ELSEVIER

Chemometrics and Intelligent Laboratory Systems 42 (1998) 3–40

Chemometrics and intelligent laboratory systems

Experimental design and optimization

Torbjörn Lundstedt [a,*], Elisabeth Seifert [a], Lisbeth Abramo [b], Bernt Thelin [c], Åsa Nyström [a], Jarle Pettersen [a], Rolf Bergman [a]

[a] *Pharmacia and Upjohn Structure-Property Optimization Center, F3A-1, SE-751 82 Uppsala, Sweden*
[b] *Pharmacia and Upjohn Lund Research Center, P.O. Box 724, SE-220 07 Lund, Sweden*
[c] *Ferring, Pharmaceutical Analysis, P.O. Box 30047, SE-200 61 Malmö, Sweden*

Received 6 November 1997; accepted 11 May 1998

Abstract

The aim with this tutorial is to give a simple and easily understandable introduction to experimental design and optimization. The screening methods described in the paper are factorial and fractional factorial designs. Identification of significant variables are performed by normal distribution plots as well as by confidence intervals. Refinements of the models are also discussed. For optimization, the simplex method, central composite designs and the Doehlert design are discussed. The paper also gives an introduction to mixture designs. The paper contains 14 hands-on examples and if anyone needs the answers on these it is just to contact the authors.

Keywords: experimental design; factorial design; fractional factorial design; response surface optimization; simplex; Doehlert design; mixture design

Contents

* Corresponding author

0169-7439/98/$19.00
PII: S0169-7439(98)00065-3

1. Field of application

Experimental design and optimization are tools that are used to systematically examine different types of problems that arise within, e.g., research, development and production. It is obvious that if experiments are performed randomly the result obtained will also be random. Therefore, it is a necessity to plan the experiments in such a way that the interesting information will be obtained.

In the following pages, experimental design and optimization are presented to give the experimentalist useful tools in the real experimental situation, as well as the necessary theoretical background.

2. Definition of aim

What is the aim?

When the aim is well defined the problem should be analysed with the help of the following questions:

What is known? What is unknown? What do we need to investigate?

To be able to plan the experiments in a rational way the problem has to be concrete.

Which experimental variables can be investigated? Which responses can be measured?

When the experimental variables and the responses have been defined the experiments can be planned and performed in such a way that a maximum of information is gained from a minimum of experiments.

3. Terminology

To simplify the communication a few different terms are introduced and defined. Others will be defined when they are needed.

Experimental domain	the experimental 'area' that is investigated (defined by the variation of the experimental variables)
Factors	experimental variables that can be changed independently of each other
Independent variables	same as factors
Continuous variables	independent variables that can be changed continuously
Discrete variables	independent variables that are changed step-wise, e.g., type of solvent
Responses	the measured value of the result(s) from experiments
Residual	the difference between the calculated and the experimental result

4. Empirical models

It is reasonable to assume that the outcome of an experiment is dependent on the experimental conditions. This means that the result can be described as a function based on the experimental variables,

$$y = f(x)$$

The function $f(x)$ is approximated by a polynomial function and represents a good description of the relationship between the experimental variables and the responses within a limited experimental domain. Three types of polynomial models will be discussed and exemplified with two variables, x_1 and x_2.

The simplest polynomial model contains only linear terms and describes only the linear relationship between the experimental variables and the responses. In a *linear model*, the two variables x_1 and x_2 are expressed as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \text{residual}.$$

The next level of polynomial models contains additional terms that describe the interaction between different experimental variables. Thus, a *second order interaction model* contains the following terms:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + \text{residual}$$

The two models above are mainly used to investigate the experimental system, i.e., with screening studies, robustness tests or similar.

To be able to determine an optimum (maximum or minimum) quadratic terms have to be introduced in the model. By introducing these terms in the model, it is possible to determine non-linear relationships between the experimental variables and responses. The polynomial function below describes a *quadratic model* with two variables:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + b_{12} x_1 x_2 + \text{residual}$$

The polynomial functions described above contain a number of unknown parameters (b_0, b_1, b_2, etc.) that are to be determined. For the different models different types of experimental designs are needed.

5. Screening experiments

In any experimental procedure, several experimental variables or factors may influence the result. A screening experiment is performed in order to determine the experimental variables and interactions that have significant influence on the result, measured in one or several responses.

5.1. Early words of advice

- Specify the problem:
  Review the whole procedure—different moments, critical steps, raw material, equipment, etc. Try to get a holistic view of the problem.
- Define the response(s):
  Which response(s) can be measured?
  Which source(s) of errors can be assumed?
  Is it possible to follow the change in responses in course of time?
- Select variables:
  Which experimental variables are possible to study?
  Review and evaluate the variables—important, probably unimportant, etc.
  Select experimental domain.
  Are all variables interesting in the selected experimental domain?
  Which interaction effects can be expected?
  Which variables are probably not interacting?

This gives a list of possible responses, experimental variables and potential interaction effects. Penetrate this list critically a few times. The time spent on planning in the beginning of a project is always paid back with interest at the end.

5.1.1. Hints on variable selection

At this point, when the variables to be investigated are selected, it is also decided which variables that should *not* be investigated. These variables have to be kept at a fixed level in all experiments included in the experimental design. However, remember that it is always more economical to include a few extra variables in the first screening, than adding one variable later.

Think about *how* the different variables should be defined. It is sometimes possible to lower the number of experiments needed, in order to achieve the important information, just by redefining the original variables. Concentrations of different starting material can for example often be redefined as molar/molar ratios:

$$\text{concentrations } [A], [B] \text{ and } [C] = \text{three variables;}$$

give the same information as

$$\text{ratios } [A]/[B] \text{ and } [C]/[B] = \text{two variables.}$$

When all aspects have been penetrated, and variables, responses as well as experimental domain are selected, then it is time for the next step in the planning procedure.

5.1.2. Select a design and plan the experiments

When a list of variables to be investigated has been completed, an experimental design is chosen in order to estimate the influence of the different variables on the result. In screening studies, linear or second order interaction models are common, such as in full factorial or fractional factorial designs. The former design is limited to the determination of linear influence of the variables, while the latter allows for interaction terms between variables to be evaluated as well. Eventually, the variables with the largest influence on the procedure can be identified.

5.2. Factorial design

In a factorial design the influences of all experimental variables, factors, and interaction effects on the response or responses are investigated. If the combinations of k factors are investigated at two levels, a factorial design will consist of 2^k experiments. In Table 1, the factorial designs for 2, 3 and 4 experimental variables are shown. To continue the example with higher numbers, six variables would give $2^6 = 64$ experiments, seven variables would render $2^7 = 128$ experiments, etc. The levels of the factors are given by $-$ (minus) for low level and $+$ (plus) for high level. A zero-level is also included, a centre, in which all variables are set at their mid value. Three or four centre experiments should always be included in factorial designs, for the following reasons:

- the risk of missing non-linear relationships in the middle of the intervals is minimised, and
- repetition allows for determination of confidence intervals.

What $-$ and $+$ should correspond to for each variable is defined from what is assumed to be a reasonable variation to investigate. In this way the size of the experimental domain has been settled. For two and three variables the experimental domain and design can be illustrated in a simple way. For two variables the experiments will describe the corners in a quadrate (Fig. 1) while in a design with three variables they are the corners in a cube (Fig. 2).

5.2.1. Signs of interaction effects

The sign for the interaction effect between variable 1 and variable 2 is defined as the sign for the product of variable 1 and variable 2 (Table 2). The signs are obtained according to normal multiplication rules. By using these rules it is possible to construct sign columns for all the interactions in factorial designs.

Example 1: A 'work-through' example with three variables

This example illustrates how the sign tables are used to calculate the main effects and the interaction effects from a factorial design. The example is from an investigation of the influence from three experimental variables

Table 1
Factorial designs

Two variables			Three variables				Four variables				
Exp. no.	Variables		Exp. no.	Variables			Exp. no.	Variables			
	x_1	x_2		x_1	x_2	x_3		x_1	x_2	x_3	x_4
1	−	−	**1**	−	−	−	**1**	−	−	−	−
2	+	−	**2**	+	−	−	**2**	+	−	−	−
3	−	+	**3**	−	+	−	**3**	−	+	−	−
4	+	+	**4**	+	+	−	**4**	+	+	−	−
			5	−	−	+	**5**	−	−	+	−
			6	+	−	+	**6**	+	−	+	−
			7	−	+	+	**7**	−	+	+	−
			8	+	+	+	**8**	+	+	+	−
							9	−	−	−	+
							10	+	−	−	+
							11	−	+	−	+
							12	+	+	−	+
							13	−	−	+	+
							14	+	−	+	+
							15	−	+	+	+
							16	+	+	+	+

Note that all variables are changed simultaneously in a controlled way, to ensure that every experiment in each design is a unique combination of variable levels.

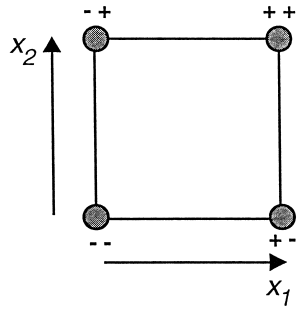


Fig. 1. The experiments in a design with two variables.

x_1, x_2 and x_3 on the yield of an organic synthesis. The variables and the experimental domain are specified in Table 3.

A sign table, or design matrix, for varying 3 variables according to a full factorial design is constructed in Table 4. The centre point, experiment number 9, is only added as a control to see if there is a non-linear relationship between the variables and the responses. If the value for the response of the centre point is very much different from the mean value, b_0, then it is necessary to include quadratic terms in the model. This means that additional experiments have to be performed.

The experiments are evaluated in order to fit a polynomial model, in this case a third order interaction model:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + b_{123} x_1 x_2 x_3$$

Table 4 is used to calculate the main effects and the mean value, b_0.

$$b_0 = 1/8(73 + 71 + 79 + 82 + 78 + 89 + 83 + 93) = 81$$

The main effects are calculated by using the signs in the corresponding columns and either add or subtract the value of the response, y. This summation is finally divided with the number of experiments (in this case 8).

$$b_1 = 1/8(-73 + 71 - 79 + 82 - 78 + 89 - 83 + 93) = 2.8$$

$$b_2 = 1/8(-73 - 71 + 79 + 82 - 78 - 89 + 83 + 93) = 3.2$$

$$b_3 = 1/8(-73 - 71 - 78 - 82 + 78 + 89 + 83 + 93) = 4.9$$

To be able to calculate the signs for the interaction effects, the corresponding columns are constructed for each effect—in this case, $x_1 x_2$, $x_1 x_3$, $x_2 x_3$ and $x_1 x_2 x_3$. The signs for these columns are achieved by multiplying the columns for the corresponding main effects (Table 5). This type of matrix is called model matrix or calculation matrix. The centre point experiment is removed since it is not used in the calculation of the effects.

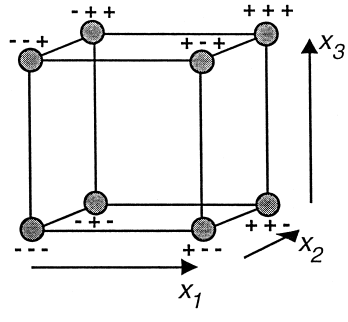


Fig. 2. The experiments in a design with three variables.

Table 2
Sign of interaction effect x_1x_2

x_1	x_2	x_1x_2
−	−	+
+	−	−
−	+	−
+	+	+

Table 3
Specification of variables and the experimental domain

Variables	Experimental domain		
	(−)-level	0-level	(+)-level
x_1: Catalyst concentration (%)	0.1	0.2	0.3
x_2: Reaction temperature (°C)	60	70	80
x_3: Reaction time (min)	20	30	40

Table 4
Design and the yield response

Exp. no.	x_1	x_2	x_3	Yield (%)
1	−	−	−	73
2	+	−	−	71
3	−	+	−	79
4	+	+	−	82
5	−	−	+	78
6	+	−	+	89
7	−	+	+	83
8	+	+	+	93
9	0	0	0	81

Table 5
Model matrix and the yield response

Exp.	I	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	Yield (%)
1	+	−	−	−	+	+	+	−	73
2	+	+	−	−	−	−	+	+	71
3	+	−	+	−	−	+	−	+	79
4	+	+	+	−	+	−	−	−	82
5	+	−	−	+	+	−	−	+	78
6	+	+	−	+	−	+	−	−	89
7	+	−	+	+	−	−	+	−	83
8	+	+	+	+	+	+	+	+	93

Table 6
Levels of the variables

	Experimental domain	
	(−)-level	(+)-level
x_1: Granv	25,37	25,93
x_2: Sfhast	650	950
x_3: Tidsfpl	60	180

The calculations of the interaction effects are done in the same way as for the main effects.

$$b_{12} = 1/8(73 - 71 - 79 + 82 + 78 - 89 - 83 + 93) = 0.5$$

$$b_{13} = 1/8(73 - 71 + 79 - 82 - 78 + 89 - 83 + 93) = 2.5$$

$$b_{23} = 1/8(73 + 71 - 79 - 82 - 78 - 89 + 83 + 93) = -1.0$$

$$b_{123} = 1/8(-73 + 71 + 79 - 82 + 78 - 89 - 83 + 93) = -0.8$$

The estimated effects are then placed in the polynomial model describing the relationship between the variables:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + b_{123} x_1 x_2 x_3$$

$$y = 81 + 2.8 x_1 + 3.2 x_2 + 4.9 x_3 + 0.5 x_1 x_2 + 2.5 x_1 x_3 - 1.0 x_2 x_3 - 0.8 x_1 x_2 x_3$$

The function above is now describing how the experimental variables and their interactions influence the response y. The model shows that variable x_3 has the largest influence on the yield. An increase of variable 3 with one scaled unit (e.g., from 0 to +1) results in an increase of the yield by 4.9%. This corresponds, in real variables, to an increase of the reaction time by 10 min.

Example 2: A 2^3-factorial design: Pharmacy, pellets

This is an example from a project within Pharmacia. A process for producing caffeine pellets with a certain size (0.71–1.4 mm) was studied. The aim was to obtain a robust process giving a yield higher than 95% of this fraction.

Three variables were investigated: amount of water in the granulation (Granv), spheronizer speed (Sfhast) and spheronizer time (Tidsfpl). A 2^3-full factorial design was used to study the robustness of the process. The levels of the variables and the design used are shown in Tables 6 and 7.

Exercise

Investigate the influence of the experimental variables on the yield using the following response model:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + b_{123} x_1 x_2 x_3$$

Example 3: A 2^3-factorial design: Pharmacy, formulation of tablets

In the formulation of a certain tablet three variables were considered to be important for the thickness of the tablets. These variables were investigated by a factorial design. The different variables were the amount of stearate (lubricant), the amount of active substance and the amount of starch (disintegrant). The experimental domain is shown in Table 8. Experimental design and results are given in Table 9.

Table 7
Design and yield response

Exp. no.	x_1	x_2	x_3	Yield (0.71–1.4 mm) (%)
1	−	−	−	97.4
2	+	−	−	98.1
3	−	+	−	97.1
4	+	+	−	97.8
5	−	−	+	98.6
6	+	−	+	98.2
7	−	+	+	98.3
8	+	+	+	98.3

Table 8
Variables and experimental domain of the formulation

Variables	Experimental domain		
	(−)-level	0-level	(+)-level
x_1: Amount of stearate (mg)	0.5	1	1.5
x_2: Amount of active substance (mg)	60	90	120
x_3: Amount of starch (mg)	30	40	50

Exercises

(a) Estimate the effects (coefficients) of the experimental variables and evaluate their influence.

(b) Determine a response model that contains only the probably significant terms. Use this model to estimate the amount of starch that has to be added to 100 mg of active substance to obtain tablets that are 5.00 mm thick.

Example 4: Screening: Candy production, 'sega råttor'

You are working in a small company producing candy and the company would like to improve the production of 'sega råttor'. You have the following recipe as a starting point for your investigation.

Recipe:

90 g of sugar
50 g of glucose
25 ml of water
40 g of gelatine solution (14 g of gelatine in 26 ml of water)

Production procedure: The gelatine powder and water are mixed and heated (carefully) until all of the gelatine is dissolved. The solution is cooled to room temperature. Sugar, glucose and water are mixed and then heated to 114°C. The mixture is then cooled to 100°C and the gelatine solution is added. Colour and flavour may be added. Let the mixture stand for 15 min at 80°C. It is important to hold the temperature constant. Then pour the mixture into the forms covered with potato-starch and let it stand for 4 h. The forms can be opened and the product is ready.

Exercises

(a) Read the recipe and make a suggestion of possible experimental variables and make a critical review. Suggest possible responses.

(b) Make a suggestion of a factorial design.

Table 9
Design and responses

Exp. no.	Variables			Thickness (mm)
	x_1	x_2	x_3	y
1	−	−	−	4.75
2	+	−	−	4.87
3	−	+	−	4.21
4	+	+	−	4.26
5	−	−	+	5.25
6	+	−	+	5.46
7	−	+	+	4.72
8	+	+	+	5.22
9	0	0	0	4.86

5.3. Fractional factorial design

To investigate the effects of k variables in a full factorial design, 2^k experiments are needed. Then, the main effects as well as all interaction effects can be estimated. To investigate seven experimental variables, 128 experiment will be needed; for 10 variables, 1024 experiments have to be performed; with 15 variables, 32,768 experiments will be necessary. It is obvious that the limit for the number of experiments it is possible to perform will easily be exceeded, when the number of variables increases.

In most investigations it is reasonable to assume that the influence of the interactions of third order or higher are very small or negligible and can then be excluded from the polynomial model. This means that 128 experiments are too many to estimate the mean value, seven main effects and 21 second order interaction effects, all together 29 parameters. To achieve this, exactly 29 experiments are enough. On the following pages it is shown how the fractions $(1/2, 1/4, 1/8, 1/16 \ldots 1/2^p)$ of a factorial design with 2^{k-p} experiments are defined, where k is the number of variables and p the size of the fraction. The size of the fraction will influence the possible number of effects to estimate and, of course, the number of experiments needed.

If only the main effects are to be determined it is sufficient to perform only 4 experiments to investigate 3 variables, 8 experiments for 7 variables, 16 experiments for 15 variables, etc. This corresponds to the following response function:

$$y = \beta_0 + \Sigma\beta_i x_i + \varepsilon$$

It is always possible to add experiments in order to separate and estimate interaction effects, if it is reasonable to assume that they influence the result. This corresponds to the following second order response function:

$$y = \beta_0 + \Sigma\beta_i x_i + \Sigma\Sigma\beta_{ij} x_i x_j + \varepsilon$$

In most cases, it is not necessary to investigate the interactions between all of the variables included from the beginning. In the first screening it is recommended to evaluate the result and estimate the main effects according to a linear model (if it is possible to calculate additional effects they should of course be estimated as well). After this evaluation the variables that have the largest influence on the result are selected for new studies. Thus, a large number of experimental variables can be investigated without having to increase the number of experiments to the extreme.

5.3.1. The principals for constructing fractional factorial designs

The model matrix from a factorial design is used to define the design matrix in fractional factorial designs. This means that columns in the model matrix **X** for a full factorial design are used to define the settings for the 'extra' variables in a series of experiments (performed according to a fractional factorial design). The number of variables that can be included is limited by the number of columns in **X**. On the following pages, Arabic numbers with bold style will be used to indicate the variable columns in the matrix. I is used to indicate the column used for calculating the mean value (the constant term in the model).

Example 5: A factorial design with two variables is shown in Table 10.

The columns **a**, **b** and **ab** define the settings for three variables, x_1, x_2 and x_3, in four experiments. The column **ab** is the product of **a** · **b**. See Table 11 for the obtained design matrix.

Table 10
Model matrix

I	**a**	**b**	**ab**
1	−1	−1	1
1	1	−1	−1
1	−1	1	−1
1	1	1	1

Table 11
Design matrix

Exp. no.	x_1	x_2	x_3
1	−1	−1	1
2	1	−1	−1
3	−1	1	−1
4	1	1	1

Table 12
Experimental plan

Exp. no.	x_1	x_2	x_3
1	−1	−1	−1
2	1	−1	−1
3	−1	1	−1
4	1	1	−1
5	−1	−1	1
6	1	−1	1
7	−1	1	1
8	1	1	1

This is a half fraction of a factorial design with three variables and it is found that the selected experiments correspond to experiment 5, 2, 3 and 8 in the factorial design (Table 12).

In Fig. 3, it is illustrated how the experiments for a half fraction are distributed in the experimental domain spanned by three variables. It is also shown that the experiments have the form of a tetrahedron. This is the largest possible volume spanned by four corners in three dimensions.

This shows another important property of the fractional factorial designs, that the experiments cover as much as possible of the experimental domain. The whole experimental domain cannot be covered by a limited number of experiments, but a fractional factorial design 'selects' those experiments that cover a maximal volume of the domain in a limited number of experiments.

Example 6: Seven variables can be studied in a 2^{7-4} fractional factorial design. The design is defined by the model matrix for a 2^3 factorial design (Table 13).

A factorial design with seven variables gives 128 experiments. The experiments in the design define 'the corners' in a hyper cube within the seven-dimensional space spanned by the seven variables. The 2^{7-4} fractional factorial design is $1/16$ of the factorial design and the eight experiments are selected in order to span the largest

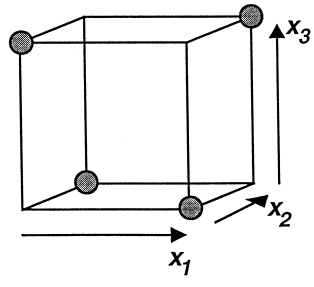


Fig. 3. Distribution of the experiments in a 2^{3-1} fractional factorial design.

Table 13
Design matrix for a 2^{7-4} fractional factorial design

Exp. no.	I	x_1	x_2	x_3	x_4	x_5	x_6	x_7
		a	**b**	**c**	**ab**	**ac**	**bc**	**abc**
1	1	−1	−1	−1	1	1	1	−1
2	1	1	−1	−1	−1	−1	1	1
3	1	−1	1	−1	−1	1	−1	1
4	1	1	1	−1	1	−1	−1	−1
5	1	−1	−1	1	1	−1	−1	1
6	1	1	−1	1	−1	1	−1	−1
7	1	−1	1	1	−1	−1	1	−1
8	1	1	1	1	1	1	1	1

possible experimental domain. In the fractional factorial design, the variables $x_4 - x_7$ are defined by the columns for the interactions between the variables **a**, **b** and **c**. The columns in a fractional factorial design are thus orthogonal. It is therefore possible to estimate the mean value and the main effects independent of each other and with a maximal precision.

In the same way as in example 6, the columns in a 2^4 factorial design can be used to define the variation of up to 15 variables in 16 experiments, i.e., a 2^{15-11} fractional factorial design. This is a $1/2048$ part of a 2^{15} factorial design.

These designs are employed to fit the experimental data to a linear model.

$$y = \beta_o + \Sigma\beta_i x_i + \varepsilon$$

The calculations are performed as for factorial designs that were described earlier.

In fractional factorial designs, many variables can be investigated without an excessive number of experiments. Less information is gained compared to full factorial designs, and the price to be paid for the few experiments is the 'contamination' of the main effects by the interaction effects. The main effects are *confounded*.

5.3.2. Confounding and aliases

To describe what confounding means, an example with three variables investigated in a 2^{3-1} fractional factor design is shown. In such a design, it is easily seen that for the experiments each variable always varies as the product of the other two variables, while column I is the product of all the three variables, $x_1 \cdot x_2 \cdot x_3$ (Table 14).

If a factorial design with three variables is chosen the following response model is used:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \varepsilon$$

Table 14
Design matrix for a 2^{3-1} fractional factorial design

Exp. no.	$x_1x_2x_3$	x_2x_3	x_1x_3	x_1x_2
	I	x_1	x_2	x_3
1	1	−1	−1	1
2	1	1	−1	−1
3	1	−1	1	−1
4	1	1	1	1

Since only four experiments are performed it is only possible to estimate four parameters in the model. In the reduced design the following relations are defined:

$$x_1 = x_2 x_3$$
$$x_2 = x_1 x_3$$
$$x_3 = x_1 x_2$$
$$I = x_1 x_2 x_3$$

If these are introduced in the model above the following is obtained

$$y = \beta_0 + \beta_{123} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_3 + \beta_{13} x_2 + \beta_{23} x_1 + \varepsilon$$

This can be reduced to

$$y = (\beta_0 + \beta_{123}) + (\beta_1 + \beta_{23}) x_1 + (\beta_2 + \beta_{13}) x_2 + (\beta_3 + \beta_{12}) x_3 + \varepsilon$$

which means that a linear model with four parameters is obtained:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$$

In this model, the parameters are linear combinations of the 'true' effects,

$$b_0 = \beta_0 + \beta_{123}$$
$$b_1 = \beta_1 + \beta_{23}$$
$$b_2 = \beta_2 + \beta_{13}$$
$$b_3 = \beta_3 + \beta_{12}$$

The parameter b_1 is an estimate of the 'true' parameter β_1, but this estimate is contaminated by the 'true' two-variable interaction β_{23}. In other words β_1 is confounded with β_{23} and b_1 is an alias of the confounded effects. This is the price to pay for using fractional factorial designs. It is always possible to estimate the effects but they will be confounded.

5.3.3. The generators of a fractional factorial design

In order to analyse which effects that are going to be confounded a new term is introduced: the *generator* of a fractional factorial design. As an example, the 2^{3-1} fractional factorial design is used. To understand why the generators are practical to use, the model matrix for a 2^3 factorial design is given (Table 15).

In four of these experiments variable **c** varies in the same way as the product of **a** and **b**. These experiments are marked by an $*$. For these experiments it is also seen that **a** = **bc, b** = **ac** and ***I*** = **abc**. It is very difficult to identify such relations in a model matrix, especially when the matrix is large. The four experiments fulfilling that **c** = **ab** are one half of the full factorial design. Four experiments correspond to a 2^2 factorial design. This means that we can use the model matrix for a 2^2 design to define the variation for the 'extra' variable **c**, and

Table 15
Model matrix for a 2^3 factorial design

	I	**a**	**b**	**c**	**ab**	**ac**	**bc**	**abc**
	1	−1	−1	−1	1	1	1	−1
*	1	1	−1	−1	−1	−1	1	1
*	1	−1	1	−1	−1	1	−1	1
	1	1	1	−1	1	−1	−1	−1
*	1	−1	−1	1	1	−1	−1	1
	1	1	−1	1	−1	1	−1	−1
	1	−1	1	1	−1	−1	1	−1
*	1	1	1	1	1	1	1	1

Table 16
Model matrix for a 2^{3-1} fractional factorial design

I	**a**	**b**	**c** **ab**
1	−1	−1	1
1	1	−1	−1
1	−1	1	−1
1	1	1	1

thus obtain the 2^{3-1} fractional factorial design. The half fraction, in which **c** = **ab**, is obtained by letting column **ab** define the variation for variable **c** in the model matrix for the 2^2 design.

The experiments above are now the same as those selected from the factorial design marked with an $*$. The matrix in Table 16 has some special mathematical properties. If the columns are multiplied to each other (in any combination) one of the other columns will always be obtained. It is also seen that

$$I \cdot \mathbf{a} = \mathbf{a} \text{ i.e., } I\mathbf{a} = \mathbf{a}$$
$$I \cdot \mathbf{b} = \mathbf{b} \text{ i.e., } I\mathbf{b} = \mathbf{b}$$
$$I \cdot \mathbf{ab} = \mathbf{ab} \; i.e., \; I\mathbf{ab} = \mathbf{ab}$$

Multiplication by I does not change anything since the elements in the column are multiplied by $(+1)$. If a column is multiplied by itself, it means that (-1) is multiplied by (-1) and $(+1)$ by $(+1)$. This will always result in column I.

$$\mathbf{a} \cdot \mathbf{a} = \mathbf{a}^2 = I$$
$$\mathbf{b} \cdot \mathbf{b} = \mathbf{b}^2 = I$$
$$\mathbf{ab} \cdot \mathbf{ab} = \mathbf{a}^2 \cdot \mathbf{b}^2 = I \quad \text{etc.}$$

Variable **c** in the fractional factorial design above is varied as the product **ab**. By doing this the effects β_3 and β_{12} will be confounded. Other effects will also be confounded.

The fractional factorial design is generated from the factorial design by defining **c** = **ab**. By using the calculation rules defined above, both sides in the relation **c** = **ab** are multiplied by **c** and the following is obtained:

$$\mathbf{c} \cdot \mathbf{c} = \mathbf{ab} \cdot \mathbf{c}$$

which gives

$$I = \mathbf{abc}$$

This relation is the *generator* and contains the information about how the different columns can be multiplied to obtain column I. The generator can now be used to identify the confounding as follows: each column in the

Table 17
Model matrix for a 2^{3-1} fractional factorial design

I	**a**	**b**	**ab**
abc	**bc**	**ac**	**c**
1	−1	−1	1
1	1	−1	−1
1	−1	1	−1
1	1	1	1

matrix is multiplied by the generator for the fractional factorial design. In this way multiplication of $I = \mathbf{abc}$ with I, **a**, **b** and **ab** gives

$I \cdot (I = \mathbf{abc})$ i.e., $I = \mathbf{abc}$

$\mathbf{a} \cdot (I = \mathbf{abc})$ i.e., $\mathbf{a} = \mathbf{bc}$

$\mathbf{b} \cdot (I = \mathbf{abc})$ i.e., $\mathbf{b} = \mathbf{ac}$

$\mathbf{ab} \cdot (I = \mathbf{abc})$ i.e., $\mathbf{ab} = \mathbf{c}$

In Table 17, the column headers provide an overview of the results of the multiplications. The model matrix is for a 2^{3-1} fractional factorial design.

When these columns are used to calculate the effects the following confounded effects are determined:

From column	I	the estimate	$\beta_0 + \beta_{123}$
	a	the estimate	$\beta_1 + \beta_{23}$
	b	the estimate	$\beta_2 + \beta_{13}$
	ab	the estimate	$\beta_{12} + \beta_3$

5.3.4. More generators

To investigate five variables in a factorial design, 32 experiments are needed. The screening study aims to determine the importance of the experimental variables. In this case, it is sufficient to start with a 2^{5-2} fractional factorial design and only eight experiments have to be performed. To define such a design the model matrix for a 2^3 factorial design is used. This matrix contains the following columns:

I **a b c ab ab bc abc**

To define the two extra variables **d** and **e**, any of the four interaction columns can be used. In this example we choose to define the 'extra' variables in the following way:

$\mathbf{d} = \mathbf{bc}$ i.e., $\mathbf{d}^2 = \mathbf{bcd}$

$\mathbf{e} = \mathbf{abc}$ i.e., $\mathbf{e}^2 = \mathbf{abce}$,

resulting in the *independent generators*:

$I = \mathbf{bcd} = \mathbf{abce}$

The multiplication rules are also valid for the independent generators. When they are multiplied with each other, column I is obtained. This means that

$I = \mathbf{bcd} \cdot \mathbf{abce}$

which is simplified to

$I = \mathbf{ade}$

Therefore a *dependent generator* has to be added to the independent generators to obtain a complete set of generators

$I = \mathbf{bcd} = \mathbf{abce} = \mathbf{ade}$

This set of generators contains four 'words'. The number of 'words' is determined by the fact that a 2^{5-2} design is $1/4$ of factorial design. From this set of generators, the following confounding pattern is found (Table 18).

Table 18
Example of confounding

I	**a**	**b**	**c**	**ab**	**ac**	**bc**	**abc**
bcd	abcd	cd	bd	acd	abd	d	ad
abce	bce	ace	abe	ce	be	ae	e
ade	de	abde	acde	bde	cde	abcde	bcde

If it is assumed that interactions of higher order than three have a negligible influence on the responses, compared to the main effects and the second order interaction effects, the following estimates of the model parameters will be calculated from the experiments:

From column	I	the estimate	β_0
	a	the estimate	$\beta_1 + \beta_{45}$
	b	the estimate	$\beta_2 + \beta_{34}$
	c	the estimate	$\beta_3 + \beta_{24}$
	ab	the estimate	$\beta_{12} + \beta_{35}$
	ac	the estimate	$\beta_{13} + \beta_{25}$
	bc	the estimate	$\beta_4 + \beta_{23} + \beta_{15}$
	abc	the estimate	$\beta_5 + \beta_{14}$

The variables **d** and **e** could have been defined by other columns, e.g., **d** = **ab** and **e** = **ac**, which would have resulted in another set of generators:

$I =$ **abd** = **ace** = **bcde**

Another pattern of confounding would then have been found.

5.3.5. Resolution of a fractional factorial design

The resolution of a fractional factorial design is defined by the shortest 'word' in the set of generators. Roman numbers usually specify the resolution. In a design of

- resolution III — the main effects are confounded with two-variable interaction effects;
- resolution IV — the main effects are confounded with three-variable interaction effects, and the two-variable interaction effects are confounded with each other;
- resolution V — the main effects are confounded with four-variable interaction effects, and the two-variable interaction effects are confounded with the three-variable interaction effects.

Fractional factorial designs of resolution higher than V are rarely used in screening.

5.3.6. How to separate confounded effects

In the previous example, a 2^{5-2} fractional factorial design was used that corresponded to a $1/4$ of a factorial design. The additional variables **d** and **e** were defined as **d** = **bc** and **e** = **abc**, which gave the independent generators

$I =$ **bcd** = **abce**

The design has the resolution III, i.e., the main effects are confounded with two-variable interaction effects. A new quarter is obtained by changing the sign on column **d** = **−bc**. Another one is obtained by changing the sign on column **e** = **−abc**. Finally, you get the last quarter by changing the signs on both columns, **d** = **−bc** and **e** = **−abc**. These different ways of defining the additional variables correspond to different sets of generators:

Design A	$I =$ **bcd** = **abce** = **ade**
Design B	$I =$ **−bcd** = **abce** = **−ade**
Design C	$I =$ **bcd** = **−abce** = **−ade**
Design D	$I =$ **−bcd** = **−abce** = **ade**

The first two are the independent generators and $\pm$**ade** is the dependent generator. Since the sets of generators are different, the confounding pattern obtained will also be different. To get an overview of the confounding, only the main effects and the two-variable interaction effects are included.

Table 19
Overview of confounding

Design A	Design B	Design C	Design D
β_0	β_0	β_0	β_0
$\beta_1 + \beta_{45}$	$\beta_1 - \beta_{45}$	$\beta_1 - \beta_{45}$	$\beta_1 + \beta_{45}$
$\beta_2 + \beta_{34}$	$\beta_2 - \beta_{34}$	$\beta_2 + \beta_{34}$	$\beta_2 - \beta_{34}$
$\beta_3 + \beta_{24}$	$\beta_3 - \beta_{24}$	$\beta_3 + \beta_{24}$	$\beta_3 - \beta_{24}$
$\beta_{12} + \beta_{35}$	$\beta_{12} + \beta_{35}$	$\beta_{12} - \beta_{35}$	$\beta_{12} - \beta_{35}$
$\beta_{13} + \beta_{25}$	$\beta_{13} + \beta_{25}$	$\beta_{13} - \beta_{25}$	$\beta_{13} - \beta_{25}$
$\beta_4 + \beta_{23} + \beta_{15}$	$\beta_4 - \beta_{23} - \beta_{15}$	$\beta_4 + \beta_{23} - \beta_{15}$	$\beta_4 - \beta_{23} + \beta_{15}$
$\beta_5 + \beta_{14}$	$\beta_5 - \beta_{14}$	$\beta_5 - \beta_{14}$	$\beta_5 + \beta_{14}$

Now suppose that a series of experiments has been performed according to the design A. In the overview in Table 19, it is shown that design B is complementary to design A in such a way that all main effects will be separated from the two-variable interaction effects. Performing new experiments according to design B would thus clear the main effects, although the two-variable interaction effects would still be confounded with each other.

When a fractional factorial design has been used, it is always possible to add a complementary fraction in order to separate confounded effects.

5.3.7. All experiments are useful

If it is concluded that a variable has no influence on the response, the amount of information available has increased. The fractional factorial design then turns in to a larger fraction due to the use of fewer variables. The experiments performed can now be used to estimate interaction effects that were confounded earlier. For three variables this is illustrated in Fig. 4.

If one or several variables have no influence on the response, the new confounding pattern can be determined by using the generators for the original design. Then it is just to remove all the elements containing the 'unimportant' variable.

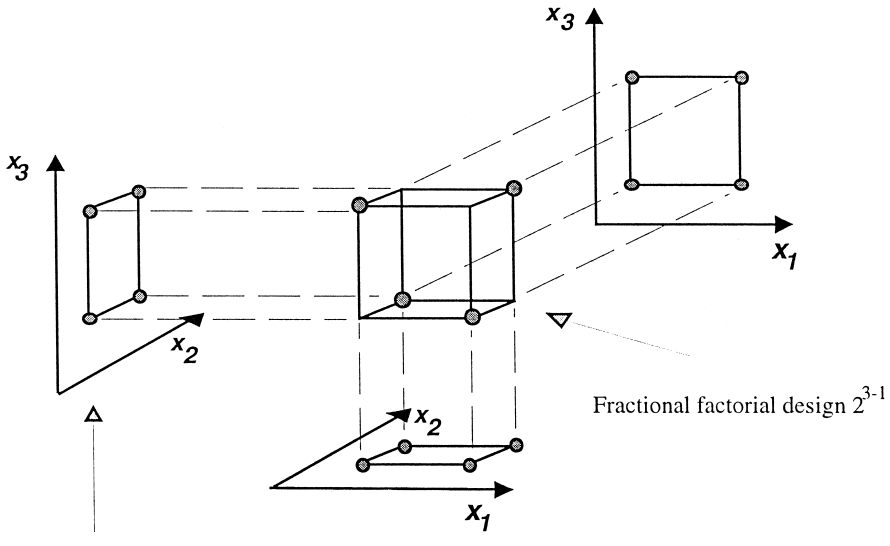


Fig. 4. The effect obtained for a 2^{3-1} fractional factorial design when a variable can be removed.

5.4. Evaluation of models

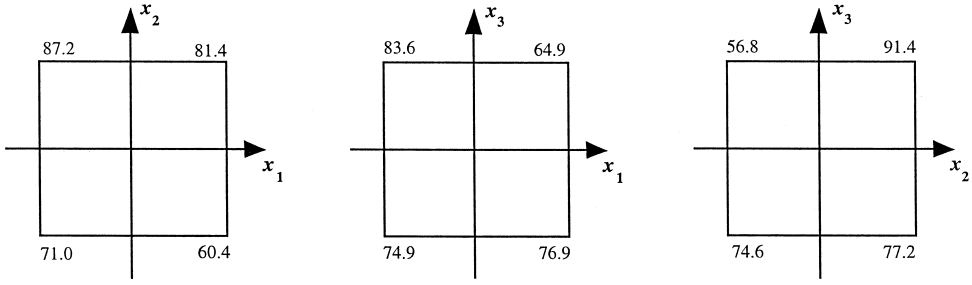
In factorial or fractional factorial designs all variables are normalised to vary between -1 and $+1$. For continuous variables, the scaling is made in such a way that the original variables vary continuously within the interval from -1 to $+1$. Since all variables used in the model are normalised in this way, the relative change of a variable is directly related to the size of its regression coefficient. This means that if the model parameters have either a large positive or negative value the corresponding variable has a large influence on the response(s).

If a response model contains interaction terms, $\beta_{ij} x_i x_j$, the evaluation of the influence of different variables on the response is simplified by projections of the response surface on to the (x_i, x_j)-plane. This means to calculate the response values for $x_1 = \pm 1$ and $x_2 = \pm 1$, while the other variables are kept constant within the experimental domain.

Example 7: The yield, y (%), of a process influenced by the three variables x_1, x_2 and x_3, is described by the following response model:

$$y = 75.0 - 4.1 x_1 + 9.3 x_2 - 0.9 x_3 + 1.2 x_1 x_2 - 5.1 x_1 x_3 + 8.0 x_2 x_3$$

In this case, the projections are made through zero, i.e., the variable not included in the projection has been kept at the 0-level.



If: $x_3 = 0$;	$x_2 = 0$;	$x_1 = 0$
Best conditions:		
$x_1 = -1$	$x_1 = -1$	$x_2 = 1$
$x_2 = 1$	$x_3 = 1$	$x_3 = 1$

It is concluded that a very high yield, $y = 99.4\%$, should be possible to obtain with the variable settings $x_1 = -1$, $x_2 = 1$ and $x_3 = 1$.

5.4.1. Normal distribution plots to identify significant effects

If an independent estimation of the experimental error has been done (i.e., determination of s^2 through repetition of an experiment several times, for example), statistical distributions can be used to compare the estimated effects with the experimental error. This kind of comparison is usually not very fruitful since the degree of freedom often is low. In screening studies, the lack of repeated experiments means that confidence intervals cannot be determined.

Another method to study the experimental error is to use normal distribution plots. It is a fast and simple method to rapidly get an indication if any of the estimated effects are diverging significantly from the normal distribution. If an effect has a large deviation from the normal distribution it probably describes something else than the experimental noise.

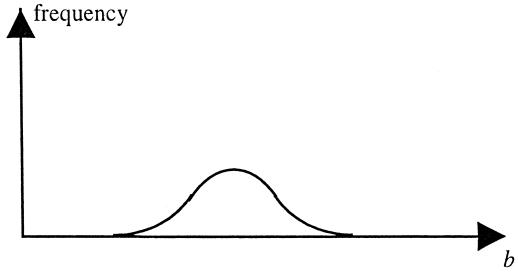


Fig. 5. A normal distribution curve.

Consider a set of estimated coefficients $(b_1, b_2 \ldots b_n)$ that is normally distributed. The frequency will then describe the bell-shaped normal distribution curve as in Fig. 5.

If instead a cumulative probability distribution is desired, i.e., the probability P (the shadowed area in Fig. 6) for a measured value less than $b = a$, this is done by plotting P vs. b. For a normal distribution, this will be an S-shaped Curve (Fig. 7).

The P-axis is then adjusted to straighten the curvature and describes the normal distribution as a straight line (Fig. 8).

After the effects of the variables have been estimated from a factorial design, their coefficients are plotted against probability, as in Fig. 8. Then it is possible to distinguish effects from normally distributed noise.

When a normal distribution plot is made, the effects have to be ordered after increasing size, mean value b_0 excluded:

$$\underset{1}{\text{smallest}}\,,\ \underset{2}{\text{second smallest}}\,,\ldots\ \underset{q}{\text{largest}}$$

If the experiments have been performed in a randomised order we can see the q effects as a randomised sample from a normally distributed population of experimental noise. This is true if the variables do not have any effect.

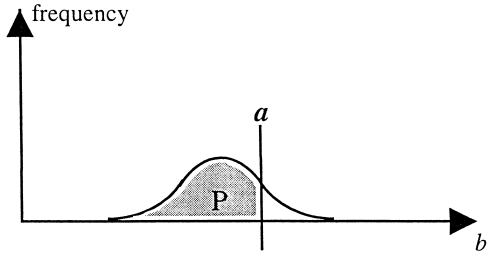


Fig. 6. Frequency vs. b.

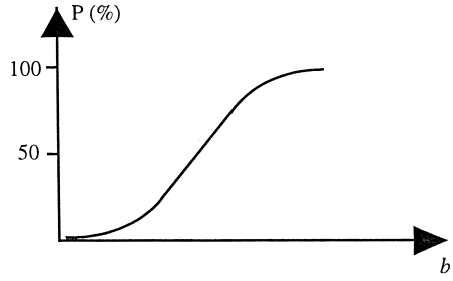


Fig. 7. Probability P vs. b.

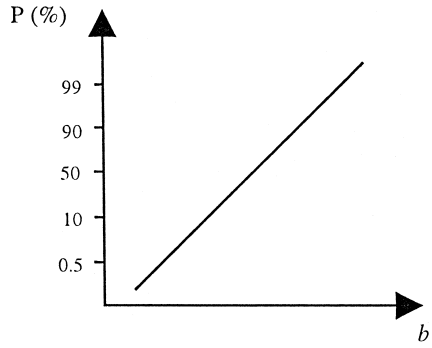


Fig. 8. Probability P, on an adjusted axis scale vs. b. Normal distribution describes a straight line.

The first (smallest) effect is assumed to represent $P(1/q \cdot 100)$% of the cumulative distribution function. The second effect represents $P(2/q \cdot 100)$%, and so on, until the last (largest) effect, which is assumed to represent $P(q/q \cdot 100)$%.

Divide the interval [0–100%] on the P-axis in q equally large intervals. Each interval will then be $100/q$. The smallest effect is then plotted in the middle point of the first interval. The second smallest effect is plotted in the middle point of the second interval, etc. This is summarised in a formula:

$$P(\%) = \frac{100\left(q - \frac{1}{2}\right)}{n}$$

where q is the ranking order of the effect and n is the number of parameters. If a normal distribution plot illustrates the distribution of the experimental noise, then the line should go through the co-ordinates (0, 50%). This is assumed, since the mean value of the experimental noise is normally distributed around zero.

5.4.1.1. Practical use of the normal distribution plot of effects. Normal probability plots of effects should only be consulted if the design is saturated, i.e., if the number of experiments is equal to the number of factors. Otherwise, if more experiments are performed than there are factors, the column plot of effects is preferable since the increased degrees of freedom allow for confidence intervals to be calculated.

A very practical use of normal probability plots of effects is to exclude one or some of the variables that have been identified in the plot as less influential. Thereby degrees of freedom are released and calculation of confidence intervals will be allowed, as was pointed out before. The continuing evaluation of the remaining variables will thus be a lot more satisfying, when having confidence intervals to consult for their significance.

Example 8: In Table 20, the parameters have been ordered according to increasing size. The table is then used to make the normal distribution plot (Fig. 9).

The three linear effects, b_2 and b_3, as well as the effect of the two-variable interaction, b_{13} and b_{23}, do not fit the straight line in Fig. 9. These effects have an influence on the result within the experimental domain. All other effects, belonging to the variables x_4 and x_5, fit the straight line well, which means that they do not influence the result (y_1) within the experimental domain. If additional experiments should be performed in order to optimize the result, only variables x_1, x_2 and x_3 should be of interest. The screening has thus lead to a focus on the most influential variables within the experimental domain.

Table 20
The effects ordered according to increasing size and the corresponding probability co-ordinate P

Ranking order	Estimate	Parameter	P (%)
1	−7.81	b_{23}	3.3
2	−6.46	b_{13}	10.0
3	−3.00	b_5	16.7
4	−2.56	b_{12}	23.3
5	−1.30	b_4	30.0
6	−0.95	b_{14}	36.7
7	−0.90	b_{24}	43.3
8	−0.73	b_{25}	50.0
9	−0.36	b_{45}	56.7
10	0.75	b_{15}	63.3
11	1.05	b_{35}	70.0
12	2.13	b_{34}	76.7
13	6.44	b_2	83.3
14	9.21	b_1	90.0
15	18.49	b_3	96.7

5.4.2. Evaluation of model fitness

5.4.2.1. Explained and predicted variation. In order to evaluate the fit of a model, values of explained variation, R^2, and predicted variation, Q^2, provide excellent guidance.

In proper terminology, the explained variation is the fraction of the total variation of the response that is explained in the model. It is calculated as

$$R^2 = (\mathrm{SS} - \mathrm{SS}_{\mathrm{resid}})/\mathrm{SS}$$

SS is the sum of squares of the total variation of a selected response, corrected for the mean. The total SS consists of two parts, one part resulting from the regression model ($\mathrm{SS}_{\mathrm{regr}}$) and another due to the residuals ($\mathrm{SS}_{\mathrm{resid}}$). Small residuals will render a high degree of explained variation.

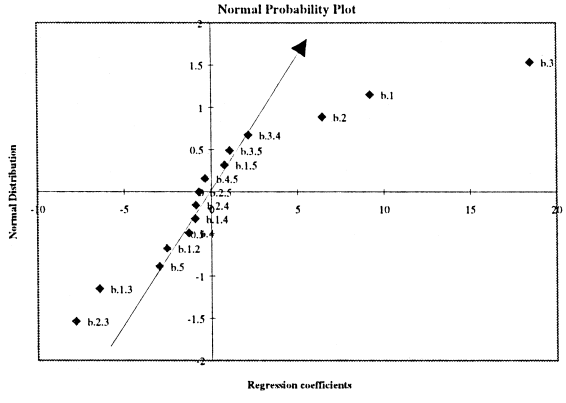


Fig. 9. Normal distribution plot of the parameters in Table 20.

Table 21
Guidelines on how to appreciate values of R^2 and Q^2

Nature of data	R^2	Q^2
Chemical	Acceptable: ≥ 0.8	Acceptable: ≥ 0.5 Excellent: > 0.8
Biological	Acceptable: > 0.7	Acceptable: > 0.4

The predicted variation, Q^2, is the fraction of the total variation of the response that can be predicted in the model. It is calculated as

$$Q^2 = (\mathrm{SS} - \mathrm{PRESS})/\mathrm{SS}$$

PRESS is the *p*rediction *r*esidual *s*um of *s*quares and is determined through cross-validation. [1] Small deviations between the actual residuals and the predicted ones will render a low PRESS and a high value of predicted variation.

Values of R^2 and Q^2 are usually between 0 and 1. Acceptable values are totally dependent on the nature of the data that are being examined. Table 21 offers some guidance.

5.4.2.2. ANOVA. In an *a*nalysis of *v*ariance, ANOVA, the total variation of the response is defined as a sum of two components; a regression component ($\mathrm{SS}_{\mathrm{regr}}$) and a component due to the residuals ($\mathrm{SS}_{\mathrm{resid}}$). The sum of squares of the total variation, corrected for the mean (SS), can thus be written as

$$\mathrm{SS} = \mathrm{SS}_{\mathrm{regr}} + \mathrm{SS}_{\mathrm{resid}}.$$

If there are replicates among the experiments, the residuals component are further divided into parts that are the sum of squares of lack of fit ($\mathrm{SS}_{\mathrm{lof}}$) and the sum of squares of pure experimental error ($\mathrm{SS}_{\mathrm{pe}}$):

$$\mathrm{SS}_{\mathrm{resid}} = \mathrm{SS}_{\mathrm{lof}} + \mathrm{SS}_{\mathrm{pe}},$$

In the statistical experimental design software Modde 4.0, [2] there are two plots that originate from analysis of variance. Those two plots, the ANOVA plot and the Lack of fit plot (Figs. 10 and 11), are very helpful in determining the fitness of a model.

In the ANOVA plot (Fig. 10), the regression component of the total variation is compared to the residual component. If the standard deviation of the response explained in the model (SD regression) is larger than the standard deviation of the residuals multiplied by the square root of the critical F ($\mathrm{RSD} * \sqrt{F_{\mathrm{crit}}}$), then the model is significant at the chosen probability level (usually $P = 0.05\%$).

In the plot of lack of fit (Fig. 11), the lack of fit component of the residuals is compared to the pure experimental error component. If the standard deviation of lack of fit (SD LoF) is larger than the standard deviation of the pure experimental error, multiplied by the square root of the critical F ($\mathrm{SD\ pe} * \sqrt{F_{\mathrm{crit}}}$), then the model suffers from a significant lack of fit.

The comparison of the bars in the ANOVA plots is similar to the ordinary variance ratio test, or F-test. [3] In an F-test, the variance ratio between lack of fit (lof) and pure experimental error (pe), is being compared to

[1] Cross validation means leaving out parts of the data during a series of model fitting. About one sixth of all values is left out at each round, and every value is left out only once. At every round, the missing values are predicted from the data that remain in the model. If the differences between the actual values and the predicted ones are small, then the predictive ability of the model is good and the value of PRESS is small.

[2] From Umetri, Umeå, Sweden.

[3] Morgan, E., Chemometrics: Experimental design, Wiley, Chichester, England (1995), or Box, G.E.P., Hunter, W.G., Hunter, J.S. in Section 8.

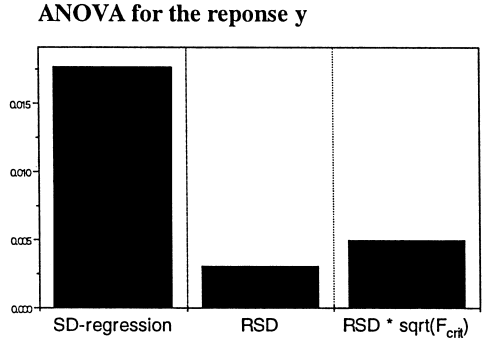


Fig. 10. ANOVA plot from Modde 4.0.

tabled values of F-distribution. If $F_{\mathrm{lof/pe}}$ exceeds the critical F, then there is a significant lack of fit at the probability level that is chosen (usually $P = 0.05$) and the model is incorrect. In the ANOVA plots in Modde, the variances are simply exchanged for the standard deviations (SD = √variance), and the comparison $\mathrm{SD_{lof}}/\mathrm{SD_{pe}}$ vs. the equally converted $\sqrt{F_{\mathrm{crit}}}$ has been rearranged to $\mathrm{SD_{lof}}$ vs. $\mathrm{SD_{pe}} * \sqrt{F_{\mathrm{crit}}}$.

5.4.3. Model refinement

Before moving on with interpreting a satisfactory model, refinement should always be attempted. Chances are that both values of explained and predicted variation will increase in the process. Refinement is primarily achieved through exclusion of the factors that are found to be insignificant in the coefficient plot.

In the coefficient plot from the software Modde 4.0, the influence of all factors on the response is displayed in bars with clearly stated confidence intervals (Fig. 12). Factors having small piles within the boundaries of the confidence interval have no significance, while factors with bigger piles are more influential. Deletion of the insignificant factors should always be made one at a time, since the deletion of one factor may influence the confidence intervals of the remaining factors.

Example 9: Refinement

An investigation covered five variables and ten cross terms. The explained variation and the predicted variation were satisfactory ($R^2 = 94.07\%$, $Q^2 = 62.14\%$). Since many of the cross terms were insignificant (Fig. 12), it was likely that the model would favour from exclusion of some of them.

One by one, the insignificant cross terms were deleted from the model. The predicted variation rose for every deletion, while the explained variation was fluctuating around its initial value. Finally, with the deletion of the

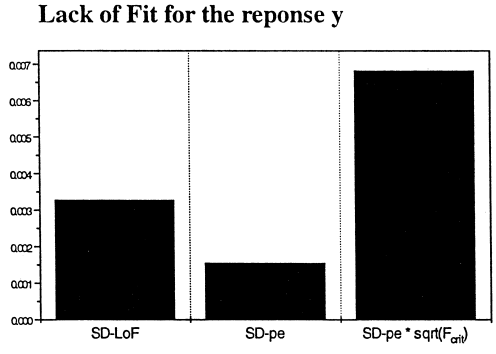


Fig. 11. Plot of lack of fit from Modde 4.0.

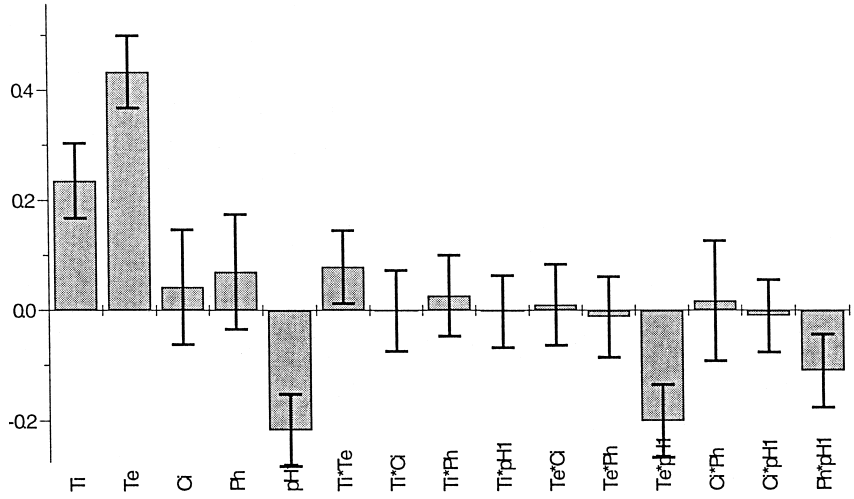


Fig. 12. Plot of scaled and centred coefficients for the response, from Modde 4.0.

seventh cross term, the explained and predicted variation suddenly decreased to very low values. Therefore, the last excluded cross term was reinserted and the model was saved.

In comparing the explained and predicted variation before and after the refinement (Figs. 13 and 14, respectively) the refinement seemed to have been successful. The explained variation increased only slightly, while the predicted variation rose from 62% to 83%.

Example 10: A 2^{5-2} fractional factorial design: Construction of experimental design

In a screening study, variables are to be investigated by a fractional factorial design.

Exercise

(a) Construct an experimental design for investigating five variables in a 2^{5-2} fractional factorial design. There is a number of possibilities.

(b) Find out how the main effects and the two-variable interaction effects are confounded in the design selected above (a).

(c) Make an additional design that will make it possible to separate the main effects from the two-variable interaction effects.

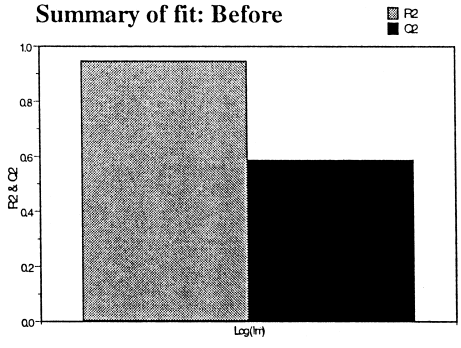


Fig. 13. Summary of fit plot from Modde 4.0, before the refinement.

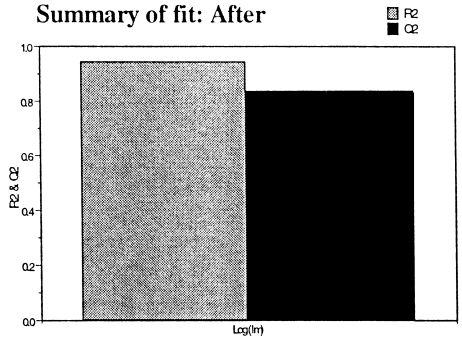


Fig. 14. Summary of fit plot from Modde 4.0, after the refinement.

Example 11: Fractional factorial design, 2^{5-1}: Pharmacy, pre-formulation

This example is taken from Leuenberger and Becher. [4] The aim of this study was to find the best suitable excipients from the point of view of stability. Therefore, the active substance was mixed with different excipients. The drug substance was analysed for degradation products (Table 22).

The response consisted of % intact drug substance as shown in Table 23.

Exercises

(a) Construct a fractional factorial design with 16 experiments. Use the unlikely significant interaction effect $x_1 x_2 x_3 x_4$ to define variable x_5.

(b) Estimate the effects (coefficients) of the experimental variables and evaluate their influence.

(c) Recommend an excipient mixture!

Example 12: A 2^{5-1} fractional factorial design: The Willgerodt–Kindler reaction (Fig. 15)

The example is taken from R. Carlson, T. Lundstedt and R. Shabana. [5] An organic reaction, the Willgerodt–Kindler reaction, was studied. Five experimental variables were investigated by a 2^{5-1} fractional factorial design.

The experimental variables, the design and the yields are given in Tables 24 and 25.

The experimental design is a 2^{5-1} fractional factorial design with the generator $I =$ **abcde**.

Exercise

Determine a second order interaction model in order to describe the yield as a function of the experimental variables. Identify probably significant variables by using a normal distribution plot and refine the model.

6. Optimization

In this part, two different strategies for optimization will be introduced; simplex optimization and response surface methodology. An exact optimum can only be determined by response surface methodology, while the simplex method will encircle the optimum.

[4] Leuenberger, H., Becher, W., A factorial design for compatibility studies in pre-formulation work, Pharm. Acta Helv, 50, 88–91 (1975).

[5] Carlson, R., Lundstedt, T., Shabana, R., Acta Chem. Scand., B 40, 534 (1986). This example is also included as an exercise in several software packages.

Table 22
Variables and experimental domain

Variables	Experimental domain	
	(−)-level	(+)-level
x_1: Filler	lactose	mannitol
x_2: Lubricant	stearic acid	magnesium stearate
x_3: Disintegrant	maize starch	microcrystalline cellulose (Avicel®)
x_4: Binder	polyvinylpyrrolidone (PVP)	gelatine
x_5: Humidity	no water added	water added

Table 23
Percentage of resulting intact drug substance after each experiment

Exp. no.	Response = y (%)	Exp. no.	Response = y (%)
1	59.6	9	54.1
2	86.4	10	45.8
3	95.0	11	92.8
4	97.0	12	96.1
5	83.4	13	53.6
6	53.8	14	64.7
7	93.7	15	94.0
8	99.7	16	96.3

6.1. Simplex optimization

A simplex is a geometric figure with $(k + 1)$ corners where k is equal to the number of variables in a k-dimensional experimental domain. When the number of variables is equal to two the simplex is a triangle (Fig. 16).

Simplex optimization is a stepwise strategy. This means that the experiments are performed one by one. The exception is the starting simplex in which all experiments can be run in parallel.

The principles for a simplex optimization are illustrated in Fig. 17. To maximise the yield in a chemical synthesis, for example, the first step is to run $k + 1$ experiments to obtain the starting simplex. The yield in each corner of the simplex is analysed and the corner showing the least desirable result is mirrored through the geometrical midpoint of the other corners. In this way, a new simplex is obtained. The co-ordinates (i.e., the experimental settings) for the new corner are calculated and the experiment is performed. When the yield is determined, the worst of the three corners is mirrored in the same way as earlier and another new simplex is obtained, etc. In this way, the optimization continues until the simplex has rotated and the optimum is encircled. A

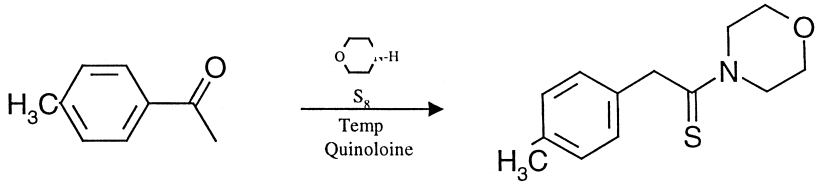


Fig. 15. The Willgerodt–Kindler reaction formula.

Table 24
Variables and experimental domain in the screening of the Willgerodt–Kindler reaction

Variables	Experimental domain	
	(−)-level	(+)-level
x_1: Ratio sulphur/ketone (mol/mol)	5	11
x_2: Ratio amine/ketone (mol/mol)	6	10
x_3: Reaction temperature (°C)	100	140
x_4: Particle size of sulphur (mesh)	240	120
x_5: Agitation speed (rpm)	300	700

fully rotated simplex can be used to calculate a response surface. The type of design described by a rotated simplex is called a Doehlert design.

6.1.1. Rules for a simplex optimization

With k variables $k+1$ experiments are performed with the variable settings determined by the co-ordinates in the simplex. For two variables the simplex forms a triangle. For three variables it is recommended to use a 2^{3-1} fractional factorial design as a start simplex.

- *Rule 1*: Reflect the co-ordinates for the lowest-achieving corner in the line/plane described by the rest of the corners and perform a new experiment by using the co-ordinates as variables settings for this experiment. A new simplex is obtained consisting of the remaining k corners together with the new one. Continue in this way until the response does not improve.
- *Rule 2*: If the new experiment is the one with the poorest result of the three corners, then according to the first rule the new experiment should be performed with the same settings as the worst point in the previous simplex. In this case the second worst point should be mirrored in the geometrical midpoint of the other corners.
- *Rule 3*: If a reflection gives a new experiment to be performed outside the possible experimental domain, then this point should be regarded as the lowest-achieving and rule 2 is used.

Table 25
Experimental design and yields

Exp. no.	Variables					Yield (%)
	x_1	x_2	x_3	x_4	x_5	y
1	−	−	−	−	+	11.5
2	+	−	−	−	−	55.8
3	−	+	−	−	−	55.8
4	+	+	−	−	+	75.1
5	−	−	+	−	−	78.1
6	+	−	+	−	+	88.9
7	−	+	+	−	+	77.6
8	+	+	+	−	−	84.5
9	−	−	−	+	−	16.5
10	+	−	−	+	+	43.7
11	−	+	−	+	+	38.0
12	+	+	−	+	−	72.6
13	−	−	+	+	+	79.5
14	+	−	+	+	−	91.4
15	−	+	+	+	−	86.2
16	+	+	+	+	+	78.6

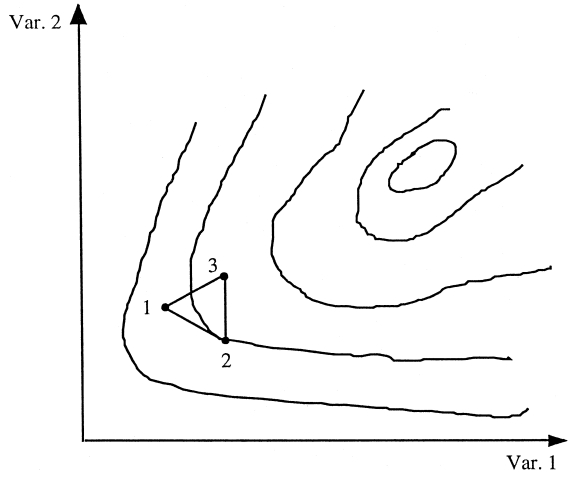


Fig. 16. A simplex in two variables.

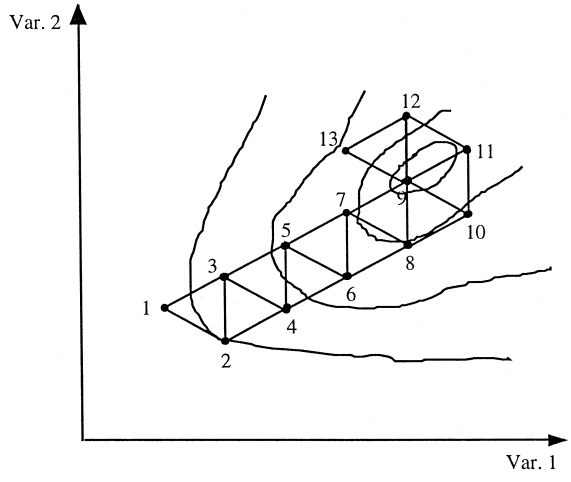


Fig. 17. Illustration of a simplex optimization with two variables.

6.1.2. Calculation of the experimental settings for a new experiment in the simplex

When $k+1$ experiments in the start simplex have been performed, the responses in the different corners are examined. In a simplex with two variables, the corner with the least desirable result is marked S, the best corner with B and the second best with N. M is the geometrical midpoint in the hyper plane that is spanned by the remaining corners when S is removed, in this case a line between B and N (Fig. 18). T is the new corner that substitutes S in the continued optimization.

In a simplified way the mirroring can be described as follows:

midpoint	$M = (N + B)/(k)$
distance	$d = M - S$
the new point	$T = M + d$

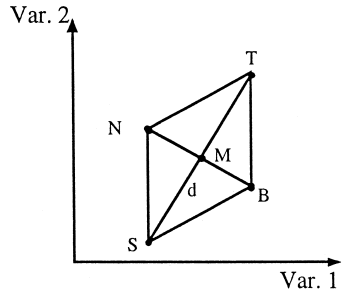


Fig. 18. Projection of the corner *S*, with poor result, to the new corner *T* in the simplex.

This gives

$$T = M + M - S = \frac{2(N + B)}{k - S}.$$

For k variables, the co-ordinates x_{iT} of the new corner T are calculated as:

$$X_{iT} = 2/K\left(\sum_{\substack{\text{all } j \\ \text{except } S}} X_{ij}\right) - x_{iS}$$

6.2. Response surface methodology

Response surfaces are used to determine an optimum. In addition, it is a good way to graphically illustrate the relation between different experimental variables and the responses. To be able to determine an optimum it is necessary that the polynomial function contains quadratic terms.

In the following pages, two types of designs will be presented that are employed to fit experimental data to a quadratic model.

$$y = \beta_0 + \sum_{1}^{k} \beta_i x_i + \sum_{1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \Sigma \beta_{ij} x_i x_j + \varepsilon$$

6.2.1. Doehlert design

When a simplex optimization with two variables comes to the point where it encircles the optimum, a hexagon is formed. Such a design is called a Doehlert design and allows the calculation of a response surface by a minimum of experiments. Another attractive feature with this design is that a neighbouring domain is easily explored by just adding a few experiments (Fig. 19).

The design matrices for Doehlert designs with two and three variables are given in Tables 26 and 27, respectively.

6.2.2. Central composite design

A full central composite design consists of the following parts:

(a) A full factorial or fractional factorial design.

(b) Experiments at the centre, i.e., $x_i = 0$ for all i.

(c) Experiments where $x_i = \pm\alpha$ and with $x_j \neq x_i = 0$. These points are situated on the axis in a co-ordinate system and with distance $\pm\alpha$ from the origin; they are *axial points*.

If the experiments are illustrated as points in a co-ordinate system, defined by the xi-axes, then the designs (for two and three variables) can be graphically illustrated as in Figs. 20 and 21.

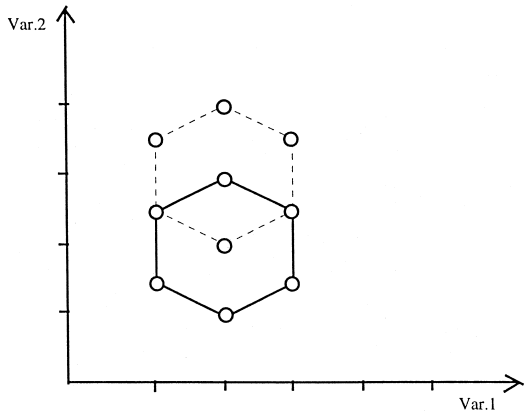


Fig. 19. A Doehlert design with two variables. The dotted design illustrates the exploration of a neighbouring domain by adding three experiments.

The design matrices for two and three variables are given in Table 28.

The value of α is varying with the number of variables. Values for up to six variables are given in Table 29.

Example 13: Simplex optimization

A reaction influenced by two experimental variables, pH and temperature, was studied. In Table 30, the first three experiments define the starting simplex.

Exercise

(a) Calculate the experimental settings for the next experiment and use the yield given in the table (experiment number four) in order to identify the new simplex. With two variables the experiments can be graphically examined on a graph paper.

(b) Use the seven experiments that encircle the optimum (this constitutes a Doehlert design) to compute a response surface model and to determine the optimal conditions for the reaction.

Example 14: Response surface: the Willgerodt–Kindler reaction (Fig. 22)

In example 11, the screening of important variables for the optimization of the Willgerodt–Kindler reaction was evaluated. The important variables identified were amount of sulphur, amount of amine and the reaction temperature. In order to find the optimal conditions for the reaction a response surface model was established.

The experimental domain, the design matrix and the yields are given in Tables 31 and 32.*Exercise*

Use a computer program to fit the experimental data to a quadratic model.

Table 26
A Doehlert design for two variables

Exp. no.	x_1	x_2
1	0.000	0.000
2	−1.000	0.000
3	−0.500	−0.866
4	1.000	0.000
5	0.500	0.866
6	−0.500	0.866
7	0.500	−0.866

Table 27
A Doehlert design for three variables

Exp. no.	x_1	x_2	x_3
1	0.0000	0.0000	0.0000
2	1.0000	0.0000	0.0000
3	0.5000	0.8660	0.0000
4	0.5000	0.2887	0.8165
5	−1.0000	0.0000	0.0000
6	−0.5000	−0.8660	0.0000
7	−0.5000	−0.2887	−0.8165
8	0.5000	−0.8660	0.0000
9	0.5000	−0.2887	−0.8165
10	0.0000	0.5774	−0.8165
11	−0.5000	0.8660	0.0000
12	−0.5000	0.2887	0.8165
13	0.0000	−0.5774	0.8165

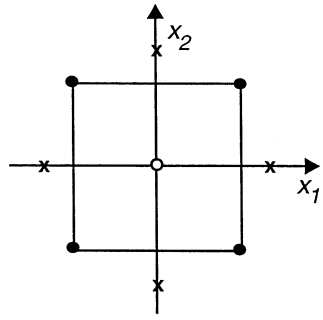


Fig. 20. Central composite designs for two variables. The different markings mean (●) factorial design, (○) centre point and (×) axial points.

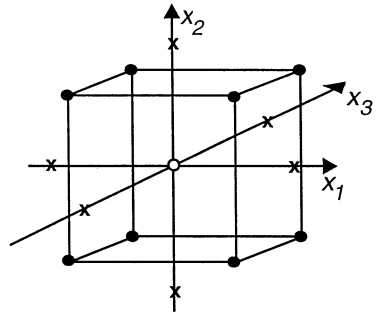


Fig. 21. Central composite designs for three variables. Markings as in Fig. 20.

Table 28
Central composite design matrices for two (left) and three (right) variables

Two variables			Three variables			
x_1	x_2		x_1	x_2	x_3	
−1	−1	Factorial design	−1	−1	−1	Factorial design
1	−1		1	−1	−1	
−1	1		−1	1	−1	
1	1		1	1	−1	
			−1	−1	1	
0	0	Centre points	1	−1	1	
			−1	1	1	
$-\alpha$	0	Axial points	1	1	1	
α	0		0	0	0	Centre point
0	$-\alpha$					
0	α		$-\alpha$	0	0	Axial points
			α	0	0	
			0	$-\alpha$	0	
			0	α	0	
			0	0	$-\alpha$	
			0	0	α	

Table 29
Central composite designs

Number of variables	2	3	4	5 2^{5-1}	5	6	6 2^{6-1}
Number of experiments in the factorial design	4	8	16	32	16	64	32
Number of axial points	4	6	8	10	10	12	12
Value of α	1.414	1.682	2.000	2.378	2.000	2.828	2.378

Table 30
Experimental variables and response

Exp. no.	pH	Temp.	Yield (%)
1	6.90	25	29
2	7.05	26	38
3	6.95	28	41
4	7.10	29	68
5	7.00	31	45
6	7.15	32	56
7	7.25	30	63
8	7.20	27	41

Evaluate and refine the model and determine the optimal conditions. Make projections of the response surface.

7. Mixture designs

In a mixture experiment, it is not the actual amount of the single ingredient that matters, but rather its proportion in relation to other ingredients. The sum of all the ingredients is a constant total T, which is equal to 100% or 1 (unless any constant mixture factors are present). The constant total T represents a constraint on mixture

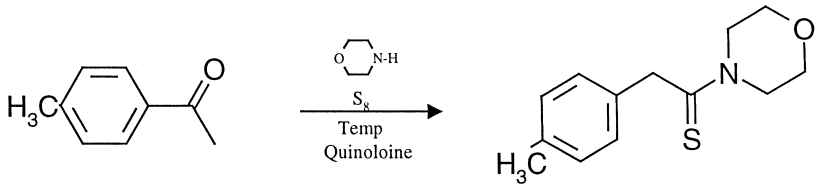


Fig. 22. The Willgerodt–Kindler reaction formula.

experiments that implies independence between all mixture factors. This independence in turn affects the experimental region and brings particular demands on the model making, which calls for mixture designs.

7.1. Factors in mixture experiments

7.1.1. Mixture factors

The mixture factors, or the formulation factors, are the ingredients that are believed to have effect on the outcome of the formulation and whose proportions are to be varied in the mixture experiment.

7.1.2. Filler

Some mixture experiments have mixture factors only, while in others the presence of a filler is mandatory. When baking, for example, yeast and sugar would render little result without the presence of flour. It may however not always be as clear-cut to define the filler among many ingredients. The following pointers may be helpful.

- The filler has no effect of its own that is of interest for the outcome of the formulation.
- It is always present in the mixture.
- It accounts for a large percentage of the mixture.
- It is added at the end to bring the mixture total to the desired amount, the constant total T.

Only one ingredient can be defined as a filler. When you specify a filler, the conditions that are to be met are checked in Modde. Then, by default, a design is created in which the amount of the filler considered but the filler itself is in other regards omitted.

7.1.3. Constant mixture factors

All mixture factors are either controlled or constant. Mixture factors are defined as constant when they are kept unchanged in the experiment. When constant mixture factors are present, the mixture total T is calculated as

$$T = 1 - \Sigma(\text{constant mixture factors})$$

instead of adding up to 1.

In Modde, an error message is issued whenever the mixture total is not equal to T or 1.

A filler can never be a constant mixture factor, as was pointed out in Section 7.1.2.

Table 31
Variables and experimental domain in the optimization of the Willgerodt–Kindler reaction

Variables	Experimental domain				
	−1.682	−1	0	1	1.682
x_1: Sulphur/ketone (mol/mol)	2.95	5.0	8.0	11.0	13.05
x_2: Amine/ketone (mol/mol)	4.63	6.0	8.0	10.0	11.37
x_3: Reaction temperature (°C)	86	100	120	140	154

Table 32
The design matrix and yields

Exp. no.	x_1	x_2	x_3	Yield (%)
1	−1	−1	−1	11.5
2	1	−1	−1	43.7
3	−1	1	−1	38.0
4	1	1	−1	75.1
5	−1	−1	1	79.5
6	1	−1	1	88.9
7	−1	1	1	77.6
8	1	1	1	78.6
9	−1.682	0	0	48.5
10	1.682	0	0	91.5
11	0	−1.682	0	58.8
12	0	1.682	0	94.7
13	0	0	−1.682	14.4
14	0	0	1.682	94.1
15	0	0	0	83.9
16	0	0	0	84.2
17	0	0	0	85.6
18	0	0	0	82.6
19	0	0	0	83.2
20	0	0	0	84.9

7.1.4. Process factors

Process factors are the experimental parameters that are not part of the actual mixture, such as temperature or pH, for example. The process factors are expressed either in a quantitative or a qualitative form. Process factors can be incorporated together with mixture factors in the same designs in Modde. [6]

7.2. Scaling

When the model is fitted with *p*artial least squares projections to *l*atent *s*tructures (PLS), factors are scaled to unit variance. Mixture factors will not be scaled when the model is fitted with *m*ultiple *l*inear *r*egression (MLR).

With PLS, the dependence between factors will be taken into account when the model is fitted. Therefore, it is preferable to MLR when mixture experiments are concerned.

7.3. The experimental region

When all mixture factors vary from 0 to T (the mixture total), the shape of the experimental region is a simplex. [7] In some cases, though, factors may be constrained in ways that prohibit values beyond certain limits. Then, some of the experimental region has been cut off and cannot be explored.

[6] A quantitative factor is measured on a continuous scale, while a qualitative factor only has discrete values, such as 'off or on' and 'with or without'. In Modde, the discrete values have to be converted to numerical values.

[7] A simplex is a geometric figure with $n+1$ corners in an n-diemsional space. A regular simplex in two dimensions ($n=2$) is an equilateral triangle. In three dimensions ($n=3$), it is a tetrahedron.

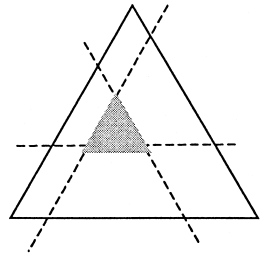


Fig. 23. Example of a restricted experimental region, resulting form lower limits in all three factors. The shape of the shadowed area is a simplex, inside the original simplex.

When all factors have only lower or only upper limits, the restricted experimental region is a small simplex inside the original simplex (Fig. 23). When factors have different limitations, the restricted experimental region becomes an irregular polyhedron inside the simplex (Fig. 24).

In order to establish the form of the experimental region in each case, the ranges of the experimental region is calculated in Modde from the range of each factor as:

$$R_U = \Sigma U_i - T$$

and

$$R_L = T - \Sigma L_{i.}$$

R_L and R_U are the lower and upper ranges for the restricted experimental region. L_i and U_i are the lower and upper limit of the ith mixture factor. The lower limit simplex is orientated as in Fig. 23, while the upper limit simplex has the opposite orientation.

7.4. Pseudo component transformations

When the experimental region is a lower or upper limit simplex, the mixture factors are by default transformed into pseudo components in Modde. Their ranges will thereby be made to vary between 0 and 1. The analysis is performed on the pseudo components, which allows for the coefficients to be directly interpreted as

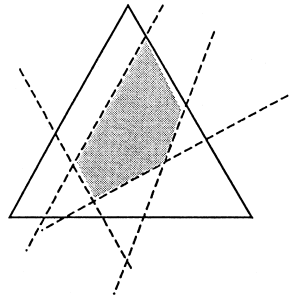


Fig. 24. Example of a restricted experimental region, resulting from different limits in all three factors. The shape of the shadowed area is an irregular polyhedron.

mixture factor effects (unless the Sheffé MLR fit method is used [8]). The worksheet is however always displayed in original units.

When the shape of the experimental region is an irregular polyhedron, pseudo component transformation is less meaningful, and is therefore not available in Modde.

7.5. Choice of design

When all factors are mixture factors and the shape of the region is a simplex, there are several classical mixture designs available in Modde. It is however always recommended that D-optimal designs be used for mixtures, since it is the only design in which the dependence between factors will be considered. In all other classical designs the factors are assumed to be independent.

Example 15: The following example is taken from a Modde 4.0 tutorial. [9] The data are available in Modde in example 'Snee8'.

A new product was made from eight ingredients ($x_1 - x_8$). None of them was defined as a filler, so they were all formulation factors. There was one response, y, for which high values would indicate good quality of the product. The design of the screening was chosen to be D-optimal with PLS as fit method. From the resulting model, a series of plots could be produced in Modde.

Exercises

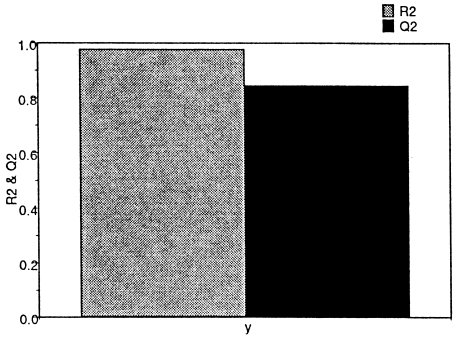
Try to find answers to the questions below with the aid of the following plots. Guidelines to the interpretation are added in the end of this exercise.

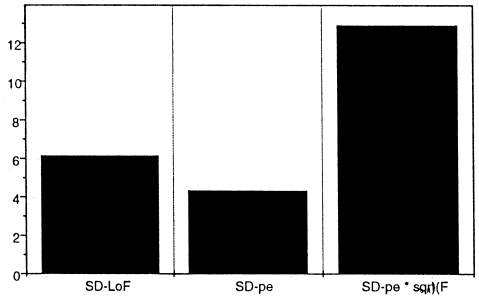
Analysis of the quality of the model:
- Would you say that the explained variation was satisfying?
- Did the predicted variation reach an acceptable level, in your opinion?
- Did the model suffer from lack of fit?
- Were there any severe outliers among the data?

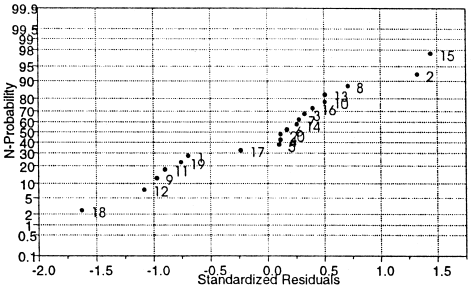
Analysis of the formulation results:
- Would you say that the response was well correlated with the mixture factors?
- Which mixture factors were the most influential on the response?
- What recommendations would you give for future formulations? Note that all variables have to be present in the mixture.

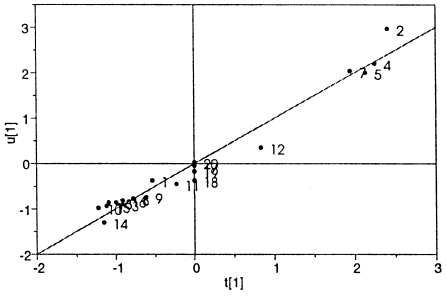
[8] Scheffé MLR is one of three fit methods available in Modde 4.0. The other two are multiple linear regression (MLR) and partial least squares (PLS).

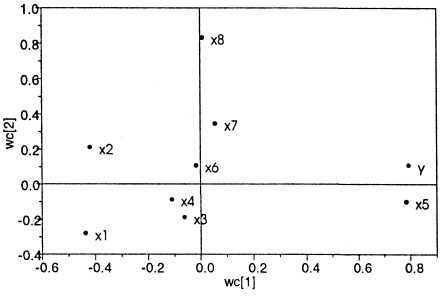
[9] Issued by Umetri, Umeå, Sweden (unknown version).

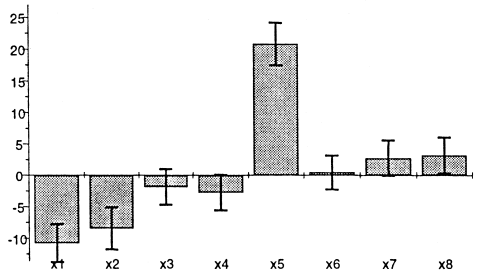


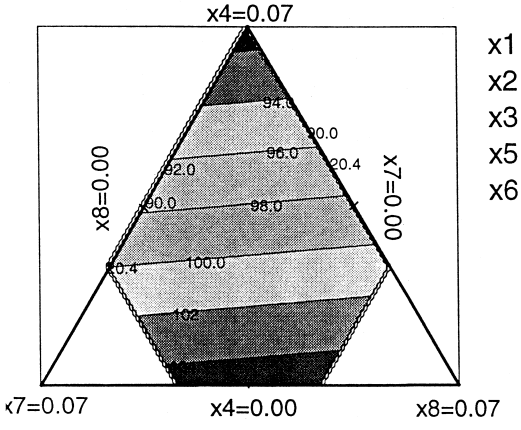












x1 = lowest level
x2 = lowest level
x3 = mid level
x5 = highest level
x6 = mid level

Interpretation guidelines

Plot 1. Summary of fit: the levels of explained and predicted variation were excellent ($R^2 = 97\%$ and $Q^2 = 84\%$).

Plot 2. ANOVA: in the plot of lack of fit, the standard deviation of lack of fit was smaller than the standard deviation of the pure error multiplied by the square root of the critical F. Hence, there was no lack of fit.

Plot 3. Residuals N-plot: the residuals were normally distributed without outliers.

Plot 4. PLS plot I: in the score plot, the formulation factors and the response showed correlation to a satisfying degree.

Plot 5. PLS plot II: in the loading plot, formulation factor x_5 turned out to be positively correlated to the response y. Other important factors were x_1 and x_2, although in negative correlation to y.

Plot 6. Coefficient plot: the conclusions drawn from the loading plot were confirmed. Increased additions of x_1 and x_2 influenced the outcome in a negative direction. Additions of x_5 had a positive impact. Other factors had insignificant effects.

Plot 7. Mixture contour plot: the desirable settings of the significant factors x_1, x_2 and x_5 were unquestionable. The settings of the less influential factors, however, needed to be established as well, since their presence in the mixture was required. The variables x_4, x_7 and x_8 were put on the three axes of the mixture contour plot. The other factors were set at their lowest (x_1, x_2), their highest (x_5) and mid level values (x_3, x_6).

Apparently, low amounts of x_4 would serve the purpose of this mixture. The amounts of x_7 and x_8 have little impact on the result.

8. Additional references

8.1. Books

G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for experimenters. Wiley, New York, 1978.
R. Carlson, Design and Optimization in Organic Synthesis. Elsevier, Amsterdam, 1992.
S.N. Deming, S.L. Morgan, Experimental design: a chemometric approach. Elsevier, Amsterdam, 1993.
R.M. Myers, Response surface methodology. Allyn and Bakon, Boston, 1971.

8.2. Papers

D.H. Doehlert, Uniform shell designs. Applied Statistics, 19 (1970) 231–239.

R. Carlson, T. Lundstedt, R. Shabana, Optimum conditions for the Willgerodt–Kindler reaction 1: reaction of substituted acetophenones. Prediction of optimum conditions for new substrates by multivariate correlation, Acta Chemica Scandinavica, B 40 (1986) 534.

Ø. Langsrud, M. Risberg Ellekjaer, T. Naes, Identifying significant effects in fractional factorial experiments, Journal of Chemometrics, 8 (1994) 205–219.

T. Lundstedt, P. Thorén, R. Carlson, Synthesis of 4-(*N*,*N*-dimethylamino)acetophenone optimized by a Doehlert design, Acta Chemica Scandinavica, B 38 (1984) 717–719.

W. Spendley, G.R. Hext, F.R. Himsworth, Sequential application of simplex designs in optimization and evolutionary operation, Technometrics, 4 (1962) 441.

R. Sundberg, Interpretation of unreplicated two-level factorial experiments, by examples (Tutorial), Chemometrics and intelligent laboratory system, 24 (1994) 1–17.