IMK-14 – Research & Developement in Heavy Machinery 20(2014)2 EN55-60
UDC 621 ISSN 0354-6829

Selection of the Optimal Mathematical Model of Multiple Regression in the Ternary Mixture Experiments

M. Kolarević[1,*] - D. Minić[2]- - M. Rajović[1] - V. Grković[1] - Zv. Petrović[1]
[1] Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia
[2] Faculty of Technical Sciences, Kosovska Mitrovica, Serbia

For a three-component system, regression models can be generally set in the form of polynomials which are usually defined by the following Scheffé canonical forms: a) linear model, b) square model, c) incomplete cube model, d) complete cube model, e) incomplete quartic model, f) complete quartic model. From a lot of models that meet the adequacy requirement it is necessary to choose a model with a rational number of variables for the purpose of easy interpretation and practical application of the model. The paper presents the criteria for evaluation of the model quality and selection of the "optimal" model composition with a "rational" number of variables.

Keywords: Optimal Model, Multiple Regression, Ternary Mixture Experiments

0. INTRODUCTION

Three-component systems can be graphically represented in 2-D space by applying ternary graphs. The main condition for application of ternary graphs is:

$$0 \le X_i \le 1; \qquad \sum_{i=1}^{3} X_i = 1. \qquad (1)$$

Xi – the relative proportion of a component in the mixture.

From the previously mentioned conditions, it is noticeable that the proportion of each component in the mixture depends on the proportion of the remaining two components.

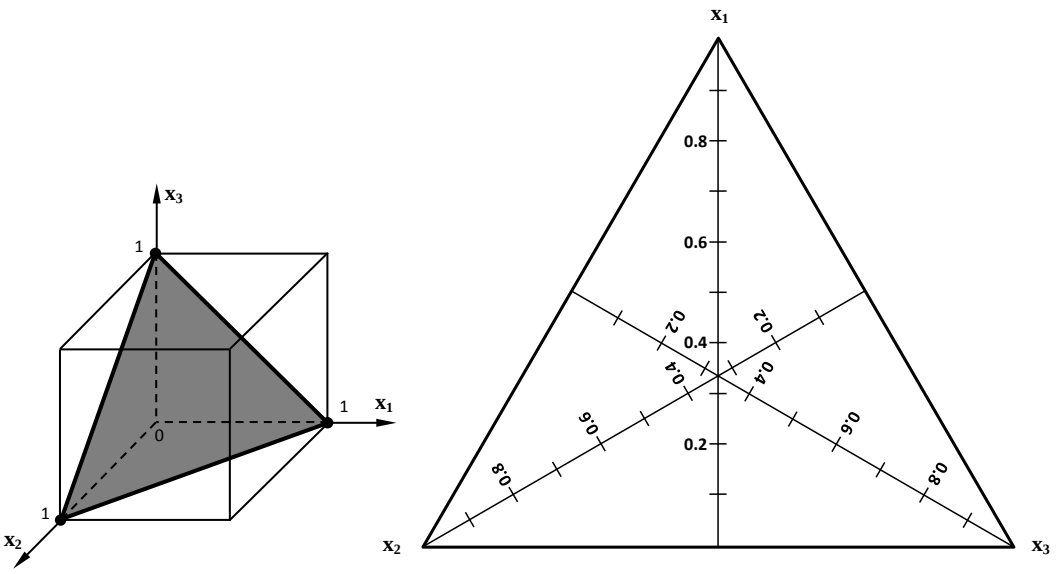

Fig. 1. *Triangular (trilinear) coordinate system and representation of vertical sections and the directions of increase in the proportion of individual components*

Each point inside the triangle represents a corresponding composition of the three-component system. The vertices of the triangle represent pure substances, while the points on the sides of the triangle represent two-component systems. For a point inside the triangle, the proportion of each component is read by drawing lines through the given point in such a way that they are parallel to the sides of the triangle (Figure 2).

For the three-component system, regression models can be generally set in the form of polynomials which are defined by the following canonical or Scheffé forms [1] [2] [7]:

*Corr. Author's Address: Faculty of Mechanical and Civil Engineering in Kraljevo, Dositejeva 19,Kraljevo, Serbia, kolarevic.m@mfkv.kg.ac.rs

1. Linear

$$\hat{y} = \sum_{i=1}^{q} \beta_i x_i \tag{2}$$

2. Quadratic

$$\hat{y} = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q-1} \sum_{j}^{q} \beta_{ij} x_i x_j \tag{3}$$

3. Special Qubic

$$\hat{y} = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q-1} \sum_{j}^{q} \beta_{ij} x_i x_j + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{ijk} x_i x_j x_k \tag{4}$$

4. Full Cubic

$$\hat{y} = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q-1} \sum_{j}^{q} \beta_{ij} x_i x_j + \sum_{i<j}^{q-1} \sum_{j}^{q} \delta_{ij} x_i x_j (x_i - x_j) + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{ijk} x_i x_j x_k \tag{5}$$

5. Special Quartic

$$\hat{y} = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q-1} \sum_{j}^{q} \beta_{ij} x_i x_j + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{iijk} x_i^2 x_j x_k + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{ijjk} x_i x_j^2 x_k + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{ijkk} x_i x_j x_k^2 \tag{6}$$

6. Full Quatric

$$\begin{aligned} \hat{y} = & \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q-1} \sum_{j}^{q} \beta_{ij} x_i x_j + \sum_{i<j}^{q-1} \sum_{j}^{q} \delta_{ij} x_i x_j (x_i - x_j) + \sum_{i<j}^{q-1} \sum_{j}^{q} \gamma_{ij} x_i x_j (x_i - x_j)^2 + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{iijk} x_i^2 x_j x_k \\ & + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{ijjk} x_i x_j^2 x_k + \sum_{i<j}^{q-2} \sum_{j<k}^{q-1} \sum_{k}^{q} \beta_{ijkk} x_i x_j x_k^2 + \sum_{i<j}^{q-3} \sum_{j<k}^{q-2} \sum_{k<l}^{q-1} \sum_{l}^{q} \beta_{ijkl} x_i x_j x_k x_l \end{aligned} \tag{7}$$

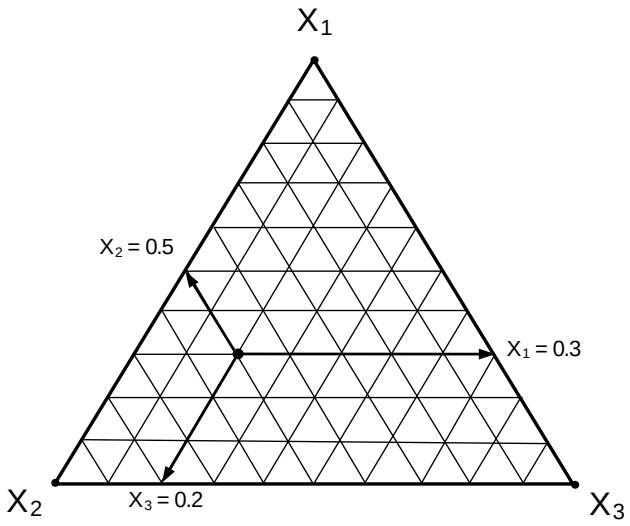

Fig. 2. *Determination of the composition of an alloy in the ternary system*

The selection of the regression model of the ternary system depends, before all, on the available number of design points. The necessary number of design points for the corresponding number of components and the required polynomial degree can be calculated based on the expression [2]:

$$N = \frac{(n+q-1)!}{n!(q-1)}. \tag{8}$$

where:

n – the polynomial degree,
q – the number of components

As there are also incomplete models of the third and fourth degrees, the number of regression coefficients for them cannot be calculated based on the expression (8), and therefore the necessary number of design points for setting up of regression models of ternary systems is presented in Table 1.

The sufficient number of design points for the highest degree model does not mean that it will be the best one. A too high polynomial degree may lead to adoption of an inadequate mathematical model (Figure 3). Besides, for a more rational use and interpretation, a model with a degree which is as low as possible should be selected.

In order to carry out the procedure of regression analysis of the three-component system and select an adequate regression model, it is necessary to respect the following phases [6]:

1. Selection of possible forms of regression models based on the available number of design points
2. Calculation of regression coefficients for all selected models
3. Checking the adequacy of selected mathematical models
4. Selection of the optimal regression model
5. Evaluation of the significance of regression coefficients of the selected model
6. Calculation of confidence limits of regression coefficients of the selected model
7. Calculation of confidence limits of the selected regression model
8. Graphical interpretation of the mathematical model by contour and surface ternary graphs.

Table 1. *Necessary number of design points for setting up of the regression model of the ternary system*

Regression model	***Response subscripts***	***Number of regression coefficients***	***Σ***
Linear	*i*	3	3
Second degree	*i* *ij*	3 3	6
Incomplete third degree	*i* *ij* *ijk*	3 3 1	7
Complete third degree	*i* *iij* *ijk*	3 6 1	10
Incomplete fourth degree	*i* *ij* *iijk*	3 3 3	9
Complete fourth degree	*i* *ij* *iiij* *ijk*	3 3 6 3	15

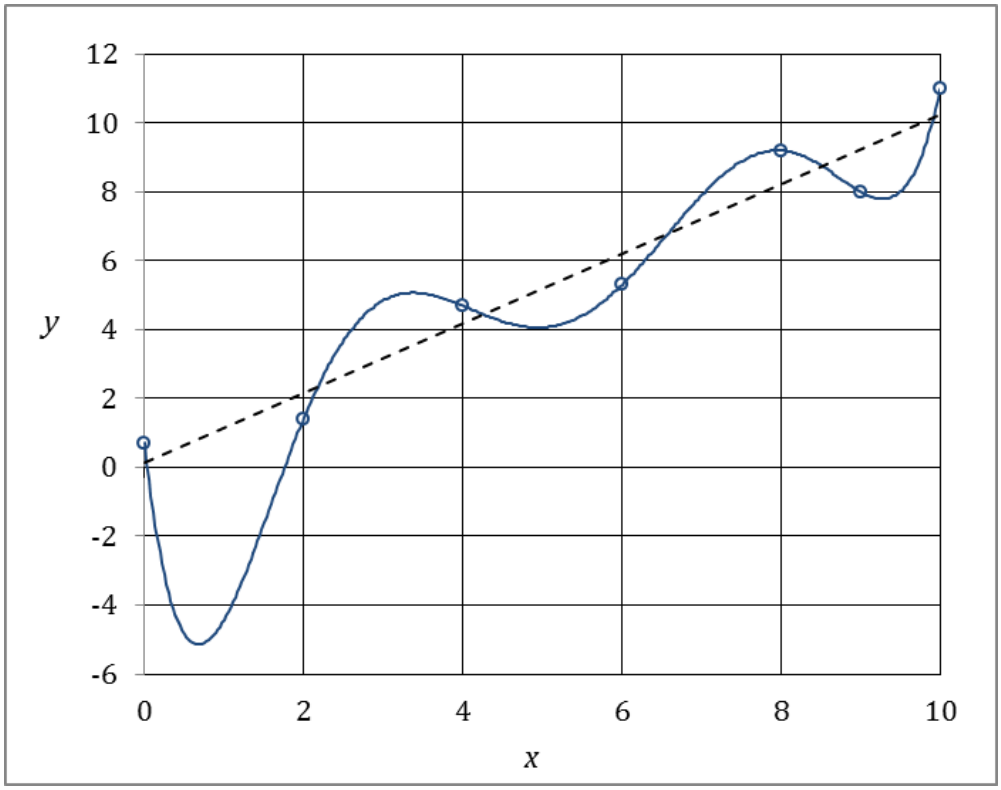

Fig. 3. *Experimental values and the look of the regression curve of the first order (dashed line) and the regression curve of the 6th order (continuous line)*

1. QUALITY INDICATORS OF THE REGRESSION MODEL

For quality evaluation of the model and selection of the "optimal" composition of the model with a "rational" number of variables, the following statistical-analytical values are available [4]:

- the coefficient of determination R^2
- the adjusted coefficient of determination R^2_{adj}
- the residual mean square-variance $\hat{\sigma}^2$
- Mallows' indicator C_p
- Prediction Sum of Squares Statistic - *PRESS*
- Akaike's Information Criterion - *AIC*
- Schwartz's Bayesian Criterion - *SBC*
- Bayes' Information Criterion - *BIC*
- Amemiya's Prediction Criterion – *PC*
- Witcomb Score - *WS*

1.1 The coefficient of multiple determination

The coefficient of multiple determination is the ratio between the sum of squares of deviation of regression values from their arithmetic mean and the sum of squares of deviation of values of the dependent variable from its arithmetic mean [3]:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \quad 0 \le R^2 \le 1. \tag{9}$$

where:

SS_R – the regression sum of squares

SS_E – the error (residual) sum of squares

SS_T – the total sum of squares

The coefficient of multiple determination can have the value between zero and one. According to this criterion, the most representative model is the one whose coefficient of determination is closer to one.

The disadvantage of this indicator is that it is a monotone non-decreasing function of the number of regression variables, so that its highest value is for the model with all available regression variables, which may lead to the selection of a model whose dimensions are too large [4].

1.2 The adjusted coefficient of multiple determination

The adjusted coefficient of multiple determination is given by the expression [3]:

$$R^2_{adj} = 1 - \frac{\frac{SS_E}{(n-k)}}{\frac{SS_T}{(n-1)}} = 1 - \frac{n-1}{n-k}\left(1 - R^2\right), \quad R^2_A \le R^2. \tag{10}$$

where:

n – the number of samples

k – the number of variables (coefficients of the regression model)

$v_2 = n-k$ –the number of degrees of freedom

The highest value of this coefficient is one, and unlike the coefficient of determination, it can also be a negative number. The value of the adjusted coefficient of determination is not a monotone increasing function of the number of variables, but it depends on the number of degrees of freedom, i.e. on the model size and it is more suitable than the coefficient of determination because it ensures that the model does not include too many variables. The highest adjusted coefficient of determination is relevant for the selection of the regression model.

1.3 Residual mean square (estimatie of variance) of regression

The residual mean square of regression is the ratio between the residual sum of squares and the number of degrees of freedom $v_2 = n-k$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k}. \tag{11}$$

The statistical representation of the model increases with the decrease of the residual mean of squares, and so this is the criterion for selection of the model with the lowest value of this indicator.

1.4 Malows' criterion C_p

Malows' criterion is given by the expression:

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - (n - 2p), . \tag{12}$$

where:

p – the number of parameters in the regression model

$SS_E(p)$ – the residual sum of squares for a model that includes p parameters

$\hat{\sigma}^2$ – the residual mean of squares for a model with the maximum number of regression variables

Most users adhere to one of the two criteria for selection of the optimal model:

- select the model with a small value of C_p
- select the model with a small positive (or negative) difference between C_p and p.

1.5 PRESS (*Prediction Sum of Squares Statistic*)

PRESS represents the residual sum of squares which is calculated in a specific way for every possible regression model [4]. If k regression variables are available, a model can be formed with one variable or a combination of two, three or more regression variables. The total number of all possible regressions is equal to 2^k-1. The *PRESS* value is calculated for each 2^k-1 regression. The PRESS statistic represents the sum of squares of n PRESS residuals [1]:

$$PRESS = \sum_{i=1}^{n}\hat{e}^2_{(i)} = \sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2. \tag{13}$$

where:

$\hat{e}_{(i)} = y_i - \hat{y}_{(i)}$ – the i-th PRESS residual.

The numerical calculation is simplified by applying the diagonal elements of the h_{ii} "hat"-matrix

$$H = X\left(X`X\right)^{-1}X` = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ . & . & . & . \\ . & . & . & . \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix} \quad (14)$$

in which the values of independent variables are contained in the matrix **X**. The *PRESS* residual deviations are given by the expression:

$$\hat{e}_{(i)} = \left(y_i - \hat{y}_{(i)}\right) = \frac{\hat{e}_i}{1-h_{ij}}, \quad (15)$$

and so the *PRESS* indicator can be shown by the expression [1]:

$$PRESS = \sum_{i=1}^{n}\left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2. \quad (16)$$

A model with the lowest *PRESS* indicator and a relatively small number of parameters is chosen for the optimal regression model.

1.6 Akake's Information Criterion – *AIC*

Akake's Information Criterion-*AIC* is given by the expression:

$$AIC = n \cdot ln\left(\frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}}{n}\right) + 2p. \quad (17)$$

where:

n – the number of values (sample size)

$\sum_{i=1}^{n}\hat{e}_i^{\,2}$ – the residual sum of squares

p – the number of parameters in the model.

Small values of this criterion are desirable.

1.7 Schwartz's Bayesian Criterion – *SBC*

Schwartz's Bayesian Criterion – *SBC* is similar to the previous criterion:

$$SBC = n \cdot ln\left(\frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}}{n}\right) + p \cdot ln\, n. \quad (18)$$

1.8 Bayes' Information Criterion – *BIC*

Bayes' Information Criterion – *BIC* is defined by the expression:

$$BIC = n \cdot ln\left(\frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}}{n}\right) + 2\left(p+2\right)q - 2q^2, \quad (19)$$

where:

$$q = \frac{\hat{\sigma}^2}{\left(\frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}}{n}\right)}.$$

$\hat{\sigma}^2$ – the estimate of variance based on a model with all variables

1.9 Amemiya's Prediction Criterion – *PC*

Amemiya's Prediction Criterion –*PC* is given by the expression:

$$PC = \frac{\sum_{i=1}^{n}\hat{e}_i^{\,2}\left(1+\frac{p}{n}\right)}{\left(n-p\right)}. \quad (20)$$

The criteria *AIC, SBC, BIC* and *PC* represent a set of similar criteria and a minimum value is desirable in all of them.

1.10 Witcomb Score

Design-Expert [6] uses the following system to score models.

1. Calculate the values (*M*) from the sequential model of the sum of squares.
   - *M = 1* *if* *p ≤ 0,5*
   - *M = 0,05/p* *if* *p > 0,5*
   - *M = 0* *if* *model is aliased*
2. Calculate the values (L) from the Lack-of-fit table:
   - *L = 1* *if* *p ≥ 0,10*
   - *L = p / 0,10* *if* *p < 0,10*
3. Combine the first two evaluations with the statistics R^2, which forms the total evaluation:
   - *Score-1 = (M)(L)($R^2_{predicted}$)*
   - *Score-2 = (M)(L)($R^2_{adjusted}$)*
   - Select the model with the maximum score. If all model scores are less than or equal to zero, select the *mean* model

where:

$R^2_{predicted} = 1 - SS_{PRESS}/(SS_{Total} - SS_{Blocks})$

Two models are mainly proposed. Design-Expert then conservatively defaults to the model scored highest on the basis of predicted *r*-squared.

2. CONCLUSION

In the process of investigation of electrical and mechanical properties of alloys, three-component systems play a very important role. Regression analysis provides a possibility to use experimental results in order to obtain the theoretical dependence of these values on the molar ratio of certain components of the mixture.

Regression models for a three-component system are generally defined by polynomials from the first to the fourth degrees, where the selection of the regression model of the ternary system depends, before all, on the available number of design points.

It is desirable to analyze several possible regression models and, out of the models with proved adequacy, select the one which best describes the given phenomenon. The fact that the higher degree models are very complicated for interpretation of the observed phenomenon and that the highest degree model does not always have to be the best one should be considered.

For quality evaluation of the model and selection of the "optimal" composition of the model with a "rational" number of variables, there is a multitude of statistical-analytical values which are described in the previous chapter. The quality of the solution depends on the criterion applied for selection of the "optimal" model. It is diffulct to say which criterion is the best one and hence it is desirable to combine several criteria to select the adequate mathematical model.

3. ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Ministry of Education and Science of the Republic of Serbia for their support to this research through the projects TR37020 & OI172037.

4. REFERENCES

[1] Cornell, *Experiments with Mixtures*, 2nd ed., John Wiley&Sons, Inc, New York, 1990.

[2] Lazić Ž. *Design of Experiments in Chemical Engineering*, Wilez-VCH Verlag GmbH&Co.KGaA, Weiheim, 2004

[3] Montgomery D., *Design and Analysis of Experiments*, 5th edition, John Wiley&Sons, INC, New York

[4] Šošić I. *Primijenjena statistika*, Školska knjiga, Zagreb, 2004

[5] M. Kolarević, M. Vukićević, B. Radičević, M. Bjelić, V. Grković: *A Methodology For Forming The Regression Model Of Ternary System*, The Seventh Triennial International Conference Heavy Machinery HM 2011, Faculty of Mechanical Engineering, Proceedings, Vrnjačka Banja, pp. E 1-6, 29 June-2 July 2011

[6] Design Expert v.8 User`s Guide, Stat-Ease, http://www.statease.com/dx8_man.html

[7] Kolarević M, Rajović.M, Bjelić M., *Ternary graph i njegova primena u regresionoj analizi*, IMK-14 Istraživanje i razvoj, časopis instituta IMK "14 OKTOBAR" - Kruševac, Godina XI, broj (22-23) 3-4, Kruševac 2005, str. 113-122