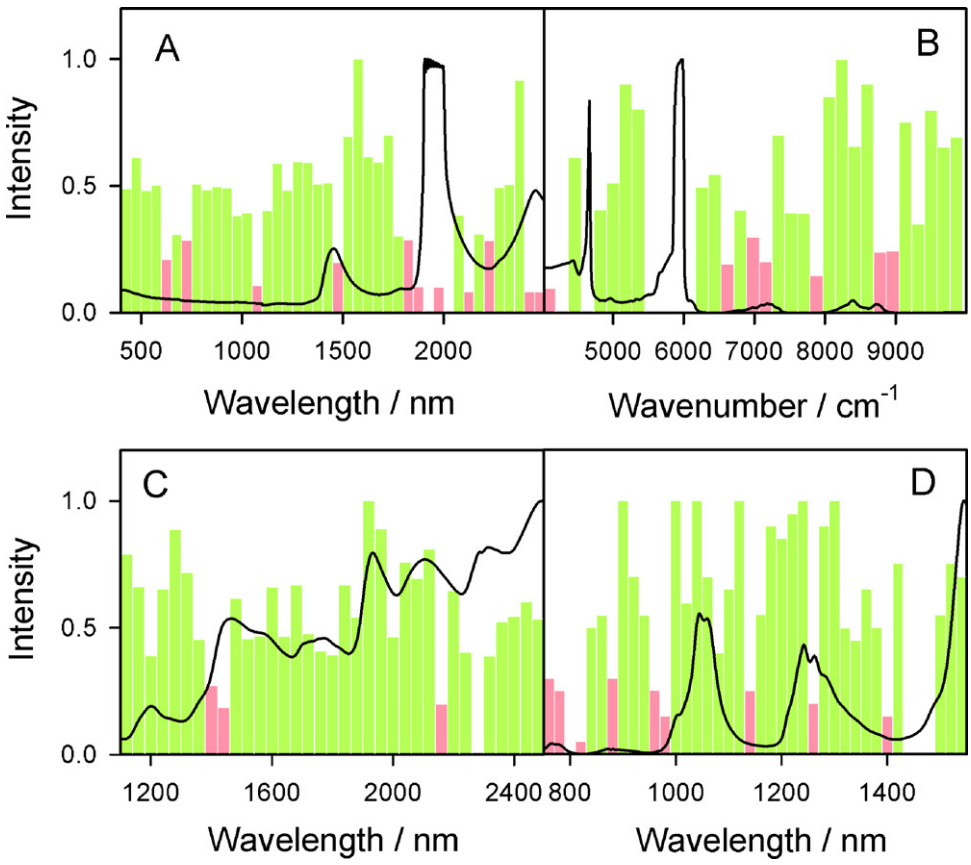

Fig. 5. GA variable selection results for the experimental data sets. (A) BRIX. (B) OCTANE. (C) CORN. (D) VISC. In all cases, the green bars show the histogram values larger than 0.3, and the red bars those with values smaller than 0.3. Superimposed are the mean calibration spectra arbitrary vertical scales. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

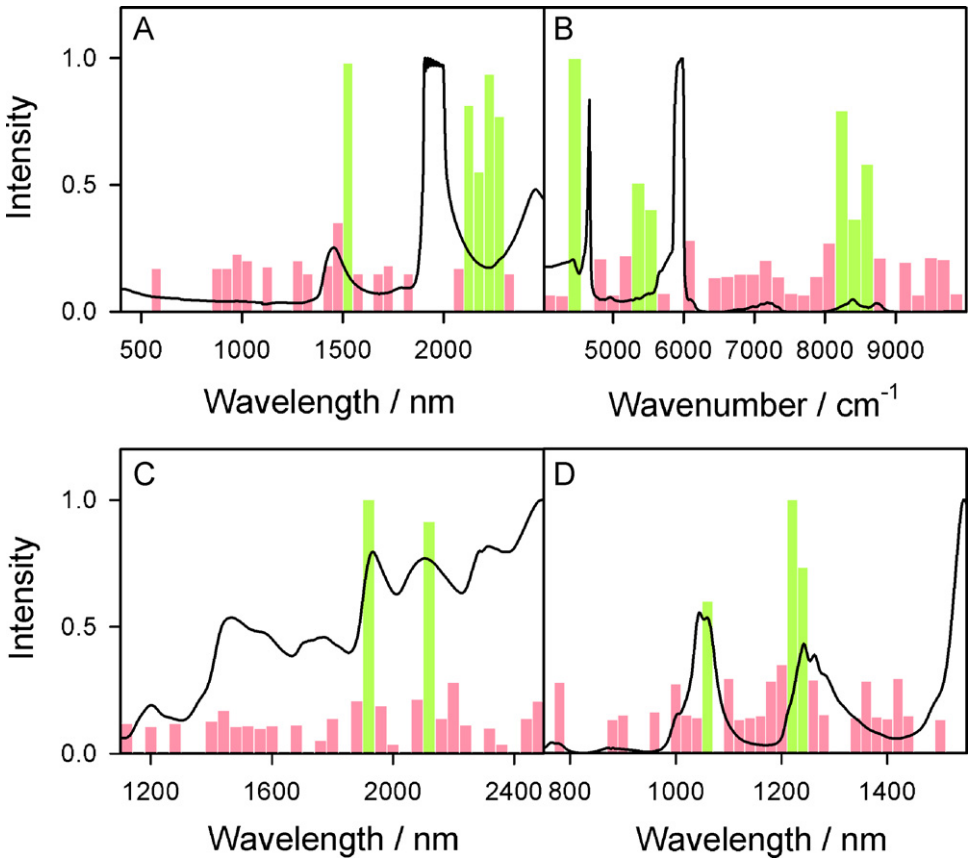

Fig. 6. ACO-3 variable selection results for the experimental data sets. (A) BRIX. (B) OCTANE. (C) CORN. (D) VISC. In all cases, the green bars show the histogram values larger than 0.3, and the red bars those with values smaller than 0.3. Superimposed are the mean calibration spectra arbitrary vertical scales. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

and (2) the number of calibration PLS latent variables does also decrease, because less responsive components occur in the selected regions in comparison with the full spectrum. Table 2 does also show that ACO-3 leads to the minimum number of latent variables and prediction error.

4.2. BRIX data

Brix analysis is a relevant industrial parameter characterizing sugar cane juice. The Brix degree is a unit representative of the sugar content of an aqueous solution. One degree Brix corresponds to 1% by weight (% w/w) of sucrose in a solution. Traditionally, the Brix degrees are measured with a refractometer, which measures the refractive index of a solution, although it is being increasingly determined by NIR spectroscopy.

The main spectral features of the BRIX data set involve a high-absorbance signal due to water (around 1950 nm), regions with significant signals at 1450 and 2500 nm, as well as regions which are mainly dominated by noise below 1300 nm (see Fig. 5A). A previous analysis of the useful regions for PLS estimation of the Brix degrees in this system was made using a GA, requiring i-PLS post-processing in order to reach a reasonable answer [29]. Fig. 5A shows the GA results without the latter post-processing, i.e., the raw sensor blocks selected when only GA variable selection is employed. As can be seen, both relevant and irrelevant regions are equally selected. The results using both ACO-1 and ACO-2 (not shown) were similar to those obtained by GA.

ACO-3 was applied with the parameters quoted in Table 1 to the BRIX data set, with results, in terms of selected sensor blocks, which are shown in Fig. 6A. The result is encouraging, because both the high water signal and the noisy regions are avoided, as expected. The selected regions (1500–1548 and 2100–2298 nm) are similar to those previously found by GA/i-PLS variable selection [29]. Using these regions to build the PLS regression model, the RMSE results for an independent test sample set, quoted in Table 2, represent a significant improvement over the full spectral results. Additional improvements in REP% and R^2 can also be seen (Table 2). The number of calibration PLS latent variables, on the other hand, decreases in going from full spectra to selected sensor blocks, on account of the removal of signals which are uncorrelated to the Brix degrees to be estimated by the model. In comparison with other selection methods, ACO-3 produces the best results, with lower calibration latent variables and prediction error.

4.3. OCTANE data

The octane number is a measure of the resistance of petrol and other fuels to autoignition in spark-ignition internal combustion engines. It is measured in a test engine, and is defined by comparison with the mixture of 2,2,4-trimethylpentane (*iso*-octane) and *n*-heptane which would have the same anti-knocking capacity as the fuel under test: the percentage, by volume, of 2,2,4-trimethylpentane in that mixture is the octane number of the fuel. The most common type of octane rating worldwide is the research octane number (RON), which is determined by running the fuel in a test engine with a variable compression ratio under controlled conditions, and comparing the results with those for mixtures of iso-octane and *n*-heptane. Much more conveniently, NIR spectroscopy allows for the fast and accurate measurement of octane number.

This OCTANE data set was previously studied using different GA versions. Without i-PLS post-processing, the results look rather

unselective (see Fig. 5B). Similar discouraging results were obtained on applying the ACO-1 and ACO-2 versions to this system. For improving these raw GA results, the combination of GA and i-PLS was developed and named RRGA (rank ranges genetic algorithm) [26].

Application of ACO-3 provides a more reasonable answer than the remaining algorithms, which agrees in some of the selected regions with more sophisticated GA versions [24,25] and other selection methodologies [38]. In particular, the sensitive regions 5300–5700 and 8200–8800 cm^{-1} are consistently selected (Fig. 6B). Table 2 implies a significant improvement in ACO-3 statistical parameters with respect to the full spectral analysis, with lower number of calibration variables and better statistical indicators in comparison with the remaining selection algorithms.

4.4. CORN data

This data set is available on the internet, and is intended for calibration of the moisture content in corn seeds. Application of GA to this data set provides the results shown in Fig. 5C. Similar results were obtained applying ACO-1 and ACO-2 methodologies. Many different blocks are selected, as with other data sets. Recently, a parallelized GA was applied to this data set, including in the chromosomes the possibility of selecting a wide variety of preprocessing steps [39]. The selection results are also similar to those herein reported.

In the case of ACO-3, it is interesting to notice that the highest intensity in the histogram corresponds to the sensor block for the ranges 1900–1938 and 2100–2138 nm (Fig. 6C). The former one is close to one of the highly absorbing NIR bands of pure water at 1940 nm [40]. This is understandable, since the calibration for moisture in this data set should be most sensitive when the analysis is focused on the water absorption band.

Table 2 shows a significant improvement in predictive ability on sensor selection, as judged from the reported statistical indicators, with a final $RMSE_{test}$ value for the test samples which is comparable to that recently reported for a sophisticated GA version [39]. Also, the results are better than those provided by other selection algorithms.

4.5. VISC data

This data set is also available on the internet, and is supposed to provide a test field for variable selection tools. When the GA is applied to this system, Fig. 5D is obtained, where it is apparent that relevant absorption bands are selected, together with regions of very low signal intensity. ACO-1 and ACO-2 provided comparable results.

When ACO-3 was applied to this data set, the selected wavelengths corresponded to well-defined absorption peaks in the NIR spectra, located at 1050–1068 and 1210–1248 nm (Fig. 6D). Building a PLS model with these selected regions led to improved viscosity prediction (Table 2). The RMSE quoted in Table 2 for the independent test sample set (0.087 units) is close the one reached by Westerhuis et al. [41] using the full spectral information analyzed by a rather sophisticated version of pre-processed partial least-squares regression called direct orthogonal signal correction (DOSC), which allowed to reach a mean error of 0.08–0.09 units. The improvements in REP% and R^2 are also apparent, comparable to those achieved by ACO-2 (Table 2).

5. Conclusions

Both simulations and experimental information show that a new variable selection model based on ant colony optimization, combined with Monte Carlo repeated calculations, is highly useful in discarding irrelevant spectral regions when partial least-squares regression analysis is performed on spectroscopic data. The results could be beneficial for other research areas such as QSAR, where variable selection is often crucial for the success of the correlation.

Acknowledgment

Financial support from the University of Rosario and CONICET (Project No. PIP 1950) is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.aca.2011.04.061.

References

[1] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-Infrared Spectroscopy: Principles, Instruments, Applications, Wiley-VCH, Weinheim, Germany, 2002.
[2] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[3] R.K.H. Galvao, M.C.U. Araujo, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Amsterdam, 2009, p. 233.
[4] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Anal. Chim. Acta 348 (1977) 71–86.
[5] Q. Ding, G.W. Small, M.A. Arnold, Anal. Chem. 70 (1998) 4472–4479.
[6] K. Hasegawa, T. Kimura, K. Funatsu, Quant. Struct. Act. Relat. 18 (1999) 262–272.
[7] A.S. Bangalore, R.E. Shaffer, G.W. Small, Anal. Chem. 68 (1996) 4200–4212.
[8] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35–44.
[9] J.-P. Gauchi, P. Chagnon, Chemom. Intell. Lab. Syst. 58 (2001) 171–193.
[10] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14–32.
[11] P.J. Gemperline, J.R. Long, V.G. Gregoriou, Anal. Chem. 63 (1991) 2313–2323.
[12] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851–3858.
[13] I.-G. Chong, C.-H. Jun, Chemom. Intell. Lab. Syst. 78 (2005) 103–112.
[14] M.B. Seasholtz, B.R. Kowalski, Appl. Spectrosc. 44 (1990) 1337–1348.
[15] O.M. Kvalheim, T.V. Karstang, Chemom. Intell. Lab. Syst. 7 (1989) 39–51.
[16] A.J. Burnham, J.F. MacGregor, R. Viveros, J. Chemom. 15 (2001) 265–284.
[17] C.D. Brown, R.L. Green, Trends Anal. Chem. 28 (2009) 506–514.
[18] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[19] H.C. Goicoechea, A.C. Olivieri, Analyst 124 (1999) 725–731.
[20] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Anal. Chim. Acta 501 (2004) 183–191.
[21] J.A. Cramer, K.E. Kramer, K.J. Johnson, R.E. Morris, S.L. Rose-Pehrsson, Chemom. Intell. Lab. Syst. 92 (2008) 13–21.
[22] R. Leardi, M.B. Seasholtz, R.J. Pell, Anal. Chim. Acta 461 (2002) 189–200.
[23] R. Leardi, A. Lupiáñez González, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[24] H.C. Goicoechea, A.C. Olivieri, J. Chem. Inf. Comp. Sci. 42 (2002) 1146–1153.
[25] C.E. Boschetti, A.C. Olivieri, J. NIR Spectrosc. 12 (2004) 85–91.
[26] H.C. Goicoechea, A.C. Olivieri, J. Chemom. 17 (2003) 338–345.
[27] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, J. Chemom. 20 (2006) 146–157.
[28] M. Dorigo, T. Stützle, Ant Colony Optimization, The MIT Press, Cambridge, MA, USA, 2004.
[29] N. Sorol, E. Arancibia, S.A. Bortolato, A.C. Olivieri, Chemom. Intell. Lab. Syst. 102 (2010) 100–109.
[30] R. Leardi, L. Nørgaard, J. Chemom. 18 (2004) 486–497.
[31] Q. Shen, J.-H. Jiang, J.-C. Tao, G.-L. Shen, R.-Q. Yu, J. Chem. Inf. Model. 45 (2005) 1024–1029.
[32] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Anal. Chim. Acta 646 (2009) 39–46.
[33] M. Goodarzi, M.P. Freitas, R. Jensen, Chemom. Intell. Lab. Syst. 98 (2009) 123–129.
[34] M. Goodarzi, M.P. Freitas, R. Jensen, J. Chem. Inf. Model. 49 (2009) 824–832.
[35] M. Dorigo, Optimization, Learning and Natural Algorithms, PhD Thesis, Politecnico di Milano, Milan, Italy, 1992.
[36] MATLAB 7.10, The MathWorks Inc., Natick, MA, 2010.
[37] ASTM Method D 2699-99, Annual Book of ASTM Standards, vol. 05.05. ASTM, West Conshohocken, PA, USA, 2001.
[38] H. Chung, H. Lee, C.-H. Jun, Bull. Korean Chem. Soc. 22 (2001) 37–42.
[39] O. Devos, L. Duponchel, Chemom. Intell. Lab. Syst., doi:10.1016/j.chemolab.2011.01.008, in press.
[40] J.A. Curcio, C.C. Petty, J. Opt. Soc. Am. 41 (1951) 302–1302.
[41] A. Westerhuis, S. de Jong, A.K. Smilde, Chemom. Intell. Lab. Syst. 56 (2001) 13–25.