Chemometric analysis of the reaction systems

TATJANA N. MOSORINAC
Oil Refinery Pancevo, Spoljnostiarsceva bb, 13000 Pancevo
SERBIA
tanjadexter@yahoo.com

Abstract: This paper presents an effective approach in study and investigation of the chemical reaction system. Experimental design matrix for active experiment was used for the new product formulation. Regression and correlation analysis give correct curves as to the relationships between process variables and optimum conditions. The derived mathematical model presents a response function which connects the optimization parameter and variables. The optimal experimental design and correlation analysis and synthesis has leaded to development advanced investigation method.

Key-Words: Design matrix, factor variable, full factorial design, regression equation, interaction coefficients, correlation analysis.

1 Introduction

In cases where lack accurate information about a system or a system is complex to a point where a deterministic model is out of the question, resort is made to experimentation and statistics. In such cases, the system in question is treated as a black box or gray box, and its input-output relationships are studied through an experiment, which may be passive or active.

A passive experiment is traditional variety in which a large number of observations are made, each variable being varied in turn. The concept of passive experiment also applies to the collection of statistical data in the normal operation of an industrial plant or processes. The data thus collected are then reduced to develop a mathematical model by the classical methods of regression and correlation analysis.

An active experiment is carried out to a predetermined the design, the design of experiments and is usually arranged so as to cover all the governing system variables simultaneously, thereby permitting the experimenter to assess the strength of interaction between the variables and to cut down the amount of experimentation.

2 Regression and correlation analysis

Regression and correlation analysis, when used judiciously, can give correct curves as to the relationships between process variables and optimum conditions.

In the both cases, the mathematical model is a response function which connects the optimization parameter which describes the outcome of an experiment to the variables which the experimenter manipulates in conducting the experiment:

\[ y = \varphi(x_1, x_2, x_3, \ldots, x_k) \]  

The independent variables \( x_1, x_2, x_3, \ldots, x_k \) are usually called factors, the coordinate space with coordinates \( x_1, x_2, x_3, \ldots, x_k \) is called the factorial space, and a geometrical portrait of response functions in the factorial space is called response surface.

If experiments were conducted and their outcomes analyzed by conventional methods, that is, by varying only one variable at a time while holding the others constant, there would be considerable probability of running into a false optimum.

With statistical methods, a mathematical model

\[ y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + k \sum_{u,j=1}^{k} \beta_{uj} x_u x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 \]  

where \( \beta_j = \frac{\partial \varphi}{\partial x_j} \bigg|_{x=0}, \beta_{uj} = \frac{\partial^2 \varphi}{\partial x_u \partial x_j} \bigg|_{x=0} \) and \( \beta_{jj} = \frac{\partial^2 \varphi}{\partial x_j^2} \bigg|_{x=0} \).

Since in real system there are always uncontrolled and unobservable variables, changes in \( y \) are random in character. This is why in processing experimental data one uses sample regression coefficients \( b_0, b_1, b_2, b_{uj}, b_{jj} \) which are estimates of the theoretical coefficients \( \beta_0, \beta_1, \beta_2 \ldots \beta_{uj}, \beta_{jj} \).
\[ y = b_0 + \sum_{j=1}^{k} b_j x_j + k \sum_{a \neq j} b_{aj} x_a x_j + \sum_{j=1}^{k} b_{jj} x_j^2 \quad (3) \]

The coefficient \( b_0 \) is the free term of the regression equation, the coefficients, \( b_j \) are linear terms, the coefficients \( b_{aj} \) are quadratic terms and the coefficients \( b_{jj} \) are interaction terms.

The coefficients are estimated by the method of the least squares from the condition

\[ \Phi = \sum_{i=1}^{N} (y_i^{cal} - y_i^{exp}) = \text{min} \quad (4) \]

where \( N \) is the size of a sample out of the entire set of values of the parameters in question. The difference between the sample size \( N \) and the number of constraints \( l \), imposed on the sample, is the number of values of the parameters in question. The difference when the function is differentiable and \( b_0, b_1, b_2, \ldots \) should be selected so that

\[ \Phi = \sum_{i=1}^{N} [y_i - f(x_i, b_0, b_1, b_2, \ldots)]^2 = \text{min} \quad (9) \]

the necessary condition for minimizing \( \Phi(b_0, b_1, b_2, \ldots) \) is to satisfy the equalities

\[ \frac{\partial \Phi}{\partial b_0} = 0, \quad \frac{\partial \Phi}{\partial b_1} = 0, \quad \frac{\partial \Phi}{\partial b_2} = 0, \ldots \quad (10) \]

or

\[ \sum_{i=1}^{N} 2[y_i - f(x_i, b_0, b_1, b_2, \ldots)] \frac{\partial f(x_i)}{\partial b_0} = 0 \]

\[ \sum_{i=1}^{N} 2[y_i - f(x_i, b_0, b_1, b_2, \ldots)] \frac{\partial f(x_i)}{\partial b_1} = 0 \]

after some manipulation, is obtained

\[ \sum_{i=1}^{N} y_i \frac{\partial f(x_i)}{\partial b_0} - \sum_{i=1}^{N} f(x_i, b_0, b_1, b_2, \ldots) \frac{\partial f(x_i)}{\partial b_0} = 0 \quad (12) \]

\[ \sum_{i=1}^{N} y_i \frac{\partial f(x_i)}{\partial b_1} - \sum_{i=1}^{N} f(x_i, b_0, b_1, b_2, \ldots) \frac{\partial f(x_i)}{\partial b_1} = 0 \]

The system Eqs.(12) should have as many equations as there are unknown coefficients \( b_0, b_1, b_2, \ldots \) in the estimated equation of regression. In mathematical statistics it is called a system of normal equations. At any values of \( b_0, b_1, b_2, \ldots \), the quantity \( \Phi \geq 0 \). Therefore, it must have at least one minimum. In other words, if the system of normal equations has a unique solution, this solution is the minimum for \( \Phi \).

### 3 Method of optimal experimental design

One of the most important aspects of empirical model building is the design of experiments, that is, the selection of strategy to obtain an adequate model with a minimum of experimentation.

There are many methods of design experiments such as factorial experimental design, fractional replicate, orthogonal second order design, rotatable second order

\[ f = N - l \quad (5) \]

In developing an equation of regression it is assumed that the number of constraints is equal to the number of coefficients sought.

The number of coefficient needed to be estimated in order to develop equations of different orders for various independent variables. The number of coefficients to be estimated rapidly increases with both the number factors and the order of the equation.

The order of the equation is found by finite difference method and the form of an equation of regression is found by trial and error method.

In establishing the average relationship between two variables associated with the same experiment, the form of the equation of regression can be deduced from the shape of an empirical line of regression. For this purpose, the entire range of changes in the independent variable \( x \) on the correlation field is broken down into equal intervals \( \Delta x \).

All points falling within a given interval \( \Delta x_j \) are referred to its centre \( x_j \). It is done by finding the sample mean

\[ y_j = \frac{\sum_{i=1}^{n_j} x_{ji}}{n_j} \quad (6) \]

where \( n_j \) is the number of points within a given sample interval \( \Delta x_j \).

\[ \sum_{j=1}^{k} n_j = N \quad (7) \]

Where \( k \) is the number of sample intervals and \( N \) is the sample size.

The estimation of the parameters in an estimated equation of regression reduces to minimizing a function of many variables. If

\[ y = f(x, b_0, b_1, b_2, \ldots) \quad (8) \]

that is, if the function is differentiable and \( b_0, b_1, b_2, \ldots \) should be selected so that

\[ \Phi = \sum_{i=1}^{N} [y_i - f(x_i, b_0, b_1, b_2, \ldots)]^2 = \text{min} \quad (9) \]

the necessary condition for minimizing \( \Phi(b_0, b_1, b_2, \ldots) \) is to satisfy the equalities

\[ \frac{\partial \Phi}{\partial b_0} = 0, \quad \frac{\partial \Phi}{\partial b_1} = 0, \quad \frac{\partial \Phi}{\partial b_2} = 0, \ldots \quad (10) \]

or

\[ \sum_{i=1}^{N} 2[y_i - f(x_i, b_0, b_1, b_2, \ldots)] \frac{\partial f(x_i)}{\partial b_0} = 0 \]

\[ \sum_{i=1}^{N} 2[y_i - f(x_i, b_0, b_1, b_2, \ldots)] \frac{\partial f(x_i)}{\partial b_1} = 0 \]

after some manipulation, is obtained

\[ \sum_{i=1}^{N} y_i \frac{\partial f(x_i)}{\partial b_0} - \sum_{i=1}^{N} f(x_i, b_0, b_1, b_2, \ldots) \frac{\partial f(x_i)}{\partial b_0} = 0 \quad (12) \]

\[ \sum_{i=1}^{N} y_i \frac{\partial f(x_i)}{\partial b_1} - \sum_{i=1}^{N} f(x_i, b_0, b_1, b_2, \ldots) \frac{\partial f(x_i)}{\partial b_1} = 0 \]

The system Eqs.(12) should have as many equations as there are unknown coefficients \( b_0, b_1, b_2, \ldots \) in the estimated equation of regression. In mathematical statistics it is called a system of normal equations. At any values of \( b_0, b_1, b_2, \ldots \), the quantity \( \Phi \geq 0 \). Therefore, it must have at least one minimum. In other words, if the system of normal equations has a unique solution, this solution is the minimum for \( \Phi \).

### 3 Method of optimal experimental design

One of the most important aspects of empirical model building is the design of experiments, that is, the selection of strategy to obtain an adequate model with a minimum of experimentation.

There are many methods of design experiments such as factorial experimental design, fractional replicate, orthogonal second order design, rotatable second order
design, Latin squares, Graeco-Latin squares, Youden squares, simplex design and EVOP-evolutionary operation design.

In factorial experiment, a certain number of fixed values, or levels, are assigned each factor or each experimental variable which is deliberately varied from experiment to experiment. If all possible combination of the factors are used, the experiment is called a full factorial design. If k variables are to be controlled at two level each, the experiment is termed a full two level factorial experiment for k factors or $2^k$ design.

The factor levels are in effect the boundaries of the area to be searched for a given variable. Let one be interested in the effect of three factors on the product yield $y$. If the upper level for the first variable is $z_1^{max}$ and the lower level $z_1^{min}$, then $z_1^0$ and $\Delta z_1$ defined by

$$z_1^0 = \frac{z_1^{max} + z_1^{min}}{2}$$
$$\Delta z_1 = \frac{z_1^{max} - z_1^{min}}{2}$$

Generally, for any factor $z_j$

$$z_j^0 = \frac{z_j^{max} + z_j^{min}}{2}$$
$$\Delta z_j = \frac{z_j^{max} - z_j^{min}}{2}$$

The point with coordinates $(z_1^0, z_2^0, ..., z_k^0)$ is called the centre point of the design, or the basic level, $\Delta z_j$ is the change interval or unit on the $z_j$ - axis. It is usually to pass from the $z_1, z_2, ..., z_k$ coordinates to a new dimensionless coordinates $x_1, x_2, ..., x_k$ by the coding equation

$$x_j = \frac{z_j - z_j^0}{\Delta z_j} \quad j=1,2,3,...,k$$

In the dimensionless coordinate system, the upper level is +1, the lower level is -1, the coordinates of the centre point of the design are zero and coincide with the origin of coordinates. In this problem k=3. The number of all possible combinations N of three factors each controlled at two levels is $N=2^k = 2^3 = 8$. The experimental design, the design matrix can reduced in Table 1. Geometrically, the coded design of Table 1 may be interpreted as a cube, in which the eight corners represent the eight experimental points.

Let write the $2^3$ design matrix and the experimental results by introducing a column for a dummy variable $x_0$ =1.

The design matrix presented in Table 1 has the following properties:

$$\sum_{j=1}^{N} x_{ui} x_j = 0 \quad u \neq j \quad u, j = 0,1,...,k \quad (16)$$

$$\sum_{j=1}^{N} x_{ji} = 0 \quad j=1,2,...,k \quad j \neq 0 \quad (17)$$

Table 1. Design matrix with dummy variable

<table>
<thead>
<tr>
<th>N</th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>$y_1$</td>
</tr>
<tr>
<td>2</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>$y_2$</td>
</tr>
<tr>
<td>3</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>$y_3$</td>
</tr>
<tr>
<td>4</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>$y_4$</td>
</tr>
<tr>
<td>5</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>$y_5$</td>
</tr>
<tr>
<td>6</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>$y_6$</td>
</tr>
<tr>
<td>7</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>$y_7$</td>
</tr>
<tr>
<td>8</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>$y_8$</td>
</tr>
</tbody>
</table>

$$\sum_{j=1}^{N} x_{ji}^2 = N \quad j=0,1,...,k \quad (18)$$

where k is the number of factors and N is the number of experiments in the design matrix.

The first property expressed by Eq.(16) is the equality of the scalar products of all column vectors to zero and called the property of orthogonality. This property reduces the difficulties in the estimation of coefficients for the regression equation because the coefficient matrix $(X^T X)$ of normal equations becomes diagonal, and its diagonal elements are equal to the number of experiments N in the design matrix. The diagonal elements of the inverse matrix $(X^T X)^{-1}$ are:

$$C_{jj} = 1/N$$

Thus,

$$B = \begin{bmatrix} b_0 \\ \cdot \\ \cdot \\ b_k \end{bmatrix} = (X^T X)^{-1} XY = \begin{bmatrix} \Sigma x_{0i} y_i/N \\ \Sigma x_{1i} y_i/N \\ \cdot \\ \cdot \end{bmatrix}$$

$$\begin{bmatrix} \Sigma x_{0i} y_i \\ \Sigma x_{1i} y_i \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \end{bmatrix} \quad (19)$$
\[ \Sigma x_i y_i \quad (\Sigma x_i y_i)/N \]

As a consequence, any coefficient \( b_j \) of the regression equation is defined by the scalar product of a \( y \) column by the respective \( x_j \) column divided into the number of experiments \( N \) in the design matrix:

\[ b_j = \frac{1}{N} \sum_{i=1}^{N} x_j y_i \]  

(20)

Using the design presented in Table 1 the first should be estimated linear regression equation

\[ y^{\text{cal}} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \]  

(21)

In the case of a fuller regression equation containing interaction coefficients

\[ y^{\text{cal}} = b_0 + h_{12} x_1 x_2 + h_{13} x_1 x_3 + h_{23} x_2 x_3 \]  

(22)

where one has to determine \( b_{12} \), \( b_{13} \), \( b_{23} \), the second order interaction term and \( b_{123} \), the third order interaction terms, the design matrix (Table 1) can be extended for products \( x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3 \).

By making replicate observations, it will be possible to determine \( s^2_{\text{exp}} \) to test the significance of the regression coefficients and, in the case of degrees of freedom, also adequacy of the equation.

Since the correlation matrix \((X^*X)^{-1}\) for a designed experiment is a diagonal matrix,

\[
(X^*X)^{-1} = \begin{bmatrix}
\frac{1}{N} & 0 & 0 \\
0 & \frac{1}{N} & 0 \\
0 & 0 & \frac{1}{N}
\end{bmatrix}
\]

the coefficients of the regression equation are uncorrelated. The significance of each coefficient can be tested on the basis of the Student t-test. Removal of an insignificant coefficient from the regression equation (22) will have no effect on the remaining coefficients.

The diagonal elements of a correlation matrix are equal to one another and so coefficients can be determined with the same accuracy.

\[ s_{b_j} = \frac{s_{\text{exp}}}{\sqrt{N}} \]  

(23)

The estimated regression equation is tested to see how it fits the observations, using Fisher’s test, that is the variance ratio, \( F = s^2_{\text{r}}/s^2_{\text{exp}} \), where

\[ s^2_{\text{r}} = \frac{\sum (y_i^{\text{cal}} - y_i^{\text{exp}})^2}{N - m} \]  

(24)

The value \( F \) should be less then the tabulated value of Fisher’s \( F_{\alpha}(f_1, f_2) \) to fit the experiment.

4 Experimental design for the new product formulation

At formation miristyl-di - methyl - benzyl - ammonium-chloride need seek to determine how yield, \( y \) depends on the temperature \( X_1 \) and the reactants content \( X_2 \) and \( X_3 \). \( \text{N,N',di-methyl-benzyl-amine} \) in the isopropyl alcohol solution reacts with miristyl-chloride and gives miristyl-di - methyl - benzyl - ammonium-chloride:

\[ C_6H_{15}N(CH_2)_2(CH_3)_2 + CH_3(CH_2)_3Cl \rightarrow [CH_3(CH_2)_3N]^+C_6H_{15}CH_2(CH_3)_2Cl^- \]

Miristyl-- di - methyl - benzyl - ammonium-chloride is cationic suface active ingredient fromgroup of quaternary ammonium salts, and is using in pharmaceutical industry as the desiscificenec.

Using the design presented in Table 2 the first need to find the coefficients of the estimated linear regression equation (21).

Table 2. Synthesis of miristyl-di-methyl-benzyl-ammonium-chloride matrix

<table>
<thead>
<tr>
<th>N</th>
<th>( x_0 )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>39.85</td>
</tr>
<tr>
<td>2</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>42.50</td>
</tr>
<tr>
<td>3</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>67.94</td>
</tr>
<tr>
<td>4</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>37.45</td>
</tr>
<tr>
<td>5</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>59.85</td>
</tr>
<tr>
<td>6</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>62.28</td>
</tr>
<tr>
<td>7</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>95.20</td>
</tr>
<tr>
<td>8</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>96.47</td>
</tr>
</tbody>
</table>

\( b_1 \) can be determined by multiplying together the \( x_1 \) and \( y \) columns and adding together the products:
\[
\sum_{i=1}^{N} x_{ij}y_i \quad b_i = \frac{\sum_{i=1}^{N} x_{ij}y_i}{N} = -3.0225
\]

Similarly, 
\[b_0 = 62.6900, \quad b_1 = -3.0225, \quad b_2 = 11.5675, \quad b_3 = 15.7625\]

In the case of a fuller regression equation containing interaction coefficients Equation (22), the design matrix (Table 2) has to be extended with products \(x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3\).

The interaction terms are determined in the same manner as the linear terms. For example,
\[
b_{12} = \frac{\sum_{i=1}^{N} (x_1x_2)y_i}{N} = -34.26
\]

The remaining coefficients are found in a similar manner and are obtained:
\[b_{12} = -14.2450, \quad b_{13} = 3.9375, \quad b_{23} = 5.8075, \quad b_{123} = 4.0125\]

By making replicate observations, it will be possible to determine \(s_{exp}^2\) to test the significance of the regression coefficients and, in the case of degrees of freedom, also the adequacy of the equation.

Since the correlation matrix \((X^TX)^{-1}\) for a designed experiment the coefficients of the regression equation are uncorrelated. The significance of each coefficient are tested on the bases of the Student t-test. All coefficient are significant, and regression equation
\[
y_{cal} = 54.57 - 11.14x_1 + 92.54x_2 + 126.10x_3
\]
\[-34.26x_1x_2 + 31.50x_1x_3 + 47.02x_2x_3 + 32.02x_1x_2x_3 (24)
\]

is adequate.

5 Results and discussion
A response surface for reaction yield vs. temperature \(X_1\) and reactant alkyl-chloride contents \(X_2\) for constant amine contents \(X_3\) has shown in Fig. 1.

A response surface for reaction yield \(y\) vs. reactants alkyl-chloride \(X_2\) and amine \(X_3\) for temperature \(X_1\) constant has shown in Fig. 2.

A response surface for reaction yield \(y\) vs. temperature \(X_1\) and reactant amine \(X_3\) for alkyl-chloride \(X_2\) constant has shown in Fig. 3.
6 Conclusion
The optimal design of experiment method has leaded to more effective investigation in the new product formulation. Correlation analysis has gave useful information about the examined reaction system.

An optimum decision strategy is formulated on the basis of an analysis of the situations that are likely to develop in the use of system under various conditions and is originally presented as a set of verbal specifications.

Chemometric methods make possible full automation of experiments when system operates in real time with computer.

Notation
- \( f_1, f_2 \): number of freedom parameter
- \( m \): number of replicate experiments
- \( N \): total number of experiments
- \( s \): deviation
- \( s_{exp}^2 \): sum of squares of deviation of the experimental values
- \( x \): factor variable
- \( X \): process variable
- \( y \): dependent variable

Subscript
- \( i \): number of experiments

References: