Talanta 97 (2012) 242–248

Contents lists available at SciVerse ScienceDirect

Talanta

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

Uncovering interactions in Plackett–Burman screening designs applied to analytical systems. A Monte Carlo ant colony optimization approach

Alejandro C. Olivieri [a,*], Jorge F. Magallanes [b]

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

[b] Comisión Nacional de Energía Atómica, Gerencia Química, Av. Gral. Paz 1499, B1650KNA San Martín, Pcia. de Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 3 January 2012
Received in revised form
9 April 2012
Accepted 18 April 2012
Available online 27 April 2012

Keywords:
Plackett–Burman designs
Factor associations
Ant colony optimization

ABSTRACT

Screening of relevant factors using Plackett–Burman designs is usual in analytical chemistry. It relies on the assumption that factor interactions are negligible; however, failure of recognizing such interactions may lead to incorrect results. Factor associations can be revealed by feature selection techniques such as ant colony optimization. This method has been combined with a Monte Carlo approach, developing a new algorithm for assessing both main and interaction terms when analyzing the influence of experimental factors through a Plackett–Burman design of experiments. The results for both simulated and analytically relevant experimental systems show excellent agreement with previous approaches, highlighting the importance of considering potential interactions when conducting a screening search.

1. Introduction

When many different experimental factors have a potential influence on a chemical response, and the specific response-factor relationship is not known, it is wise to resort to statistical techniques known as design and optimization of experiments (DOE) [1]. Applications in chemistry abound, specifically in analytical chemistry studies, as has been beautifully described in a recent tutorial [2]. Common application areas for DOE comprise chromatography [3], capillary electrophoresis [4], phosphorimetry [5], electrochemistry [6], etc.

When the number of potentially influencing factors on a given response is large, it is advisable to conduct a screening study before the optimization phase, in order to detect the relevant factors. Screening models intend to explain the system behavior as a function of the experimental factors [7–10], usually by rationally devising a reasonably small number of experimental runs. Two-level full factorial designs, for example, are appealing in this regard, but are only capable of describing linear relationships and associations between factors. They require 2^k runs to build a model, where k is the number of factors. However, for more than three factors, the number of runs may be uneconomical, and therefore fractional factorial designs have been devised, which save experiments by dividing the number of runs by powers of 2. As a consequence, the possibility of independently estimating main and two-factor interacting terms is lost, and only estimations of confounded effects can be obtained [8,9].

A very popular and extremely economic design in terms of number of experimental runs is the one devised by Plackett and Burman (PB) [11]. A 12-experiment PB design allows one to study the effect of up to 11 factors. In comparison, two-level full factorial designs require 32 runs for 5 factors, 64 for 6 factors, etc. PB designs can only estimate main factors, while association terms are confounded with main effects or other associations. The PB confounding pattern is complex: every main factor is partially confounded with all possible two-factor interactions not involving the factor in question. It is important to notice that the validity of a PB screening design to estimate the main factor effects depends on the assumption that the interaction effects are negligible [8]. Ignoring interactions when they are indeed present may lead to the following undesired effects: (1) missing important effects, (2) including irrelevant effects in subsequent optimization stages, and/or (3) mistaking effect signs, leading to incorrect factor levels.

Because PB designs are very appealing due to the apparent economy of experimental runs, it is of interest to probe for strategies that allow to estimate the importance of interactions terms. The latter can be uncovered in PB designs by: (1) regression guided by the alias matrix [12], (2) frequentist analysis, (3) Bayesian–Gibbs analysis [13–20], (4) Danzig selection [21], and (5) genetic algorithms [14]. The latter belongs to a group of variable selection chemometric activities mimicking natural processes [22], such as particle swarm optimization [23] and the recently introduced ant colony optimization (ACO) [24]. The ACO approach has been employed for feature

* Corresponding author. Tel./fax: +54 341 4372704.
E-mail address: olivieri@iquir-conicet.gov.ar (A.C. Olivieri).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.04.025

selection in QSAR [25–27] and in partial least-squares (PLS) regression models [28,29].

The ACO algorithm mimics the behavior of ant colonies in the search for the best path to food sources [24]. Features to be selected are identified with space dimensions defining the available paths followed by ants, with allowed coordinates of 1 or 0 (selected and unselected features, respectively). A given path is thus connected to a number of selected variables, which in turns corresponds to a given value of the objective function to be minimized. In each generation, ants deposit a certain amount of pheromone in their paths, which increases with decreasing values of the objective function. During the evolution of the ant colony, they find new and better paths, based on a probabilistic combination of the following factors: (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.

In the present report we describe a strategy based on Monte Carlo ant colony optimization to uncover factor associations in PB designs with analytical implications. A suitably adapted algorithm has been applied to the analysis of a simulated PB design and also to experimental ones which are relevant to analytical chemistry.

2. Theory

2.1. Screening designs

The purpose of an experimental design is to build a statistical model for estimating the system response as a function of the values of a certain number of factors. When two-level models are considered for evaluating main factors and two-factor interactions, the following expression can be employed for modeling:

$$y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} \sum_{j=i+1}^{k} b_{ij} x_i x_j + e \quad (1)$$

where y is the system response, b are coefficients to be estimated, x_i and x_j are model factors, k is the total number of factors, and e collects the model error. The model matrix includes the k columns corresponding to the main factors x_i (i=1, 2, …, k), and the [k (k–1)/2] columns corresponding to the interacting terms (x_i x_j) (for $i \neq j$) (the need of the intercept b_0 is usually removed by mean centering the data). The total number of main and interacting terms is thus [k(k+1)/2]. For $k > 4$, this latter number is larger than the 12 runs required by the minimum PB design, leading to a rank-deficient model matrix which cannot solve for both main and two-factor associations directly.

Classicallle during PB analysis. The PB model matrix $\mathbf{X}_0$ (size $12 \times k$), with k columns corresponding to the main factors only, is full-rank. It is employed to find the $\mathbf{b}$ coefficients through:

$$\mathbf{b} = \mathbf{X}_0^{+}\mathbf{y} \quad (2)$$

where $\mathbf{y}$ is the vector of responses and the superscript '+' stands for the generalized inverse of a matrix. However, if interactions are present, the model may lead to erroneous results concerning the significance of the coefficients.

In practice, certain models including only some of the main factors and their two-factor interactions may lead to full-rank model matrices $\mathbf{X}$ (size $12 \times s$, where s is the number of included main terms and two-factor interacting terms). This may allow for a direct least-squares solution of the selected model coefficients. These models are usually assumed to comply with the following principles: (1) factor sparsity, meaning that only a few factors are significant, and (2) heredity, implying that significant interactions occur between factors when at least one of them is in itself significant. The existence of these models can be searched guided by a significant improvement in model fit, i.e., a decrease in the root mean square error (RMSE), which is given by:

$$\mathrm{RMSE} = \|\mathbf{y} - \mathbf{X}\mathbf{X}^{+}\mathbf{y}\| / (\mathrm{DOF})^{1/2} \quad (3)$$

where $\| \ \|$ indicates the Euclidean norm of a vector, $\mathbf{X}$ the model matrix including a certain number of main and two-factor interacting terms, and DOF the degrees of freedom, equal to the number of experimental runs minus the number of coefficients to be estimated.

The above discussion implies that feature selection algorithms can be adapted to find the best fitting models, as was the case with genetic algorithms [14].

2.2. Ant colony optimization

The ACO flow chart shown in Fig. 1 compactly illustrates the proposed algorithm steps. The aim is to find a suitable model including s terms, selected from the full number of terms N (i.e., the sum of the numbers of main terms and two-factor associations). The setup of the algorithm is similar to that recently discussed for PLS regression, except that in the present case certain combinations of features are not allowed. This occurs if the resulting model matrix $\mathbf{X}$ (size $12 \times s$) is not full rank, or if the association terms do not comply with the heredity principle. These latter models are easily discarded by assigning them arbitrarily large values of the objective function to be minimized.

At the start, a vector $\mathbf{p}$ of size $N \times 1$ is defined, whose generic element $p(n)$ represents the amount of ant pheromone at a given evolution time, associated to the nth. term. Initially, all elements

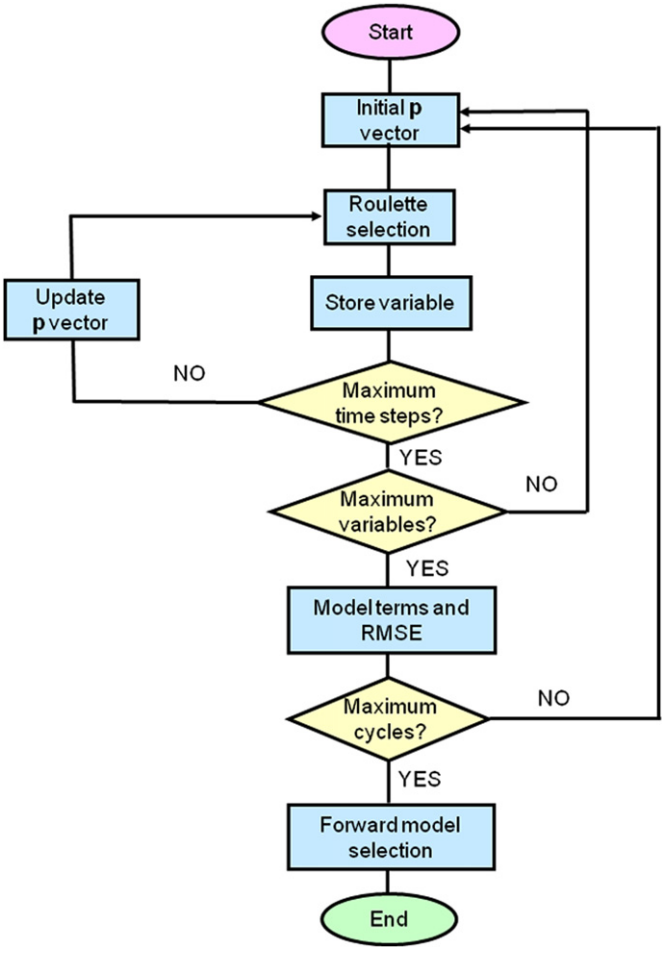

Fig. 1. Flow chart of the Monte Carlo ant colony optimization algorithm for feature selection in screening designs.

of **p** are equal to 1, meaning that all terms have the same probability of being selected. Then, s terms are selected from the available N according to the pheromone content $p(n)$, using the roulette-wheel selection mode. In this selection method, a fitness value equal to $p(n)$ is assigned to each term, and its probability $prob(n)$ of being selected is:

$$prob(n) = \frac{p(n)}{\sum_{n=1}^{N} p(n)} \tag{4}$$

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 (normalized to a sum of fitness values equal to 1). Then, a random selection is made similar to how the roulette wheel is rotated, but with wheel sections having areas proportional to $prob(n)$. After selection of a given term, its $p(n)$ value is set to zero to avoid duplication, and the selection starts again following the same roulette scheme, until s terms have been selected. Notice that in the first step, all variables have the same probability of being selected, but as **p** is updated in successive steps, these probabilities will differ.

Having selected the s terms, a model matrix **X** is built joining the corresponding columns of selected main and two-factor interacting terms, and the coefficients are estimated by least-squares, leading to a given RMSE value [Eq. (3)]. If the model matrix does not meet the heredity principle or is not full-rank, then an arbitrarily large RMSE is assigned.

The vector **p** is updated at successive time steps through the vector $\Delta\mathbf{p}$ accounting for pheromone changes. A given ant contributes to the change in pheromone associated to the nth. selected term according to:

$$\Delta p_{an} = -\log(\mathrm{RMSE})_a \quad \text{if the } n\text{th. term is included} \tag{5}$$

$$\Delta p_{an} = 0 \quad \text{otherwise} \tag{6}$$

where a identifies a particular ant. Observe that the pheromone content increases with decreasing error [Eq. (5)]: the lower the error, the higher the pheromone content deposited by a specific ant in the corresponding term.

The pheromone changes are then summed for all terms and all ants in order to obtain the $\Delta\mathbf{p}$ vector:

$$\Delta\mathbf{p} = \sum_{a=1}^{A}\sum_{n=1}^{N} \Delta p_{an} \tag{7}$$

where A is the total number of ants. The vector **p** is then updated according to:

$$\mathbf{p}(t) = (1-\rho)\mathbf{p}(t-1) + \Delta\mathbf{p} \tag{8}$$

where t is the current time step and ρ is the rate of pheromone evaporation ($\rho < 1$). The latter parameter controls the speed at which the trail left by ants disappears. If ants deposit pheromone continuously on a certain path, this effect tends to reinforce the selection of the path; conversely, if they do not visit a given path for a certain time, pheromone evaporation may erase the path. Therefore, the parameter ρ controls that paths are not found randomly, but are selected if they are consistently better than others.

The above scheme is applied for values of s ranging from 1 to a certain maximum. According to the factor sparsity principle, the maximum value of s should be kept as small as possible; in practice we found it sensible to set it at 10 for a 12-run PB design, and then let a forward selection procedure described below to select minimal models having satisfactory statistical indicators.

2.3. The Monte Carlo approach

The Monte Carlo approach implies that the above ACO calculations are repeated a number of times for each of the analyzed cases. A normalized histogram is then built on the average, over the Monte Carlo cycles, of the absolute value of the coefficient terms, weighted inversely with the RMSE model, in order to give comparatively more importance to better models. Then a forward selection procedure is implemented, which involves creating gradually augmented models, adding successive terms appearing in the histogram, in decreasing order of intensity, until a certain minimum level is attained. Each of these models is checked for leading to full-rank model matrices, and for compliance to the heredity principle, otherwise they are discarded.

The standard deviations of the least-squares coefficients solving for the candidate augmented models are estimated in order to judge their statistical significance. Specifically, the confidence interval for each coefficient is computed as [10]:

$$\mathrm{CI}(b_n) = t_{\nu,\alpha/2}\, s(b_n) = t_{\nu,\alpha/2}\,\mathrm{RMSE}\sqrt{[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]_{nn}} \tag{9}$$

where t is a Student coefficient for ν degrees of freedom (number of experiments minus number of estimated coefficients) and a $(1-\alpha)\times 100\%$ confidence level, and $s(b_n)$ is the standard error in the estimation of b_n. A term is not considered significant when $b_n \pm \mathrm{CI}(b_n)$ includes the value of zero.

The RMSEs for all the valid forward-selected models are successively compared using Van der Voet's randomization test [30]. Significance is established if the probability for the applied test is smaller than 0.05.

The required parameters for running this ACO version for variable selection are (suggestions in brackets): (1) ρ parameter (0.65), (2) number of ants (20), (3) maximum time steps (50), (4) number of Monte Carlo cycles or repeated calculations for histogram building (10), (5) minimum level to include terms in forward selection (0.1). These parameters were estimated as optimum for the present examples on a trial and error basis. They can be employed as starting values for future work, with suitable adaptations on account of the particular characteristics of each system.

2.4. Software

All routines for performing the presently described calculations were written in MATLAB 7.10 [31]. They are available from the authors on request, including data files containing the model matrices and responses herein studied.

3. Data

3.1. Simulated system

A simulated model, including nine main factors, has been studied. Notice that the total number of main and two-factor associations in this system is 45, i.e., larger than the number of the PB screening experiments, and thus the system cannot be directly solved because the full **X** matrix of size $12 \times N$ (N=45) is rank deficient. For the simulated model, the following expression was employed to compute the responses:

$$y_1 = b_1x_1 + b_7x_7 + b_{13}x_1x_3 + b_{17}x_1x_7 \tag{10}$$

The b coefficients were selected so that factors 1 and 7 were significant, as well as an association between a significant (1) and a non-significant main factor (3), and another association between two significant main factors (1 and 7). Factors 2, 3, 4, 5, 6, 8 and 9 were not considered significant, as all associations

except 1–3 and 1–7. The values of b_1, b_7, b_{13} and b_{17} were 1, −1, 2 and 1, respectively, leading to the following normalized values: 0.38, −0.38, 0.76 and 0.38, respectively.

With the vector of coefficients **b** normalized, a PB design of 12 runs was created at two factor levels (−1 and 1). Once the twelve responses were calculated, they were scaled in the range 0−1 and random Gaussian noise was added with zero mean and 0.05 standard deviation. The model was evaluated in order to assess the terms recovered as statistically significant.

3.2. Experimental systems

Two experimental systems taken from the analytical literature have been analyzed with the present ACO approach. In both cases the influence of eight different factors were studied through a PB design of 12 runs, meaning that the total number of main and two-factor interacting terms (36 terms) greatly outnumber the available experiments. In these examples, the original works resorted to classical PB analysis, with no particular emphasis on interacting terms. After reinvestigation, interactions were indeed found, leading to the detection of previously unseen relevant factors [13].

In the experimental example 1, the potential influence of eight different factors was analyzed on several responses associated to the high-performance liquid chromatographic (HPLC) determination of pharmaceutical mixtures [32]. For one of the responses, namely the recovery of the analyte ridogrel, no significant relationship between the latter and any of the factors was originally detected. However, Bayesian–Gibbs analysis identified some of the factors as being relevant, including a two-factor interaction [13]. The factors were: the mobile phase flow rate, the buffer pH, the column temperature, the type of column, the % of one organic solvent in the mobile phase at the gradient start, the % of another organic solvent at the gradient end, the buffer concentration and the detection wavelength.

In the experimental example 2, a methodology for the chemical characterization of white grapes was developed by simultaneously determining phenolic compounds and organic acids [33]. Eight variables were studied as to their relevance regarding the analytical response (chromatographic peak area/amount of sample); in the present report focus is directed towards the analyte kaempferol-3-*O*-rutinoside. The traditional approach identified a single factor as important, although the confidence for this selection was rather low, given the poor fit of the standard PB evaluation. Significant interactions were then found by Bayesian–Gibbs analysis, pointing to additional effects as significant [13]. The experimental factors studied were: the extractive solvent, the extraction volume, the extraction time, the temperature, the extraction type (ultrasonic energy or stirring), the sorbent type (end capped or non end capped C18) and the elution volume.

4. Results and discussion

4.1. Simulated data

It is already known that standard PB evaluation may lead to increasingly incorrect results regarding the significance of the main factors as the interaction increases. In particular, for the presently analyzed simulated system, no main terms are found to be significant using classical PB analysis, because all associated probabilities with the main terms were larger than 0.05. This is in clear contrast to the building of the simulated system, where four significant terms (two main and two interactions) were definitely included in the model.

Table 1 collects the ACO results for the simulated system. The Monte Carlo histogram shows the most intense peaks corresponding to the four terms included in the simulation, in good agreement with expectations (Fig. 2). Forward selection confirms the applied strategy for building the final model: if only the most

Table 1
Model matrix, responses and ACO forward selection results for the simulated system.

Experiment	Factors									Response
	1	2	3	4	5	6	7	8	9	
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0.97
2	−1	1	1	−1	1	−1	−1	−1	1	0.61
3	−1	−1	1	1	1	−1	1	1	−1	−0.07
4	1	1	1	−1	1	1	−1	1	−1	1.01
5	1	1	−1	1	−1	−1	−1	1	1	0.55
6	1	−1	1	1	−1	1	−1	−1	−1	1.00
7	−1	−1	−1	1	1	1	−1	1	1	1.00
8	1	−1	−1	−1	1	1	1	−1	1	0.46
9	1	−1	1	−1	−1	−1	1	1	1	1.01
10	1	1	−1	1	1	−1	1	−1	−1	0.43
11	−1	1	1	1	−1	1	1	−1	1	0.04
12	−1	1	−1	−1	−1	1	1	1	−1	0.58

Model	Terms included	Normalized coefficients	Comments[a]
Forward model selection			
1	Empty design	None	1–3 alone does not meet the heredity principle
2	7	b_7=−1.00	1–3 does not meet the heredity principle
			RMSE=0.29
			r^2=0.3784
3	1, 7 and 1–3	b_1=0.36	RMSE=0.13
		b_7=−0.46	r^2=0.8926
		b_{13}=0.82	p value for 2 and 3 < 0.05
4	1, 7, 1-3 and 1–7	b_1=0.33	RMSE=0.05
		b_7=−0.42	r^2=0.9870
		b_{13}=0.77	Selected model
		b_{17}=0.34	p value for 3 and 4 < 0.05

[a] The p value corresponds to Van der Voet's test for comparing RMSEs (see text).

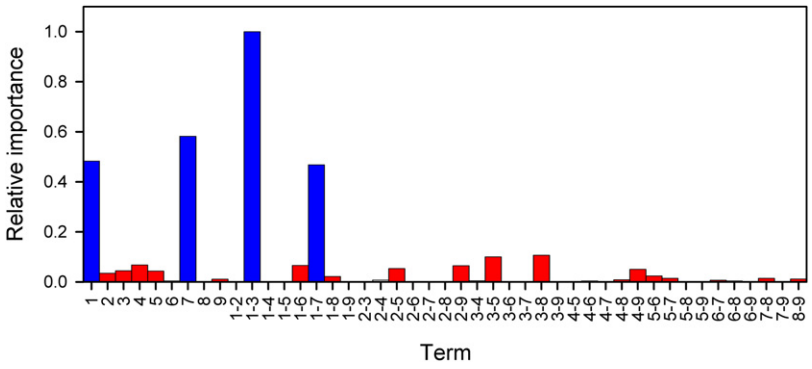

Fig. 2. (A) Histogram of Monte Carlo coefficient values for the simulated system 1. (B) Analogous results for the simulated system 2. Blue bars correspond to the terms included in the final selected model. (For interpretation of the references to color in this figure legend, the reader is reffered to the web version of this article.)

Table 2
Model matrix, responses and ACO forward selection results for the HPLC recovery experiment.

Experiment	Factors								Response
	1	2	3	4	5	6	7	8	
1	1	1	−1	1	1	1	−1	−1	101.6
2	1	1	1	−1	−1	1	1	1	101.7
3	1	−1	1	−1	1	−1	−1	1	101.6
4	1	−1	−1	−1	1	−1	1	1	101.9
5	1	−1	−1	1	−1	1	1	−1	101.8
6	−1	1	1	−1	1	−1	1	−1	101.1
7	−1	1	−1	−1	1	1	−1	1	101.1
8	−1	−1	1	−1	1	1	1	−1	101.6
9	−1	−1	1	1	−1	1	−1	1	98.4
10	−1	1	−1	1	−1	−1	1	1	99.7
11	1	1	1	1	−1	−1	−1	−1	99.7
12	−1	−1	−1	−1	−1	−1	−1	−1	102.3
Model	Terms included			Normalized coefficients				Comments[a,b]	
Forward model selection									
1	4–5			None				4–5 alone does not meet heredity principle	
2	4–5 and 4			$b_4=-0.58$				RMSE=0.16	
				$b_{45}=0.81$				$r^2=0.7373$	
3	4–5, 4 and 5			$b_4=-0.53$				RMSE=0.10	
				$b_5=0.42$				$r^2=0.8949$	
				$b_{45}=0.74$				Selected model	
								p value for 2 and 3 < 0.05	

[a] RMSE values are given in arbitrary units because responses were mean centered and scaled before applying ACO selection. See text for values in the original response scale.

[b] The p value corresponds to Van der Voet's test for comparing RMSEs (see text).

intense term is considered (the interaction 1−3, see Fig. 2), the model is empty, because this term alone does not meet the heredity principle (Table 1). The gradually augmented models display the final coefficients shown in Table 1, with RMSE values decreasing and r^2 improving on increasing the model complexity. The comparison of models shows that a definitely smaller RMSE is reached at model No. 4. No models beyond the latter one were found, which is thus selected by the present analysis (Table 1). The normalized coefficients nicely agree with those employed for simulation, while the final RMSE value is compatible with the noise level introduced in the responses (0.05 units).

Therefore, a correct solution was found in the simulated case using ACO analysis, including the finding of the significant interacting term and main factors. Agreement was found in the absolute values and signs of the significant coefficient terms (see Table 1). It is noteworthy that the model described by Eq. (2), which hypothetically has 45 different coefficients, was analyzed with only 12 runs.

As already reported [14], it is important to recall that standard PB analysis of this system would produce the wrong impression that the factors had no influence on the response, whereas clear effects of several terms are found by the present methodology.

4.2. Experimental data

4.2.1. HPLC recovery experiment

Standard PB evaluation of this experimental system indicates no statistically significant terms at $p<0.1$ level [32]. Only factor 4 is marginally important ($p=0.17$), leading to a rather poor correlation coefficient of 0.2519. This suggests that the conclusions drawn from simple PB analysis may be dubious.

Using the presently discussed ACO methodology, three terms were found to be relevant: 4, 5 and their mutual interaction 4–5 (Table 2). This is supported by the Monte Carlo histogram shown in Fig. 3(A), which leads to the forward selection results displayed in Table 2. The first forward selected model only includes the interaction 4–5, and does not meet the heredity principle, because no main terms appear to justify the interaction (Table 2). The subsequent model No. 2, now also including term 4, meets all requirements but provides a rather poor fit. In fact, the best model No. 3 provides a

clear improvement in r^2 with respect to model No. 2 and a reasonably low RMSE value. This result is in good agreement with Ref. [13] in terms of the selection of terms for the finally selected model:

$$y = 101.0 - 0.6x_4 + 0.4x_5 + 0.8x_4x_5 \quad (11)$$

which, in original units, yields an RMSE = 0.44%. It should be noticed that classical PB evaluation would indicate absence of influencing effects on the response, leading to misinterpretation of the experimental data, whereas further analysis using the present methodology strongly suggests the presence of two significant factors.

Interaction between experimental factors means, in general, that the effect of one factor depends on the level of a second factor. The mathematical model leads to the result that the effect of changing column type (i.e., different manufacturers) depends

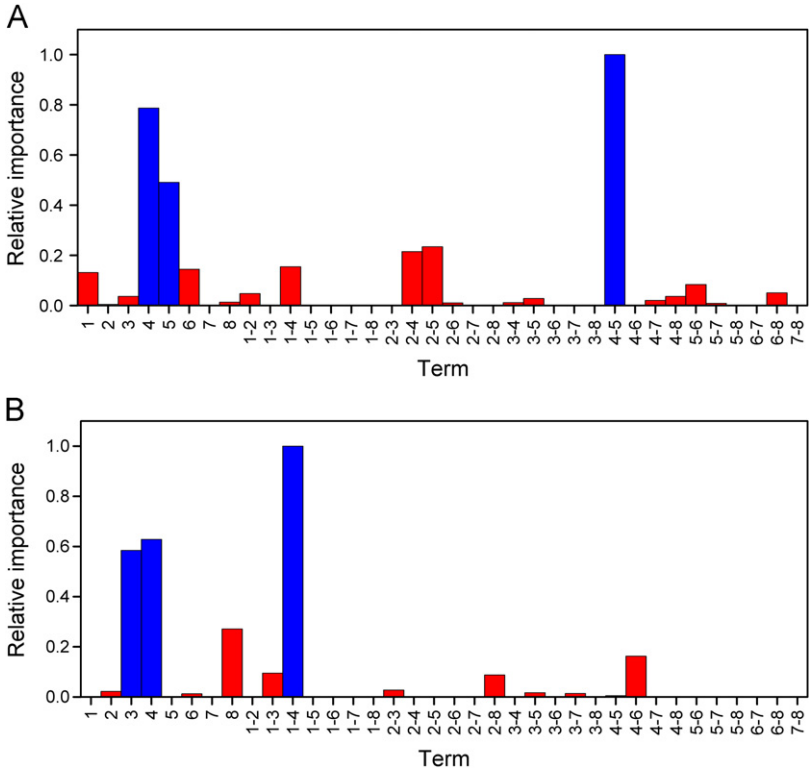

Fig. 3. (A) Histogram of Monte Carlo coefficient values for the HPLC recovery experimental system 1. (B) Analogous results for the extraction/purification experimental system 2. Blue bars correspond to the terms included in the final selected model; red bars to those not included. (For interpretation of the references to color in this figure legend, the reader is reffered to the web version of this article.)

Table 3
Model matrix, responses and ACO forward selection results for the extraction/purification experiment.

Experiment	Factor								Response
	1	2	3	4	5	6	7	8	
1	1	−1	1	−1	−1	−1	1	1	6.98
2	1	1	−1	1	−1	−1	−1	1	5.31
3	−1	1	1	−1	1	−1	−1	−1	9.67
4	1	−1		1	−1	1	−1	−1	6.45
5	1	1	−1	1	1	−1	1	−1	5.23
6	1	1	1	−1	1	1	−1	1	5.34
7	−1	1	1	1	−1	1	1	−1	4.03
8	−1	−1	1	1	1	−1	1	1	3.76
9	−1	−1	−1	1	1	1	−1	1	2.10
10	1	−1	−1	−1	1	1	1	−1	2.65
11	−1	1	−1	−1	−1	1	1	1	7.40
12	−1	−1	−1	−1	−1	−1	−1	−1	7.14

Model	Terms included	Normalized coefficients	Comments[a]
Forward model selection			
1	1–4	None	1–4 alone does not meet heredity principle
2	1–4 and 3	None	1–4 and 3 do not meet heredity principle
3	1–4, 3 and 4	b_3 = 0.48	RMSE = 0.
		b_4 = −0.49	r^2 = 0.9271
	086	b_{14} = 0.75	Selected model

[a] RMSE values are given in arbitrary units because responses were mean centered and scaled before applying ACO selection. See text for values in the original response scale.

on the % of one organic solvent in the gradient. The physical basis of this factor interaction may be found in chemical/physical interactions between the mobile phase and the column material. Whether this is a real effect or not would certainly merit further investigation.

4.2.2. Extraction/purification experiment

In this experimental system there are no significant terms if standard PB analysis is applied. Only factor 4 appeared to be marginally significant ($p=0.16$), but the poor correlation coefficient (0.2435) might indicate the presence of interactions.

The present ACO analysis helps to uncover main and interacting effects in this system, according to the histogram shown in Fig. 3(B), which leads to the forward selection results presented in Table 3. According to this Table, the first two forward selected models do not meet the heredity principle. The subsequent model No. 3, however, achieves a low RMSE value, by considering factors 3 and 4 as well as the interaction between factors 1 and 4. With no further models beyond model No. 3, the present ACO results are entirely consistent and supportive of the model proposed in [13]. The selected model leads to the following expression for explaining the response:

$$y = 5.5 + 1.1x_3 - 1.0x_4 + 1.7x_1x_4 \quad (12)$$

leading to a final RMSE of 0.69 units (12% with respect to the mean response value).

Notice that the specific factors identified as being important were the temperature (x_4), the extraction time (x_3) and the extracting solvent (x_1). As with the first experimental example, the finding of factor interactions calls for the search of physical interpretations. In the present example, the existence of an interaction between extracting solvent and temperature is perfectly understandable on a chemical basis. Of course this must be proved experimentally with separate experiments.

5. Conclusions

A study about the possibility of using the Plackett–Burman experimental design to build models that include associated terms has been carried out. In this regard, ant colony optimization provides an efficient tool for estimating the significant terms, including the values of the model coefficients directly. From this point of view, the Plackett–Burman design could not only be considered as a screening design, but as a design which allows one to build models with a great economy of runs, provided it is complemented with the appropriate approaches for uncovering factor interactions.

Acknowledgments

The University of Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project No. PIP 1950) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT 2010-0084) are gratefully acknowledged for financial support.

References

[1] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley & Sons, New York, 1978.
[2] R. Leardi, Anal. Chim. Acta 652 (2009) 161.
[3] E. Nemutlua, S. Kira, D. Katlanb, M.S. Beksac, Talanta 80 (2009) 117.
[4] L. Vera-Candioti, A.C. Olivieri, H.C. Goicoechea, Anal. Chim. Acta 595 (2007) 310.
[5] J.A. Arancibia, G.M. Escandar, Analyst 126 (2001) 917.
[6] C.R.T. Tarley, G. Silveira, W.N. Lopes dos Santos, G. Domingues Matos, E.G. Paranhos da Silva, M. Almeida Bezerra, M. Miró, S.L.Costa Ferreira, Microchem. J. 92 (2009) 58.
[7] R. Cela, Screening strategies, in comprehensive, in: S.D. Chemometrics, R. Brown, B. Tauler (Eds.), Walczak, Vol. 1, Elsevier, Amsterdam, 2009, p. 251.
[8] J.C.F. Wu, M. Hamada, Experiments. Planning, Analysis, and Parameter Designs Optimization, Wiley-Interscience, New York, 2000 Chapter 8.
[9] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997 Chapters 21–24.
[10] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, John Wiley & Sons, New York, 1978.
[11] R.L. Plackett, J.P. Burman, Biometrika 33 (1946) 305.
[12] J. Lawson, Comput. Stat. Data Anal. 39 (2002) 227.
[13] F.K.H. Phoa, W.K. Wong, H. Xu, J. Chemometrics 23 (2009) 545.
[14] J.F. Magallanes, A.C. Olivieri, Chemom. Intell. Lab. Syst. 102 (2010) 8.
[15] M. Hamada, C.F.J. Wu, J. Qual. Technol. 24 (1992) 130.
[16] H. Chipman, M. Hamada, C.F.J. Wu, Technometrics 39 (1997) 372.
[17] P.J. Brown, M. Vannucci, T. Fearn, J. R. Stat. Soc. B 60 (1998) 627.
[18] D.K.J. Lin, N.R. Draper, Technometrics 34 (1992) 423.
[19] C.S. Cheng, Ann. Stat. 23 (1995) 1223.
[20] C.S. Cheng, Biometrika 85 (1998) 491.
[21] F.K.H. Phoa, Y.H. Pan, H.J. Xu, Stat. Plan. Infer. 139 (2009) 2362.
[22] R.J. Leardi, J. Chemometrics 14 (2000) 643.
[23] L. Xu, J.H. Jiang, H.L. Wu, G.L. Shen, R.Q. Yu, Chemom. Intell. Lab. Syst. 85 (2007) 140.
[24] M. Dorigo, T. Stützle, Ant Colony Optimization, The MIT Press, Cambridge, MA, USA, 2004.
[25] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Anal. Chim. Acta 646 (2009) 39.
[26] M. Goodarzi, M.P. Freitas, R. Jensen, Chemom. Intell. Lab. Syst. 98 (2009) 123.
[27] M. Goodarzi, M.P. Freitas, R. Jensen, J. Chem. Inf. Model. 49 (2009) 824.
[28] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, J. Chemometrics 20 (2006) 146.
[29] F. Allegrini, A.C. Olivieri, Anal. Chim. Acta 699 (2011) 18.
[30] H. van der Voet, Chemom. Intell. Lab. Syst. 25 (1994) 313.
[31] MATLAB 7.10, The MathWorks Inc., Natick, MA, 2010.
[32] Y. Vander Heyden, M. Jimidar, E. Hund, N. Niemeijer, R. Peeters, J. Smeyers-Verbeke, D.L. Massart, J. Hoogmartens, J. Chromatography A 845 (1999) 145.
[33] M.S. Dopico-Garcia, P. Valentao, L. Guerra, P.B. Andrade, R.M. Seabra, Anal. Chim. Acta 583 (2007) 15.