Analytica Chimica Acta 721 (2012) 44–54

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

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

Fractions of Rechtschaffner matrices as supersaturated designs in screening experiments aimed at evaluating main and two-factor interaction effects

R. Cela[a,*], R. Phan-Tan-Luu[b], M. Claeys-Bruno[c], M. Sergent[c]

[a] Analytical Chemistry Department, Institute of Research and Food Analysis, University of Santiago de Compostela, Spain
[b] LPRAI, 40, boulevard Icard, 13010 Marseille, France
[c] Institut des Sciences Moléculaires de Marseille, AD2EM, Université Paul Cézanne Aix-Marseille III, 13397 Marseille Cedex 20, France

ARTICLE INFO

Article history:
Received 13 October 2011
Received in revised form 20 January 2012
Accepted 31 January 2012
Available online 8 February 2012

Keywords:
Experimental design
Screening designs
Supersaturated designs
Rechtschaffner matrices

ABSTRACT

Optimal fractions of resolution V design matrices proposed by Rechtschaffner in 1967 are developed and applied as supersaturated designs in screening experiments. Rechtschaffner matrices allow evaluation of all main factors and two-factor interactions, which in many real-world studies are of practical significance. However, the number of experimental runs increases rapidly with the number of factors in the matrices, which are therefore impractical for more than 5–6 factors. On the contrary, saturated fractions based on Hadamard matrices, which are commonly applied in screening studies, cannot evaluate the interaction effects. Here, a procedure for selecting the optimum fractions of Rechtschaffner matrices is presented and provides supersaturated matrices that are well adapted to a variety of problems, thus allowing the development of screening studies with a relatively small number of experiments. The procedures developed to derive the size-reduced matrices and to evaluate the active factors are discussed and compared in terms of efficiency and reliability, by means of simulation studies and application to a real problem. These fractions are the first supersaturated design matrices capable of estimating interaction effects. Additionally, one important advantage of these supersaturated matrices is that they enable development of follow-up procedures in cases of inconclusive results, by enlarging the matrix and eventually resolving the full Rechtschaffner matrix of departure when it is necessary to evaluate the active factors and their interactions.

1. Introduction

Screening experiments are applied when the objective is to identify a few significant factors from among a large number of potentially active factors (the so-called sparsity effect [1]) using as few runs as possible. In general, the hypothesis is that the main active factors can be identified by screening experiments and that further experiments will evaluate the exact influence of these factors as well as their possible interactions. Thus, saturated experiments (resolution III designs requiring only as many runs as there are parameters in the model to be estimated) that focus on main effects are frequently applied for screening purposes. In saturated experiments, we assume an additive model for the process under study. In other words, no interaction factors are significant. Moreover, since saturated experiments do not provide an estimate of experimental error, we assume that a prior estimate of this experimental error is available. Otherwise, no statistical decisions can be made.

Resolution III experiments may be regular or non-regular. In regular resolution III experiments, two-factor interaction effects are orthogonal or fully confounded with main effects, so that the only way of estimating the interaction effects is to increase the resolution by augmenting the design matrix. On the other hand, non-regular resolution III experiments have complex confounding patterns with two-factor interaction effects partially confounded by many main effects. Thus, hidden projective properties [2,3] of these non-regular designs may be used to estimate some two-factor interactions, even though the design does not have the appropriate resolution. Sparsity effects must therefore be admitted, and the effects of some heredity rules [4,5] also need to be applied. These heredity rules imply that interaction effects can only be significant if at least one of the main factors involved is active (weak heredity) or when both factors involved are active (strong heredity). Application of both the sparsity and the heredity rules make it possible to reduce the number of runs in the experiment to amenable limits because if only a few main factors are active and only the interactions between these active factors need to be screened, then there

* Corresponding author.
E-mail address: rafael.cela@usc.es (R. Cela).

0003-2670/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2012.01.060

will be far fewer parameters in the model than in the full factorial model.

As regards the limit, if we can reasonably assume that the ratio of active factors to total factors falls below 10–15%, supersaturated design matrices may be used in screening experiments. Supersaturated designs are experimental designs in which there are fewer run (rows) than factors (columns). This means that main factors are aliased, and therefore cannot be accurately estimated because the conventional estimates may be correlated. Thus, the practical application of supersaturated designs is only recommended if strong sparsity effects are anticipated. It is apparent that no attempt at evaluating interaction effects can be made in supersaturated designs.

Supersaturated designs were initially proposed by Satterthwaite [6] in 1951, but their development and application only became important in the 1990s [7–10]. In general, much work has been done regarding the construction of several new types of supersaturated designs, and much less concerning the practical resolution of these designs [11] and their application to real-life problems. Because in real world situations we cannot be sure about the fulfillment of the conditions required for the application of supersaturated designs on starting experimentation, one important aspect is the possibility of follow-up experiments if the initial design does not appear to provide conclusive results. Most of the supersaturated matrices proposed in the last two decades do not allow this possibility. However, supersaturated matrices constructed by branching Plackett-Burman designs [7,9] will allow follow-up processes when necessary, so that the effort expended is always profitable.

In cases where we suspect the existence of two-factor interactions, which in real-life problems occur frequently, resolution V experiments are required to estimate both the main effects and the effects of interactions. In such cases, Rechtschaffner designs [12] (resolution V saturated and balanced matrices), requiring only a few experiments, are of interest.

Rechtschaffner [12] proposed eight classes of designs that can be constructed with three design generators: generator I is either the treatment combination (1, . . ., 1), in which each factor is at its highest level, or (−1, . . ., −1), in which each factor is at its lowest level (one row in the design matrix). Generator II are treatment combinations in which one of the n factors is at its highest (lowest alternatively) level while the others are at their lowest (highest) level (n rows in the matrix). Generator III are treatment combinations in which two of the n factors are at their highest (lowest alternatively) level while the other are at their lowest (highest) level ($n(n-1)/2$ rows in the matrix). Thus, the total number of runs is $1+n+n(n-1)/2$, which equals the number of parameters in the general model:

$$y = b_0 + \sum_{1}^{n} b_i x_i + \sum_{1}^{n-1} \sum_{j=i+1}^{n} b_{ij} x_i x_j$$

The statistical properties of Rechtschaffner designs have been studied by Qu [13], in comparison with other resolution V designs and also by Qu and Wu [14], and applied by several authors in simulated [15] and real-life problems [16–18].

Although Rechtschaffner matrices require the minimum number of runs in resolution V experiments, the experimental effort increases considerably when a large number of factors is considered. For example, considering only designs at two levels, the Rechtschaffner matrix for 5 factors involves 16 runs, but the matrix for 9 factors requires 46 runs, and for 20 factors 211 runs are necessary. These numbers go beyond the possibilities of real-life experimental studies. Thus, we propose here the use of fractions of Rechtschaffner matrices as supersaturated matrices in screening studies where the interest is not only in the main effects but also the two-factor interactions. The use of a fraction of a resolution V design has the intrinsic advantage of allowing follow-up processes when necessary; estimation of main and two-factor effects, with a drastically reduced experimental effort, is also possible under the sparsity and heredity principles.

In the following, we describe the procedures used to obtain optimal fractions of Rechtschaffner matrices (Section 2) and two proposals for evaluating the effects in the supersaturated matrices obtained (Section 3), and also test these proposals with simulated and real-life data (Sections 4 and 5). Finally, some conclusions are reached about the practical applications of our proposal.

2. Construction of optimal fractions of Rechtschaffner matrices

In light of the generators used to construct the Rechtschaffner matrices, it is evident that a simple branching procedure to produce half or approximately half-fractions cannot be used. Moreover, using signs to make branching would produce fractions that mainly correspond to rows produced by generator I and II, or by generator III. Thus, in theory, these fractions may perform undesirably as supersaturated fractions. With the aim of using fractions of Rechtschaffner matrices in screening experiments to evaluate main and two-factor effects, a measure of the potential ability of these fractions to provide the needed information is needed. Here, we propose the use of the alias matrix calculated from the fractions.

The alias matrix, defined by Box and Wison [19], can be applied to determine the aliases of main effects with two-factor interactions in saturated designs [20,21]. The alias matrix can be calculated [22] as:

$$A = (X_1'X_1)^{-1}(X_1'X_2)$$

where X_1 is an $n \times n$ matrix of contrast coefficients for the intercept, main effects and estimable two-factor interactions and X_2 is an $n \times u$ matrix of contrast coefficients for the inestimable two-factor and higher interactions, considering a model such as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where β_1 and β_2 are respectively the corresponding $n \times 1$ and $u \times 1$ vectors of regression coefficients of estimable and inestimable effects. Thus, it is assumed that the regression solution of the main effects model from the design matrix is biased by the existence of inestimable effects:

$$y = X_1\beta_1 + e$$

$$\widehat{\beta}_1 = (X_1'X_1)^{-1}X_1'y$$

$$E(\widehat{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2$$

Thus, according to Lawson [22], the alias matrix shows how linear combinations of the inestimable two-factor interactions coefficients bias the regression coefficients for the intercept, the main effects and the estimable two-factor interactions, if any. The rows of the alias matrix correspond to the biased regression coefficients and the columns to the regression coefficients for inestimable two-factor interactions. The values in each row are the multipliers for the linear combination of interaction regression coefficients that bias each estimable regression coefficient.

The alias matrix has been used by Lawson [22] as an alternative way of interpreting the projective properties of non-regular designs. Instead of considering only the interaction effects derived from the main factors identified as active (strong heredity rule), cells in the alias matrix with large values are further considered to

screen possible active two-factor interactions, finally assessed by the heredity principle.

Thus cells with large values correspond to interaction effects that mainly bias the main factor estimations. Following this reasoning, at the time of selecting a fraction of a Rechtschaffner matrix, one must seek fractions with the most uniform alias matrix with small values. This means that bias in the estimation of main effects due to two-factor interactions will be approximately equal to all main factors. Of course, smaller values in the alias matrix will lead to less bias in the main effects estimated. Obviously this second condition will drive the search for full Rechtschaffner matrices because they are of resolution V. The fewer the number of rows in the matrix fraction, the larger the values in the alias matrix will be. After fixing a particular number of rows, a search procedure is required to select the matrix with the most uniformly distributed alias matrix. In this sense we can call these matrices optimal.

Development of the required procedure was aided by an evolutionary algorithm implemented in house [23] using Delphi 2007 for windows programming language. In this application, a population of chromosomes whose elements are the row positions in the original full Rechtschaffner matrix considered, are allowed to evolve under the effects of the usual genetic operators [24] ($k = 2$ tournament selection, simple crossover recombination with $p = 0.7$, uniform mutation ($p = 0.01$) and elitism). The fitness function is a linear combination of measures of alias matrix uniformity. In this case, a combination of the absolute average value, the range and the variance of cell values in the alias matrix corresponding to each individual was used as a fitness function to drive the evolutionary process.

The alias matrices corresponding to (a) a random fraction and (b) the optimal fraction obtained using the above described procedure for a 33 rows fraction of the $2^{11}//67$ Rechtschaffner matrix are graphically compared in Fig. 1. As expected, the overall values in (b) are lower than in (a), and positive and negative values are quite uniform, unlike the randomly constructed fraction. Moreover, in (b) it can be seen that higher values for the alias matrix cells correspond to the partial aliasing with the intercept. In (a) several large values in the alias matrix are also associated with the intercept value, although several equally important values are distributed among the main effects. Thus, the fitness function developed to optimize the construction of these fractions favors aliasing the inestimable two-factor interactions with the intercept, while managing to avoid aliasing the main effects coefficients.

Several optimal fractions of different Rechtschaffner matrices, constructed using the RFRACT application and following the procedure described, have been compiled for practical purposes (Table 1). Other fractions can easily be constructed using the RFRACT application, which is a freeware program that can be obtained from the authors on request.

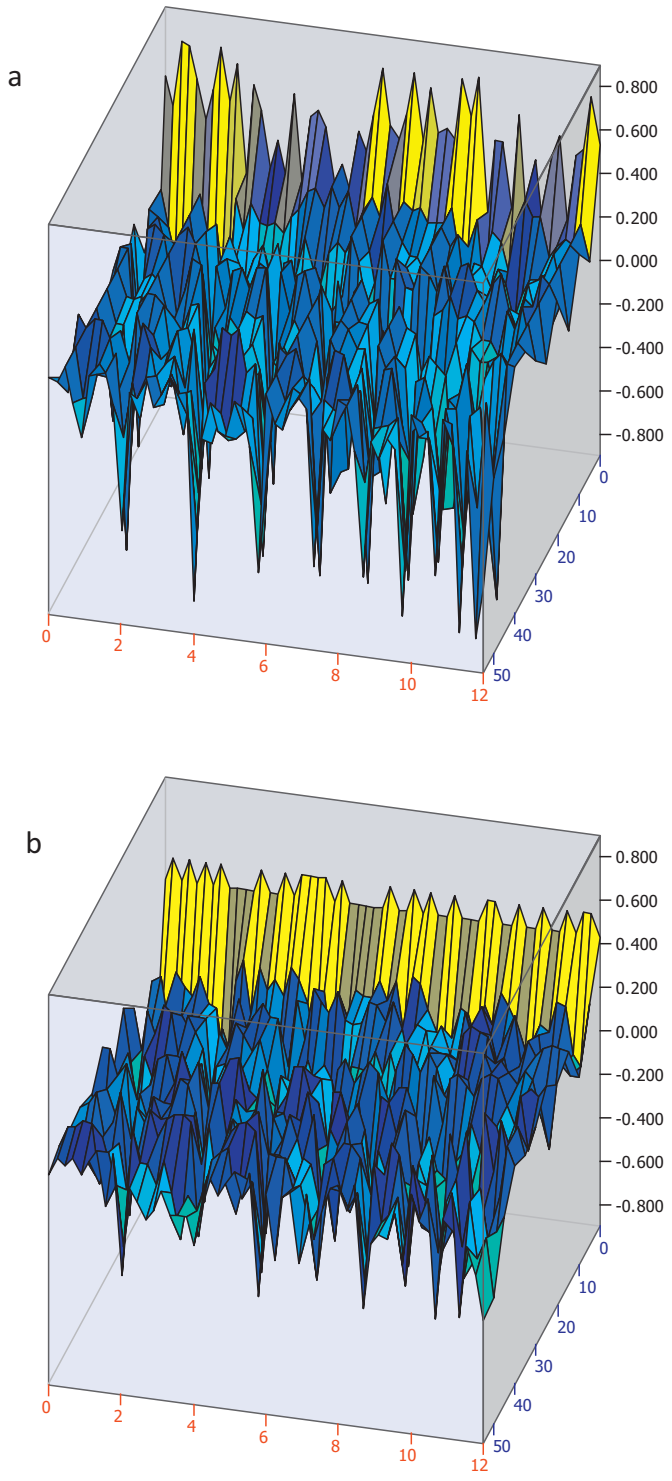

Fig. 1. Graphical representation of alias matrices of 33 row fractions of the $2^{11}//67$ Rechtschaffner matrix: (a) randomly size-reduced matrix, and (b) optimal size-reduced matrix.

3. Evaluating effects in the size-reduced Rechtschaffner matrices

The accurate evaluation of effects in supersaturated matrices has not yet been fully resolved. At the beginning of 1990s, forward selection in regression analysis [7,25], the stepwise regression procedure [8], all subsets regression [11,26] and Bayesian variable selection procedures [27] were proposed with this aim. Forward and stepwise regression have been shown to fail rapidly if the number of factors is larger than the number of runs, or if there is excessive collinearity. However, as we will see later, in sequential approaches based on projective properties, stepwise regression can perform quite satisfactorily. Although all subsets regression is one the most accurate procedures for analyzing supersaturated matrices (because in theory all possible models can be evaluated), this method has been rejected by several authors because of the enormous computational costs associated with the combinatorial nature of evaluating each potential subset.

Here, two different approaches have been applied to evaluate effects in the supersaturated matrices obtained when fractions of Rechtschaffner matrices are constructed. One of the procedures uses all subsets regression assisted by evolutionary algorithms, whereas the other combines biased initial regression guesses (e.g.

Table 1
Some optimal fractions of Rechtschaffner matrices produced by following the proposed evolutionary search procedure.

Rechtschaffner matrix of departure		Optimal supersaturated fraction	
Factors	Runs	Runs	Rows taken from the matrix of departure
5	16	7	2,4,6,7,8,11,13
		8	1,2,4,6,8,11,12,14
		9	1,2,6,7,8,11,12,14,16
		10	2,4,6,8,9,11,12,13,14,16
		11	1,2,4,6,7,8,9,11,13,14,16
6	22	10	3,4,5,6,8,9,10,11,19,22
		11	2,3,4,5,6,8,9,10,11,15,20
		12	2,3,4,5,6,7,10,11,13,16,19,20
7	29	14	3,4,5,6,7,8,16,18,19,21,22,24,26,28
		15	2,3,4,5,6,7,8,9,13,15,17,23,24,25,29
		16	2,3,4,5,6,7,8,9,13,15,17,23,24,25,26,29
8	37	18	2,3,4,5,6,7,8,9,11,14,15,18,19,20,23,31,33,37
		20	2,3,4,5,6,7,8,9,10,11,12,21,22,24,25,28,31,33,35,36
		22	2,3,4,5,6,7,8,9,10,12,16,19,20,21,22,23,25,26,28,34,36,37
9	46	20	2,3,4,5,6,7,8,9,10,13,14,21,25,27,28,35,36,41,42,45
		23	2,3,4,5,6,7,8,9,10,13,14,18,22,24,25,27,28,29,33,34,39,44,45
		26	2,3,4,5,6,7,8,9,10,11,15,17,18,20,21,23,26,28,29,30,35,37,38,39,45,46
10	56	22	2,3,4,5,6,7,8,9,10,11,19,26,27,34,40,43,44,49,52,53,54,56
		28	2,3,4,5,6,7,8,9,10,11,12,13,19,20,24,25,26,29,30,36,38,39,42,45,48,52,55,56
		32	1,2,3,4,5,6,7,8,9,10,11,12,13,14,18,20,23,24,25,29,30,32,33,37,40,42,46,48,52,53,54,56
11	67	24	2,3,4,5,6,7,8,10,11,12,18,20,21,25,26,27,35,39,41,44,45,46,52,56
		33	2,3,4,5,6,7,8,9,10,11,12,14,16,18,20,25,28,30,31,32,33,35,41,44,46,49,53,54,57,60,63,65,66
		40	1,2,3,4,5,6,7,8,9,10,11,12,14,16,17,18,21,22,23,25,26,27,28,32,38,39,40,41,43,46,50,51,54,55,58,59,61,63,65,66
12	79	26	2,3,4,5,6,7,8,9,10,11,12,13,23,24,29,34,38,45,46,52,60,63,69,72,75,78
		36	2,3,4,5,6,7,8,9,10,11,12,13,16,17,21,22,28,29,30,33,39,41,42,43,48,50,51,52,53,56,59,61,65,70,78,79
		40	2,3,4,5,6,7,8,9,10,11,12,13,16,17,18,20,22,26,27,28,29,33,37,38,39,40,42,47,50,51,53,55,62,65,67,73,74,75,76,78
14	106	30	2,3,4,5,6,7,8,9,10,11,12,13,14,15,19,20,28,32,37,40,51,53,70,71,72,74,75,96,105,106
		40	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,25,28,30,35,37,42,43,48,51,56,58,66,68,70,72,73,77,81,83,84,87,90,94,99,105
		53	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,20,21,22,24,28,29,32,33,36,37,41,45,49,50,52,53,54,56,58,62,66,68,69,78,79,80,84,88,89,92,93,94,98,99,102,103,105,106
16	137	36	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,23,28,32,38,50,67,70,81,82,84,93,94,98,106,107,118,123,130,136,137
		48	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,21,26,27,31,33,34,36,44,49,52,55,62,63,65,74,77,79,88,90,93,96,101,106,115,116,118,121,124,129,134,137
		64	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,19,24,27,28,30,32,35,38,39,40,41,45,48,49,50,55,56,60,63,66,67,68,70,76,79,80,83,84,85,86,89,94,96,100,101,107, 108,111,112,116,118,119,120,125,131,134,135,137
20	211	48	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,32,40,46,47,55,66,71,72,76,77,82,83,96,100,116,130,132,139,145,147,161,166,178,181,187,189,204,209
		72	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,24,32,33,37,38,41,42,49,50,58,59,60,66,81,90,92,94,103,105,106,107,109,116,121,124,129,132,134,142, 144,147,149,152,159,160,165,166,168,176,177,181,185,193,194,196,197,198,199,201,202,209

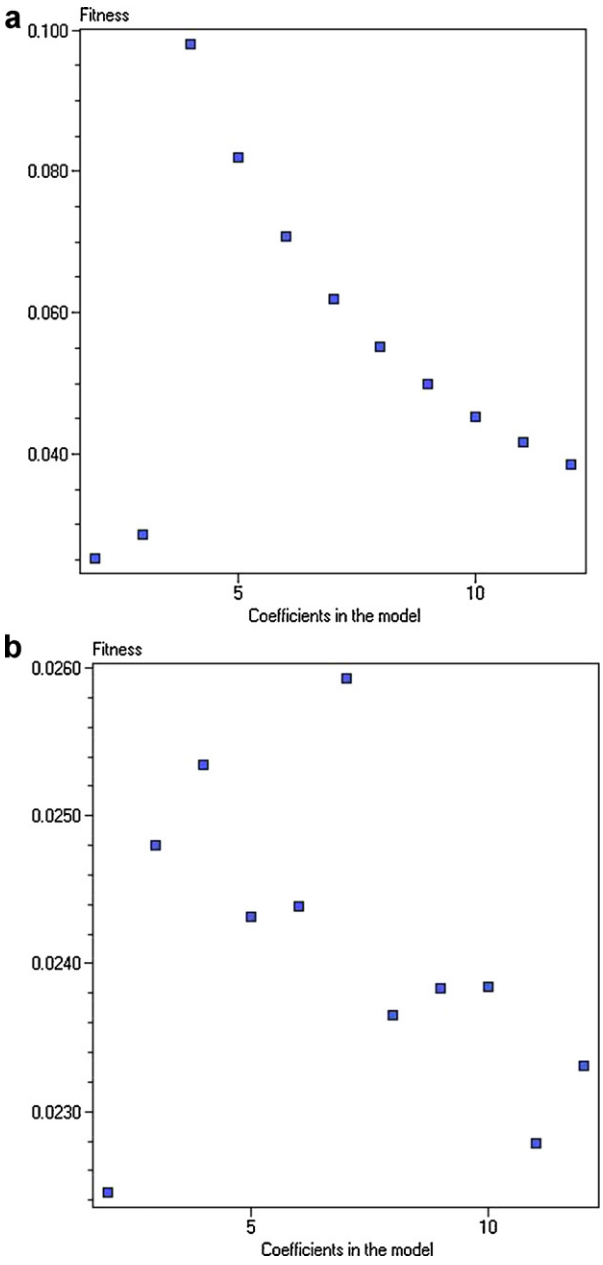

Fig. 2. Graphical tools in the evaluation of effects using all subsets regressions driven by evolutionary algorithms (island plots).

using stepwise regression or ridge regression) with projective properties of the design matrix and a final reduced all-subsets regression process to identify the active effects in the system accurately.

3.1. Procedure A: all subsets regression driven by genetic algorithms

Sudjianto et al. [26,28] were the first authors to propose the use of genetic algorithms to analyze supersaturated design matrices. The basic idea is to represent each subset of variables by a binary string, so that for the full additive model involving k variables:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

would be represented by a string of $k+1$ ones and the void model by a string of $1+k$ zeros. Thus, each subset will be represented by a unique string. These strings constitute the chromosomes of a population of n individuals that would evolve to an optimal point, which will be the best subset of variables for a given problem. To perform the evolutionary process, each individual is assigned a fitness value (a convenient measure of regression quality), which will be used by the selection operator to choose, in any given generation, pairs of individuals that perform better than the population average, to survive, be mated and produce by means of the recombination operator new individuals who by exchange of bits in the parent structure will exploit the best solutions to the problem. In addition, from time to time, small modifications in the chromosomes are introduced by the mutation operator, thus providing a means of exploring all the search space in the pursuit for the optimal solution [29].

In 2002, Cela [30] developed freeware software called SUPERSAT, which is based on the same ideas, using an island-type evolutionary algorithm without transfer of individuals coded in real numbers. Each island in the algorithm corresponds to a fixed complexity subset so that the final output provides the optimal subsets for the different subset sizes [31].

Although a population is maintained on each island, only the best individual is shown at any time, in the so-called island plots shown in Fig. 2. An island plot is a simple representation of the fitness function (vertical scale) of the best individual on each island (horizontal scale, meaning the number of variables in each subset). The island plot sometimes shows a clear break in the fitness values for a certain number of variables in the subset (Fig. 2(a)). This break should be an indication that the optimum subset of this particular size represents the optimal solution for the regression problem. Decreases in fitness values as more and more variables enter the regression subset indicate that noise is now being regressed, so that the increase in complexity of the model does not justify the regression quality. However, an island plot, such as that shown in Fig. 2(b) is sometimes obtained where no clear break appears. This indicates that the number of active factors in the system under study is large (the sparsity principle is violated) or that the noise in the response vector is too high to derive a clear optimal subset. Even in such cases, some important information can be gained by studying so-called screen maps (Fig. 3), which graphically compare the optimum models in each island once the evolutionary process has ended. In screen maps, the patterns of factors may suggest which factors are actually active. Provided the problem under study is well suited to the application of supersaturated designs, the hypothesis in SUPERSAT is that active factors will tend to be retained in the optimum subsets in each island, so that a pattern such as that shown in Fig. 3(a) should appear. Here we see that some factors are almost systematically included in most optimal subsets. Moreover, the sequence of factors entering the optimal models when the model complexity grows, and the estimated values of coefficients for factors (represented by colors according the scale), may also give an indication the activity and significance of factors if previous estimate of experimental error is available. A contrasting situation in which we see that the optimal subset for each island mainly involved different factors and no permanent pattern appears (Fig. 3(b)). This suggests that the number of active factors is large or, more frequently, that the response vector is unable to show the factor activity clearly. These graphical tools are used to evaluate the number of active factors and to give an initial estimation of the corresponding coefficients for the optimal regression model finally adopted.

SUPERSAT can be used to evaluate main effects and also to evaluate regression models including main and two-factor effects (or even a full quadratic model), so that it can be applied directly to evaluate fractions of Rechtschaffner matrices to analyze the model including interactions, or can be applied in two stages by studying a main effects model and then a model of main and two-factor interactions.

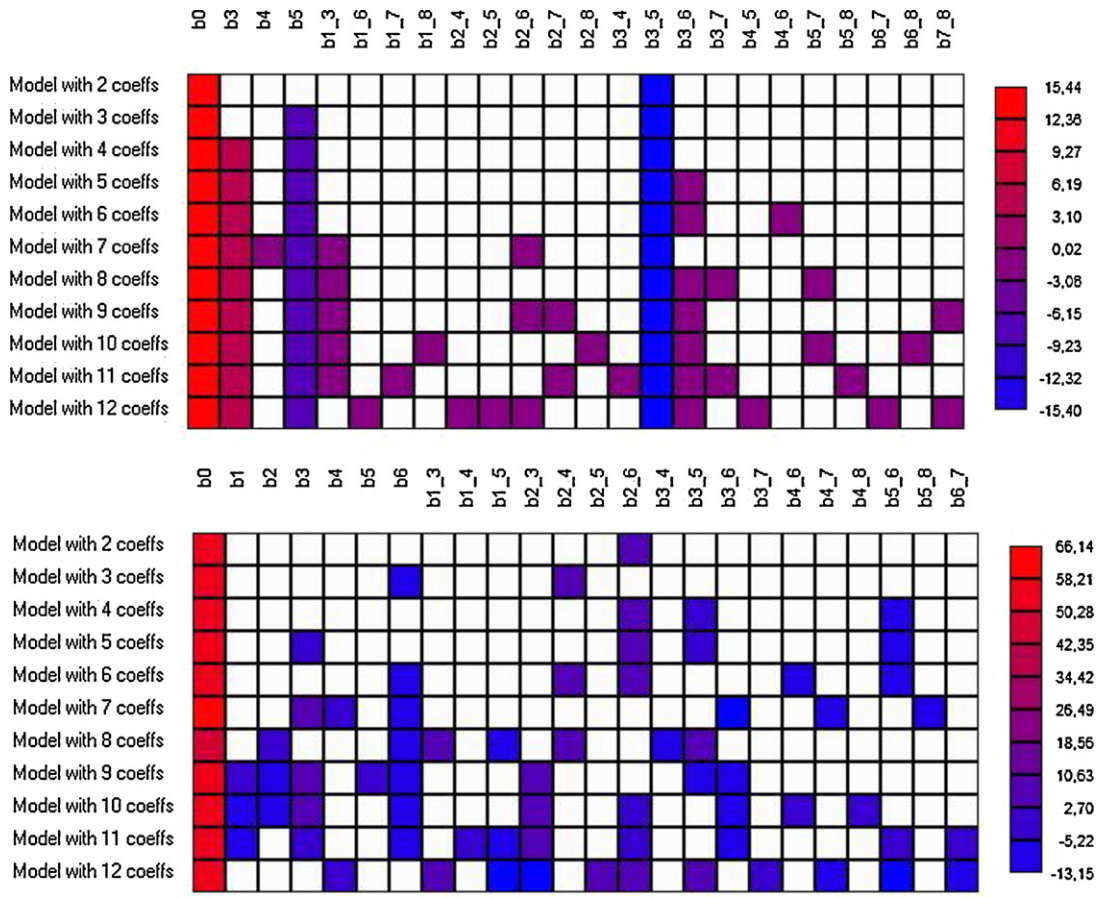

Fig. 3. Graphical tools in the evaluation of effects using all subsets regressions driven by evolutionary algorithms (screen maps).

Evolutionary-driven all subset regression procedures overcome the main drawback of conventional all subsets regression and the matrix evaluation can be carried out within seconds in a modern personal computer. Another frequent criticism of the use of these non-conventional regression procedures is the need for specific software rather than procedures commonly available in standard statistical packages. Since SUPERSAT is freeware, this second drawback is avoided.

3.2. Procedure B: iterative strategy based on stepwise and all-subsets conventional regression procedures and projective approach

There are different problems associated with conventional stepwise and all-subsets regression procedures when solving supersaturated design matrices. As shown in some studies, stepwise regression often fails to detect active factors, but tends to select many inactive factors [11,25]. However, conventional all-subsets regression rapidly becomes unfeasible when the number of factors increases to even moderate numbers, and does not identify model sizes automatically. More elaborate strategies for identifying active factors have recently been developed [31], including the so-called *iterative strategy* [32], which is highly efficient when the number of factors is large. In this strategy, the variables (principal + two-factor interaction effects) are split into several sets (F_1, F_A, F_B, . . ., F_Z). Variables in set F_1 are obtained by calculation of biased coefficients for the original variables (e.g. by stepwise regression), and selection of the apparently most important variables by sorting the coefficients in descending order; k_1 variables are then placed in the F_1 set and the remaining ones are distributed among the other sets (F_A, . . ., F_Z) by placing k_2 variables in each additional sets (thus, the number of sets depends on the original number of variables as well as on the k_1 and k_2 values adopted). The size of K_1 may be defined by computation using Lenth's pseudoscale error [33] for supersaturated matrices with an orthogonal base [7,34,35], or the variables corresponding to the largest $f_1 = |2k/3|$ absolute β_t for matrices without an orthogonal base (e.g. Lin matrices [9]), where k is the number of variables in the matrix; and $k_2 = k_1 - N_{Vmax}$, where N_{Vmax} is the anticipated maximum number of active variables in the study, which should be smaller than the total number of variables considered when supersaturated matrices are used, as previously discussed. For example, in the 2^{11} Rechstchaffner matrix considered in previous discussions, the F_1 set is first constructed with the 12 highest coefficient estimates obtained by forward stepwise regression. If we accept that in this study, 10 is a reasonable maximum number of significant variables, the additional sets will be formed by placing groups of 10 variables in descending order of coefficients. The all-subsets regression with $k' = 10$ is then performed in F_1 and the resultant factors are added to set F_A. Again, all-subsets regression with $k' = 10$ is carried out on the F_A set, taking the best solutions in each case, as before. This procedure is repeated with F_B, F_C and so on until F_Z. Note that the final group of variables is, as expected, very similar to the F_1 group. Then we carried out in this group stepwise regression; calculation of criteria such as BIC (Bayesian Information Criterion) [36], AIC (Akaike's Information Criterion) [37] and mAIC [38] and all-subsets regression.

The final stage consists in projecting the variables detected as significant by the stepwise and all-subsets regression subject to acceptable quality criteria (variance inflation factor <2). The resulting matrix leads to classical treatment (multi-linear regression), and allows quantification of principal effects b_i and two-factor

interaction effects b_{ij}. In most cases, this procedure enables reliable identification of the significant variables.

4. Results and discussion

4.1. Comparison between optimal size-reduced matrix and randomly size-reduced matrices

A series of simulations were initially carried out to confirm the basic hypothesis regarding the use of size-reduced Rechstchaffner matrices in factor screening studies. The first simulations were carried out with three random 20 row size-reduced matrices departing from the 8 factor Rechstchaffner matrix. These matrices included the following rows in the original 2^8//37 design:

Matrix	Rows taken at random from the Rechstchaffner matrix
RS1	2,3,6,10,11,12,15,16,17,21,23,24,25,29,30,31,33,34,35,37
RS2	1,2,4,5,7,9,12,15,16,18,20,22,26,27,29,30,32,33,34,36
RS3	4,6,7,8,11,15,16,17,20,21,22,23,24,28,29,30,31,34,35,37

Simulations were carried out with these matrices, and with two different sets of six randomly chosen active factors, as shown in Table 2. Two different patterns were developed with each set of active factors. In the first pattern, all effects except those of the active factors were assigned a zero value, whereas in the second pattern, small random values ranging from 0 to 1.0 were assigned to the remaining effects by using the Box–Muller algorithm [39]. Furthermore, four simulations were developed using each pattern, by adding increasing levels of noise (standard deviation values from 1 to 4 in the Box–Muller formulae). A total of 48 simulations were produced in this way.

It should be noted that the applied coefficient patterns are clearly higher than the recommended limits in the application of supersaturated matrices because, in both cases, 6 active effects were included in matrices with only 20 rows. Thus, these series of simulations were considered challenging enough to derive the real possibility of using size-reduced Rechtschaffner matrices as saturated matrices in screening studies involving main effects and two-factor interactions.

The results of these simulations are summarized in Table 3. Runs were evaluated by the previously described resolution procedures (direct evolutionary driven all-subsets regression (procedure A) and the combination of stepwise and all-subsets regression with projective approach (procedure B)).

For example, the results obtained for the matrix RS1, response "Pattern 1A" (noise level 1) are reported for the procedure B. Stepwise regression (Fig. 4a), calculation of criteria such as BIC (Bayesian Information Criterion) [36], AIC (Akaike's Information Criterion) [37] and mAIC [38] (Fig. 4b), confirm that the number of significant variables is probably six. All-subsets regression (Fig. 5), with k' values ranging from 2 to 10 (or less) variables clearly shows that coefficients b_3, b_5, b_8, b_{35}, b_{38} and b_{58} are significant. For accurate identification of the values of the coefficients, these variables are projected. The resulting matrix (VIF_{max} = 1.48) leads to the estimations of the model coefficients b_i and b_{ij} using multilinear regression (Fig. 6).

The results shown in Table 3 are expressed as the percentage success in the identification of the real active factors. Values in

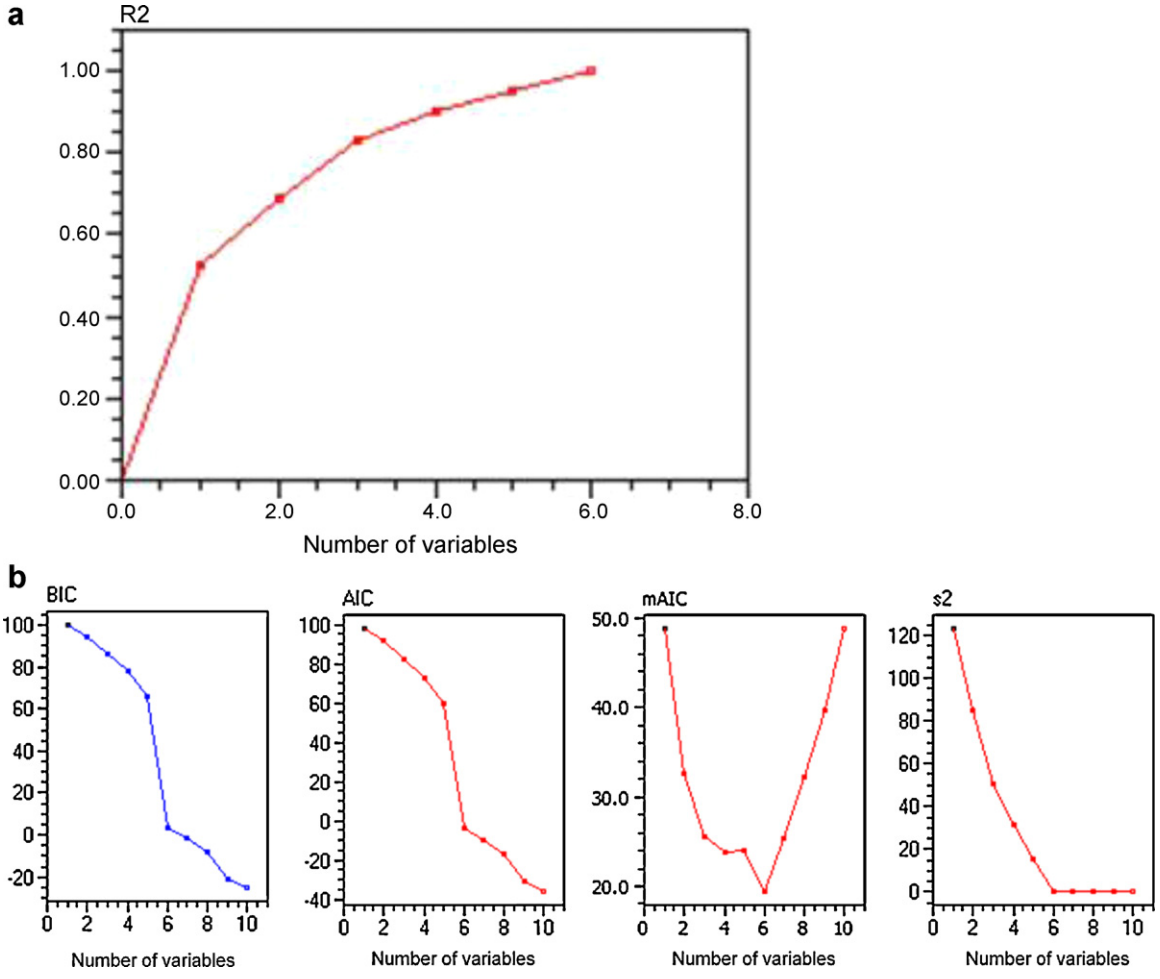

Fig. 4. (a) Stepwise regression to 10 variables and (b) graphs of s^2, AIC, BIC and mAIC.

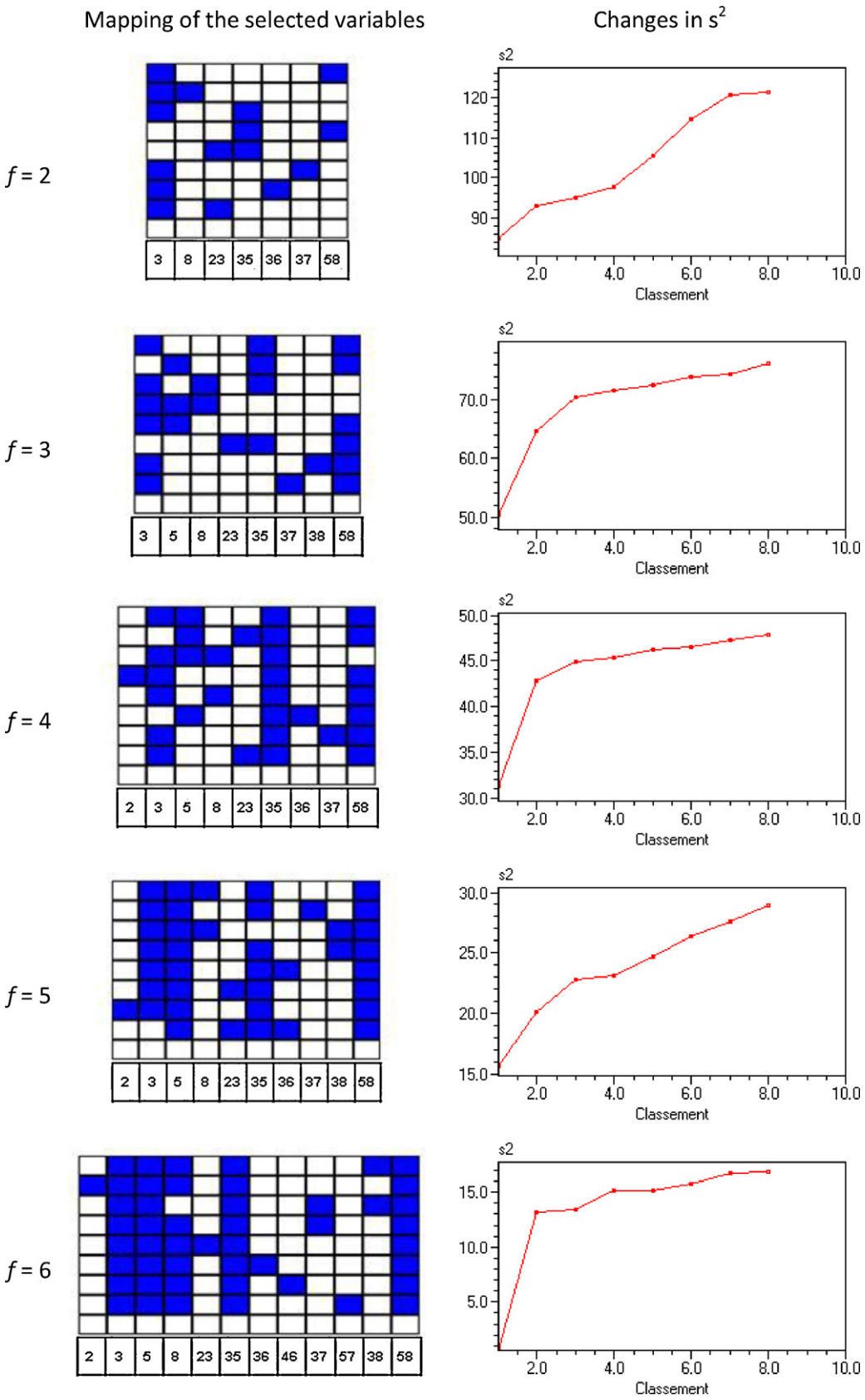

Fig. 5. All subset regressions for $f = 2$–8 variables. The first column in the table shows the number f of selected variables in each model. The second column shows the mapping of the selected variables where the presence of the variables is marked with a dark square. The third column shows the changes in s^2 for a fixed number of variables, corresponding to different models, numbered 1–8, with increasing s^2.

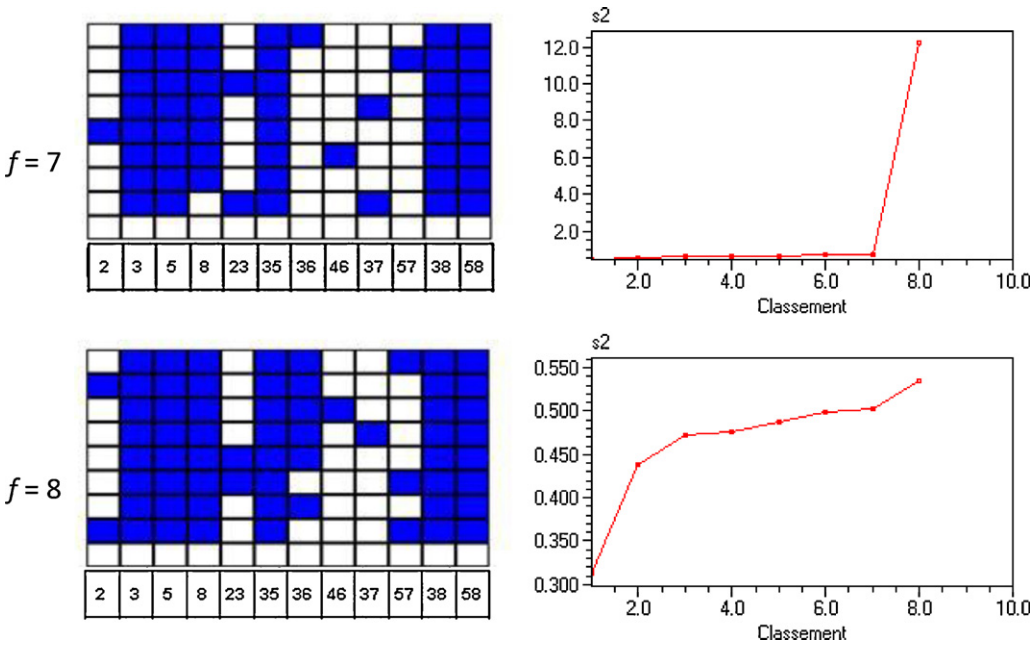

Fig. 5. (*Continued*)

Table 2
Patterns of effects in simulations.

Effect	Pattern 1A	Pattern 2A	Pattern 1B	Pattern 2B
b_0	60.00	60.00	60.00	60.00
b_1	0.00	0.00	−0.11	−0.11
b_2	0.00	−4.00	1.22	−4.00
b_3	6.00	7.00	6.00	7.00
b_4	0.00	0.00	−0.27	−0.27
b_5	7.00	0.00	7.00	0.09
b_6	0.00	−6.00	−1.28	−6.00
b_7	0.00	0.00	−0.01	−0.01
b_8	−5.00	0.00	−5.00	0.94
b_{12}	0.00	0.00	−0.32	−0.32
b_{13}	0.00	0.00	0.73	0.73
b_{14}	0.00	0.00	−0.65	−0.65
b_{15}	0.00	0.00	−1.34	−1.34
b_{16}	0.00	0.00	0.84	0.84
b_{17}	0.00	0.00	0.04	0.04
b_{18}	0.00	0.00	0.91	0.91
b_{23}	0.00	6.00	−0.33	6.00
b_{24}	0.00	0.00	−0.14	−0.14
b_{25}	0.00	0.00	0.71	0.71
b_{26}	0.00	5.00	−0.90	5.00
b_{27}	0.00	0.00	1.31	1.31
b_{28}	0.00	0.00	0.60	0.60
b_{34}	0.00	0.00	−0.18	−0.18
b_{35}	−5.00	0.00	−5.00	0.58
b_{36}	0.00	−6.00	0.64	−6.00
b_{37}	0.00	0.00	0.01	0.01
b_{38}	−4.00	0.00	−4.00	0.87
b_{45}	0.00	0.00	0.64	0.64
b_{46}	0.00	0.00	0.66	0.66
b_{47}	0.00	0.00	−0.32	−0.32
b_{48}	0.00	0.00	−0.23	−0.23
b_{56}	0.00	0.00	−1.18	−1.18
b_{57}	0.00	0.00	−0.46	−0.46
b_{58}	7.00	0.00	7.00	−0.02
b_{67}	0.00	0.00	−0.52	−0.52
b_{68}	0.00	0.00	0.27	0.27
b_{78}	0.00	0.00	−1.11	−1.11

parenthesis indicate false positive results that eventually appear in the process. The results corresponding to these same simulations using the optimal 20 rows size-reduced 2^8//37 matrix (see Table 1) are also included in Table 3 for comparison.

In general the results indicate quite good performance, even for randomly size-reduced matrices. Similar results were obtained for matrices RS1, RS2 and RS3, although matrix RS3 (fitness value = 1.2782) appeared slightly better than matrices RS1 (fitness = 1.4677) and RS2 (fitness = 1.3182). When the noise levels applied in simulation were not too high, the success rate was equal to or close to 100%, with a limited number of false positive results. As expected, performance decreased when noise increased, and when simulation pattern B was applied, which can be explained by smaller differences between the coefficients for active and non-active factors than those corresponding to pattern A (non-active factors have zero effect). Both resolution procedures were also able to identify the active factors (both main effects and interactions) in all simulation sets, with procedure B showing excellent performance.

These results encouraged us to develop a strategy to search for optimal size-reduced matrices. It was apparent that the basic hypothesis can be confirmed with randomly size-reduced matrices, although it was also apparent that the different matrices perform differently when tested against the same simulation runs, so that

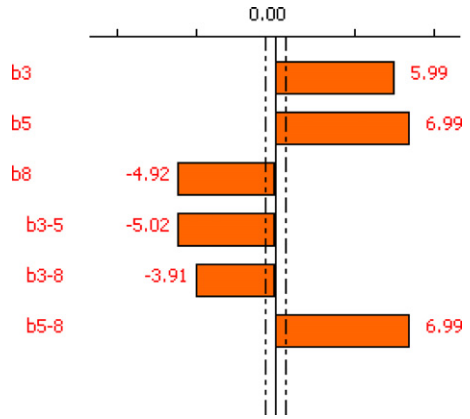

Fig. 6. Effects plot for the response "Pattern 1A" in the matrix RS1 (Noise level 1). The importance (coefficient values) of each variable is proportional to the length of the bar. The signification limits are drawn as vertical continuous line.

Table 3
Results of the preliminary simulation study, with fractions of 20 rows obtained from the 2^8//37 Rechstchaffner matrix. Comparison with results produced by the optimal 20 rows fraction matrix.

Coefficients pattern[a]	Resolution procedure[b]	Noise level (SD)	Percentage of success in active factors recognition (number of false positive results)			
			Matrix			
			RS1	RS2	RS3	Optimal
1A	A	1	100	50(2)	100	100
		2	100	100	100	100
		3	83	100	100	100
		4	83	100(1)	100	100
	CB	1	100	100	100	100
		2	100	100	100	100
		3	100	100	100	100
		4	100	100	100	100
1B	A	1	66	100(2)	83	100(1)
		2	83	100(2)	83(1)	100(2)
		3	66	100(2)	83(1)	100
		4	33(3)	50(3)	66(1)	100(1)
	CB	1	100	100	100	100
		2	100	100	100	100
		3	100	100	100	100
		4	100	83(1)	100	100
2A	A	1	66 (3)	66	100	100
		2	100 (2)	83(1)	100	100
		3	66	83(1)	100	100
		4	100	83(1)	83(1)	100(1)
	CB	1	100	100	100	100
		2	100	100	100	100
		3	100	100	100	100
		4	100	100	100	100
2B	A	1	50(1)	83	83	100
		2	50	83(1)	83(1)	100(2)
		3	83(1)	83(2)	83	100(2)
		4	50(1)	83	83	100(1)
	CB	1	100	100	100	100
		2	100	100	100	100
		3	100	100	100	100
		4	100	100	100	100

[a] See Table 2.
[b] A: direct evolutionary driven all subsets regression; CB: combined stepwise-all subsets regression plus projective properties procedure.

an optimization procedure was needed in building size-reduced supersaturated matrices.

Once the optimal matrices were available, a comparative study was carried out with these early experiments. The results corresponding to the optimal size-reduced matrix (Table 3, last column) clearly demonstrated the importance of the matrix chosen. In all cases, all active factors were accurately identified and there were very few false positive results, especially when using procedure B, which appears more robust in this sense. Furthermore, application of the heredity principle revealed that most of these false positive results can be disregarded [4,5]. Again, it is clear that the two resolution procedures allow accurate identification and estimation of active effects, while drastically reducing the work involved because the number of rows in the matrix was reduced by almost one half.

4.2. Testing against real world data

Once efficient procedures to build and solve the size-reduced Rechtschaffner matrices were established, a second test was conducted with recently published real world data. Nielloud et al. [17] described the optimization of operational variables in the formation of stable oil-in-water submicron emulsions used as vehicles for the drug delivery in different dermatological pharmaceutical routes. The authors applied a 2^5//16 Rechtschaffner matrix to study the influence of the following five factors: F_1 = oil

Table 4
Application of several fractions of the 2^5//16 Rechtschaffner matrix to data published by Niellou and al. [17].

Resolution procedure	Number of rows in the matrix[a]	Active factors identified				False positive results
Reference [17] results	16	F_1	F_3	F_5	F_{35}	
Evolutionary driven	7		F_3			F_{24}
all-subsets regression	8		F_3		F_{35}	F_{13}, F_{24}
(A)	9		F_3		F_{35}	F_{13}, F_{15}
	10	F_1	F_3	F_5	F_{35}	
	11	F_1	F_3	F_5	F_{35}	
Combination of	7		F_3			
stepwise and	8	F_1	F_3	F_5		
all-subsets regression	9		F_3			F_{24}
helped by projective	10		F_3	F_5		
properties (B)	11		F_3	F_5		F_{23}, F_{25}

[a] Using optimal fraction matrices defined in Table 1.

type; F_2 = mode of formulation; F_3 = % surfactant + co-surfactant; F_4 = ratio co-surfactant and F_5 = % oil, and their 10 two-factor interaction effects. Two responses were measured: particle size after 7 days of formation and visual evaluation after 50 days of formation, although the second response was imprecise and was therefore of little help. The experiment was not replicated, so the authors resorted to Bayesian approaches [1], to detect the active factors. The final conclusions were that factors F_1, F_3 and F_5 as well as the interaction F_{35} were active in the studied process.

Fortunately, this paper included all response data, which allowed us to use these data to check the performance of size-reduced $2^5//16$ matrices. The results of different size-reduced matrices are summarized (Table 4). Optimal $2^5//16$ size-reduced matrices with 7–11 rows were tested (see Table 1). The results demonstrate the need to use a sufficient number of rows for accurate regression when using the procedure A whereas procedure B appears less sensitive to this factor. The number of rows needed depends not only on the number of factors considered, but especially on the ratio of active factors and the total number of factors and the available degrees of freedom. Here, we see that only size-reduced matrices with at least 8–10 rows allow estimation of the real active factors without false positive results. Taking into account the results of the simulations and the experiments with real data, it is recommended to use size-reduced matrices with at least twice as many rows as the number of active factors when using the procedure A to evaluate the factors. Thus, in the case of the data reported by Niellou et al. [17], the matrix must have at least 10 rows. This means that the advantages of using size-reduced Rechtschaffner matrices as supersaturated matrices in screening studies are more evident with a relatively large number of main factors. For 5 factors, as in the example, the reduction factor would be only 1.6, however for 11 or 12 factors, which is quite frequent in practical studies, screening can be completed in 22–24 runs, which means a reduction factor equal to or greater than three. On the other hand, procedure B was not able to identify all active factors in the Niellou example. Of course, as initially discussed, one of the fundamental advantages of this approach is the possibility of carrying out a follow-up study to complete the original Rechstchaffner matrix in cases of doubtful results. This is a distinctive characteristic of the proposed approach in contrast with many currently used supersaturated design matrices.

5. Conclusions

For the first time, fractions of resolution V Rechstchaffner matrixes are proposed as supersaturated designs for screening studies. Efficient procedures for building optimal fractions and for the resolution of these fractions in practical situations have been found to exhibit very good performance, even in situations that largely exceed the limits for the application of supersaturated matrices. Simulation studies and work with real-world data clearly demonstrate that this tool opens up new perspectives in experimental studies aimed at detecting important factors in the early stages of process development, and essentially open up new perspectives for considering two factor interactions in these initial studies, something that has not previously been considered. Thus, the tool developed may be advantageous as regards reducing the experimental effort and providing enhanced, valuable information from the outset of screening studies.

Acknowledgments

Part of this research was developed under the support of the Spanish Ministry of Science and Innovation (MICIN), project no. CTQ2009-08377, and FEDER funds.

References

[1] G.E.P. Box, R.D. Meyer, Technometrics 28 (1986) 11–18.
[2] J.C. Wang, C.F.J. Wu, Stat. Sinica 5 (1995) 235–250.
[3] D.K.J. Lin, N.R. Draper, Technometrics 34 (1992) 423–428.
[4] M. Hamada, C.F.J. Wu, J. Qual. Technol. 24 (1992) 130–137.
[5] C.F.J. Wu, M. Hamada, Experiments Planning, Analysis, and Parameter Design Optimization, Wiley-Interscience, New York, 2000.
[6] F.E. Satterthwaite, Technometrics 1 (1959) 111–137.
[7] C.F.J. Wu, Biometrika 80 (1993) 661–669.
[8] Lin, Technometrics 37 (1995) 191–201.
[9] D.K.J. Lin, Technometrics 35 (1993) 28–31.
[10] W.W. Li, C.F.J. Wu, Technometrics 39 (1997) 171–179.
[11] B. Abraham, H. Chipman, K. Vijayan, Technometrics 41 (1999) 135–141.
[12] R.L. Rechtschaffner, Technometrics 9 (1967) 569–575.
[13] X. Qu, J. Statist. Plann. Inference 137 (2007) 2156–2164.
[14] X. Qu, C.F.J. Wu, J. Statist. Plann. Inference 131 (2005) 407–416.
[15] J.P.C. Kleijnen, Ö. Pala, Simulation 73 (1999) 168–173.
[16] M. Cesa, B. Campisi, A. Bizzotto, C. Ferraro, F. Fumagalli, P.L. Nimis, Arch. Environ. Contam. Toxicol. 55 (2008) 386–396.
[17] F. Nielloud, J.P. Mestres, R. Fortuné, S. Draussin, G. Marti-Mestres, Polym. Int. 52 (2003) 610–613.
[18] C. Cau Dit Coumes, S. Courtois, Chemom. Intell. Lab. Syst. 80 (2006) 167–175.
[19] G.E.P. Box, K.B. Wison, J. R. Statist. Soc. B 13 (1951) 1–38.
[20] C. Daniel, Applications of Statistics to Industrial Experimentation, Wiley, New York, 1976.
[21] N.R. Draper, D.K.J. Lin, Commun. Statist. A: Theory Methods 24 (1995) 775–795.
[22] J. Lawson, Comput. Statist. Data Anal. 39 (2002) 227–241.
[23] R. Cela, RFRACT: Application to Generate Optimal Fractions of Rechtschaffner Matrices, Santiago de Compostela, Spain, 2008.
[24] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, 3th ed., Springer, Berlin, 1996, pp. 97–105.
[25] P.H. Westfall, S.S. Young, D.K.J. Lin, Stat. Sinica 8 (1998) 101–107.
[26] A. Sudjianto, G.S. Wasserman, H. Sudarbo, Comput. Ind. Eng. 30 (1996) 839–849.
[27] H. Chipman, M. Hamada, C.F.J. Wu, Technometrics 39 (1997) 372–381.
[28] G.S. Wasserman, A. Sudjianto, Comput. Ind. Eng. 27 (1994) 489–492.
[29] D.E. Golberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, Reading, MA, 1989.
[30] R. Cela, SUPERSAT: All Subsets Regression Driven by Evolutionary Algorithms, Santiago de Compostela, Spain, 2002, www.usc.es/gcqprega/.
[31] R. Cela, M. Claeys-Bruno, R. Phan-Tan-Luu, Screening strategies, in: T.R. Brown, R. Tauler, S. Walczak (Eds.), Comprehensive Chemometrices, Elsevier, Oxford, 2009, pp. 251–300.
[32] D. Mathieu, J. Nony, R. Phan-Tan-Luu, NemrodW®, V-2007-03, LPRAI, Marseille, France.
[33] R.V. Lenth, Technometrics 31 (1989) 469–473.
[34] K.H.V. Booth, D.R. Cox, Technometrics 4 (1962) 489–495.
[35] B. Tang, C.F.J. Wu, Can. J. Stat. 25 (1997) 191–201.
[36] G. Schwarz, Ann. Statist. 6 (1978) 461–464.
[37] H. Akaike, IEEE Trans. Automat. Control 19 (1974) 716–723.
[38] F.K.H. Phoa, Y.-H. Pan, H. Xu, J. Statist. Plann. Inference 139 (2009) 2362–2372.
[39] G.E.P. Box, M.E. Muller, Ann. Math. Stat. 29 (1958) 610–611.