ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2009, Vol.38, No.3

INVESTIGATION OF THE STATISTICAL MODEL BASED OPTIMIZATION ACCURACY UNDER EXPERIMENTAL ERRORS. CASE STUDY: OPTIMIZATION OF NUTRIENT MEDIA FOR MICROORGANISMS' CULTIVATION PROCESS

Donatas Levišauskas, Tomas Tekorius

Process Control Department, Kaunas University of Technology
Studentų St. 50, LT–51368 Kaunas, Lithuania
e-mail: tomas.tekorius@ktu.lt

Abstract. The response surface methodology based process optimization procedure, including design of factorial experiments, development of statistical model and estimation of the optimum response point is simulated in order to investigate influence of experimental errors and experimental design area on the optimization accuracy. The investigated practical problem is optimization of nutrient media composition for cultivation of microorganisms' culture. The optimization results under a priori estimated experimental errors and various factor variation ranges in factorial experiments are investigated and the relationships between the factors variation range and the confidence interval of the optimum point estimations are determined.

Keywords: statistical model, experimental design, optimization.

1. Introduction

Recently, many practical problems of process optimization are solved by applying the response surface methodology (RSM). The RSM is a collection of mathematical and statistical techniques that are used for design of factorial experiments, development of models based on observed experimental data and analysis of problems, in which the desired response (output) of investigated process depends on several factors (independent variables). The objective of investigation is to optimize this response [1].

In most RSM problems the true functional relationship between the desired response and the independent variables is unknown and the first step in the RSM application is to find a suitable approximation for this relationship. In many practical cases the first-order and the second-order multiple polynomial models demonstrate an adequate approximation of response surfaces over relatively small regions of interest [2-5].

The parameters of the multiple polynomial models are typically estimated by the method of least squares. The method produces an unbiased estimation of the parameters if experimental results and conditions meet some requirements: there is no correlation between experimental results at separate points of factorial experiment; dispersion of the results does not depend on their absolute values; the desired experimental conditions are set precisely.

However, the cases are quite common in practice, in which the last-mentioned requirement is not satisfied, i.e., conditions of factorial experiments are realized with some deviations. The accuracy of measurements along with the action of uncontrollable factors also influences the observed experimental data. These experimental errors influence accuracy of the experimental data-based response surface model and the model-based optimization of the investigated process response.

The above problem arises when solving optimization problems of nutrient medium composition for cultivation particular cultures of microorganisms. Parallel factorial experiments for identification the response surface models are usually realized in flasks, in which the components of nutrient media are dosed manually. Estimation of dosage accuracy in repeated experiments has shown that the standard error totals up to 1% of absolute values. The experimental results are also corrupted by uncontrollable factors and the errors of measurement methods. By applying the sequential RSM procedure to the nutrient medium optimization problem it was noticed that narrowing the investigation area of factorial experiments around the supposed location of optimum point in order to increase accuracy of the response surface model and the model-based estimation of optimum point did not

lead to the desired goal. Location of the calculated optimum point varied in a wide range depending on the experimental design area and experimental data that was used for the statistical model identification.

In this paper, we investigate the influence of experimental errors and ranges of the independent variables variation in factorial experiments on the accuracy of statistical model-based estimation of the optimum response point. An object of investigation is the problem of nutrient medium optimization for cultivation the microorganisms' culture *Enterobacter aerogenes 17 E13* in order to maximize production of physiologically active biomass for degradation of the grease wastes. The investigated problem is a first stage of the overall optimization of the biotechnological process [6].

The investigation is based on analysis of statistical data, generated by the simulation experiments of the RSM-based optimization procedure for various factor manipulation ranges at particular experimental errors.

2. Statistical model-based optimization procedure

Statistical models for approximation the desired response surfaces are developed using data of factorial experiments [1,2]. In practical applications, the second order polynomial models are widely used [1-5]:

$$Y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_{ij} x_i x_j , \qquad (1)$$

where Y is predicted response, x_i are independent variables, n is the number of independent variables, a_o are the model parameters.

Parameters of the polynomial model (1) are identified using the least squares method [2]:

$$\mathbf{A} = \left[\mathbf{F}^T \mathbf{F}\right]^{-1} \mathbf{F}^T \mathbf{Y} , \qquad (2)$$

where $\mathbf{A}$ is the parameter vector of the model (1);

$\mathbf{F}$ is the matrix of independent variables of the model (1):

$$\mathbf{F} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{11}^2 & \cdots & x_{11}x_{21} & \cdots & x_{n-1,1}x_{n1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \cdots & x_{1N}^2 & \cdots & x_{1N}x_{2N} & \cdots & x_{n-1,N}x_{nN} \end{bmatrix};$$

$\mathbf{Y}$ is the vector of experimental results, $\mathbf{Y} = \left[y_1 \ldots y_N\right]^T$;

N is the number of experimental points.

Using the identified model (1), the point $\mathbf{x}^* = \left[x_1^* \quad \ldots \quad x_2^*\right]^T$ is calculated, at which the predicted response takes extreme value. If the extreme point lies on the boundary of experimental design area, a normalized gradient vector at this point is calculated:

$$grad_n Y\left(\mathbf{x}^*\right) = \frac{\nabla Y(\mathbf{x})}{\left\|\nabla Y(\mathbf{x})\right\|_{\mathbf{x}=\mathbf{x}^*}} , \qquad (3)$$

$$\nabla Y(\mathbf{x}) = \left[\frac{\partial Y(\mathbf{x})}{\partial x_1} \quad \ldots \quad \frac{\partial Y(\mathbf{x})}{\partial x_n}\right]^T .$$

The gradient vector (3) determines the search direction of the optimum point outside the experimental design area. Along the calculated direction, the expected location of optimum point is predicted and the new cycle of factorial experiment and response surface estimation around the predicted point is performed.

If the calculated extreme point lies inside the experimental design area, the point is assumed to be an optimum point and the test experiment is carried out at that point.

By applying the statistical model-based optimum point search procedure it is important to sum up prediction accuracy of the optimum point location, as the identified values of model parameters (2) depend on experimental data that are corrupted by experimental errors. To analyze an influence of the errors, the vector $\mathbf{Y}$ of experimental results can be represented as follows:

$$\mathbf{Y} = \mathbf{Y}_{mean} + \Delta\mathbf{Y} + \mathbf{E} , \qquad (4)$$

where $\mathbf{Y}_{mean} = y_{mean}\left[1 \quad \cdots \quad 1\right]^T$, y_{mean} is the mean of experimental results, $\Delta\mathbf{Y} = [y_1 - y_{msan} \; \cdots \; y_N - y_{msan}]^T$ is a vector of deviations of experimental results from the mean, $\mathbf{E} = e\left(\sigma_e^2, 0\right)\left[1 \quad \cdots \quad 1\right]^T$, e is random experimental error with dispersion σ_e^2 and zero mean.

The estimated value of free term a_0 of the model (1) is mainly related to the mean value of experimental results y_{mean}, while the values of the other parameters are related to the variations $\Delta\mathbf{Y}$. The range of variations $\Delta\mathbf{Y}$ depends on the experimental design: it decreases with narrowing the design area. Assuming that statistical characteristics of experimental errors do not depend on location of experimental points, the ratio of experimental errors $\mathbf{E}$ to the variations $\Delta\mathbf{Y}$ increases with narrowing of experimental design area. With a higher level of relative errors in experimental data, accuracy of the identified statistical model and the model-based prediction of the response surface decreases.

Therefore, by solving practical optimization problems it is important to choose reliable experimental design areas in order to obtain the best solution under specific experimental errors. The reliable areas are related to statistical characteristics of errors as well as a shape of response surface, and, therefore, they are specific for each optimization problem.

3. Optimization of nutrient media for microorganisms' cultivation process

The discussed optimization procedure has been applied for optimization of nutrient medium for cultivation the microorganisms culture *Enterobacter aerogenes 17 E13* in order to maximize the cells' biomass growth rate.

Close to D-optimal (B_3) experimental design matrix [3] and results at the last stage of the sequential search procedure are presented in Table 1. The repeated experiments at the centre point are carried out to estimate statistical characteristics of experimental errors.

Table 1. Experimental design matrix, experimental results and the model (5) predictions

No	Code values			Real values			Experimental results	Model predictions
	x_1	x_2	x_3	X_1 g/l	X_2 g/l	X_3 g/l	Y_{exp} $\times 10^9$ CFU/ml	Y_{mod} $\times 10^9$ CFU/ml
1	1	1	1	8	10	7	3.19	3.01
2	-1	1	1	6	10	7	3.95	3.95
3	1	-1	1	8	8	7	3.40	3.23
4	-1	-1	1	6	8	7	2.62	2.64
5	1	1	-1	8	10	5	1.74	1.64
6	-1	1	-1	6	10	5	2.86	2.95
7	1	-1	-1	8	8	5	3.65	3.58
8	-1	-1	-1	6	8	5	3.26	3.36
9	1	0	0	8	9	6	3.58	4.11
10	-1	0	0	6	9	6	4.67	4.47
11	0	1	0	7	10	6	4.01	4.20
12	0	-1	0	7	8	6	4.39	4.51
13	0	0	1	7	9	7	4.13	4.47
14	0	0	-1	7	9	5	4.16	4.15
15	0	0	0	7	9	6	5.33	4.96
16	0	0	0	7	9	6	4.91	4.96
17	0	0	0	7	9	6	5.28	4.96

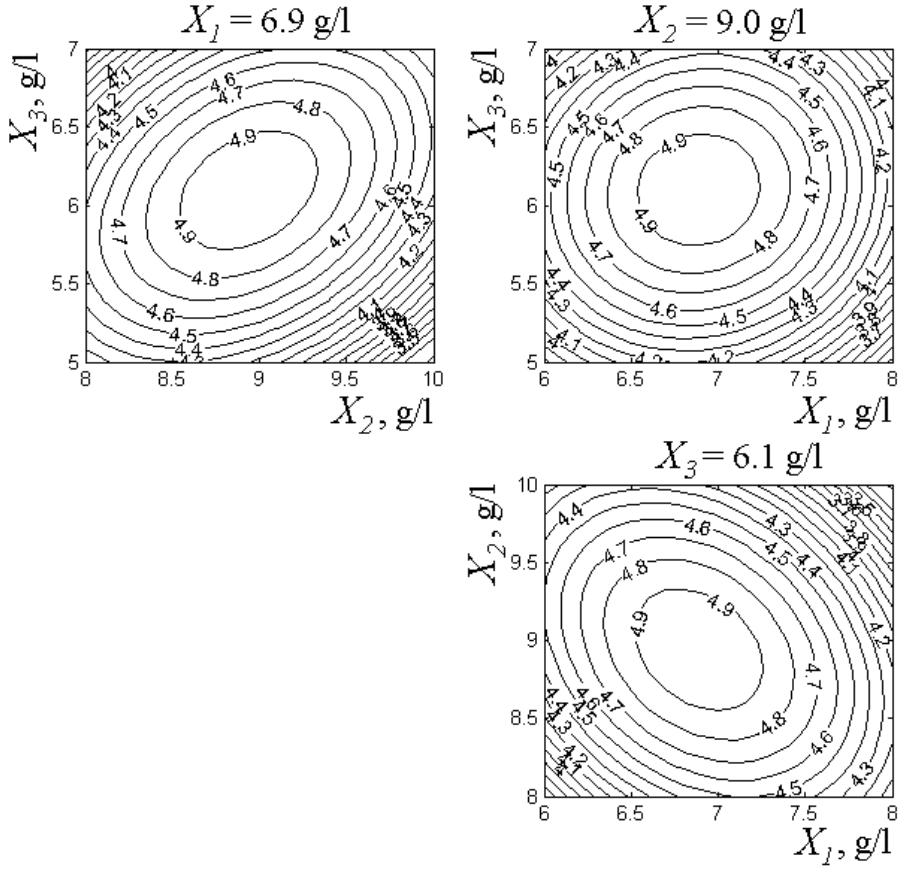

Figure 1. Sectional views of the predicted response surface of cells' concentrations in the vicinity of optimum point

With reference to the experimental data, the following response surface model has been identified:

$$Y = 4.961 - 0.180x_1 - 0.156x_2 + 0.161x_3 - 0.668x_1^2 - 0.598x_2^2 - 0.648x_3^2 - 0.382x_1x_2 + 0.093x_1x_3 + 0.429x_2x_3 \quad (5)$$

where Y is predicted response (cells' concentration after 4h of cultivation, colonies forming units per millilitre (CFU/ml)), x_1, x_2, x_3 are code values of initial concentrations of sunflower oil, casein hydrolyzate and yeast extract, respectively, (nondimensional).

The statistical test for model adequacy [1] proved that the model (5) is suitable for prediction the cells' growth rate. The model-based predictions at the experimental design matrix points are presented in Table 1. The predicted optimal point in the scale of real values is as follows: sunflower oil – 6.9 g/l, casein hydrolysate – 9.0 g/l and the yeast extract – 6.1 g/l.

The calculated sectional views of response surface in the vicinity of the predicted optimum point are presented in Figure 1.

The shape of response surface in Figure 1 shows that the optimum point lies inside the experimental design area. However, the predicted location of optimum point depends on the model (5) parameter values that are identified using the particular set of experimental data.

Due to experimental errors, with the other sets of experimental data, obtained by testing the same technological process, predicted location of optimal point will change.

In order to evaluate an influence of experimental errors and experimental design area to the optimum point prediction accuracy, simulation experiments of the statistical model-based optimization procedure were carried out. In the simulation experiments, the response surface model (5) was employed to model reactions of the real process under specific experimental errors at various factor manipulation ranges in factorial experiments.

4. Simulation of statistical distribution of the optimal solution points under experimental errors

In the simulation experiments, factors are varied according to the experimental design plan for code values presented in Table 1. The factor manipulation ranges ($\Delta X_i = X_{i\max} - X_{i\min}$) were gradually narrowed around the optimum point ($X_{1opt} = 6.9$ g/l, $X_{2opt} = 9.0$ g/l, $X_{3opt} = 6.1$ g/l) from 3.0 g/l to 1.2 g/l with a step of 0.1 g/l.

The experimental errors are simulated by adding a noise to the factor values at experimental points and to the process output:

$$X_{ij}^e = X_{ij} + \sigma_X \; rand(0,1),\ i=1,2,3,\ j=1,...,N, \quad (6)$$

$$Y_j^e = Y_j + \sigma_Y \; rand(0,1), \quad (7)$$

where X_{ij}^e is simulated value of the i-th factor at the j-th experimental point, Y_j^e is simulated value of process output at the j-th experimental point, σ_X is standard deviation of setting the experimental conditions, σ_Y is standard deviation of experimental results, *rand* (0,1) is a number from Gaussian random number sequence with zero mean an unit variance.

The standard deviations related to experimental errors were estimated using experimental data of repeated experiments: $\sigma_X = 0.1$ g/l, $\sigma_Y = 0.2 \times 10^9$ CFU/ml.

For each range of factor manipulation, 10^4 factorial experiments are generated. Using the simulated experimental data, parameters of the response surface model (1) are identified and location of optimum point is estimated. As a result, the set of the optimal point estimations, scattered around the true optimal point, is obtained. Distribution of the estimated optimum points at the factor manipulation ranges $\Delta X_i = 2$ g/l, (i=1,2,3) is exemplified in Figure 2. The area marked by dashed lines in Figure 2 indicates the experimental design area, experimental data from which are used for fitting the response surface model. The presented results demonstrate that scattering of the optimal concentration estimations around the true values is of the same order as the factor manipulation ranges in factorial experiments.

In Figure 3, the histograms of the simulated optimum point estimations are presented at various factor variation ranges. The histograms demonstrate that dispersion of the optimum point estimates increases with decreasing the factor variation ranges.

Using the generated statistical data ($n = 10^4$ realizations), the 95 % confidence intervals for location of the estimated optimum point coordinates $X_{i,opt}$ are evaluated by calculation the 2.5 and 97.5 percentiles, taking into account uncertainty in sample estimates.

The 95 % confidence intervals for the calculated percentiles are estimated by using the following procedure [7]:

1) ranking the n sample observations in increasing order of magnitude;

2) calculating the quantities r_q and s_q for the quantiles $q = 0.025$ and $q = 0.975$:

$$r_q = nq - \left[z_{1-\alpha/2} \times \sqrt{nq(1-q)} \right], \quad (8)$$

$$s_q = 1 + nq + \left[z_{1-\alpha/2} \times \sqrt{nq(1-q)} \right], \quad (9)$$

where $z_{1-\alpha/2}$ is the appropriate value from the standard Normal distribution for the $100(1-\alpha/2)$ percentile ($z_{0.975} = 1.96$);

3) rounding r_q and s_q to the nearest integers;

4) r_q-th and s_q-th observations in the ranking determine the $100(1-\alpha)\%$ confidence interval for the quantile.

For the investigated statistical data, the calculated quantities that define margins of the confidence intervals for optimum point estimations are $r_{0.025} = 219$ and $s_{0.975} = 9782$. The calculated confidence intervals and dispersions of the optimum point estimations at the factor manipulation ranges ΔX_i = 1.2; 2.0, 3.0 [g/l] are given in Table 2.

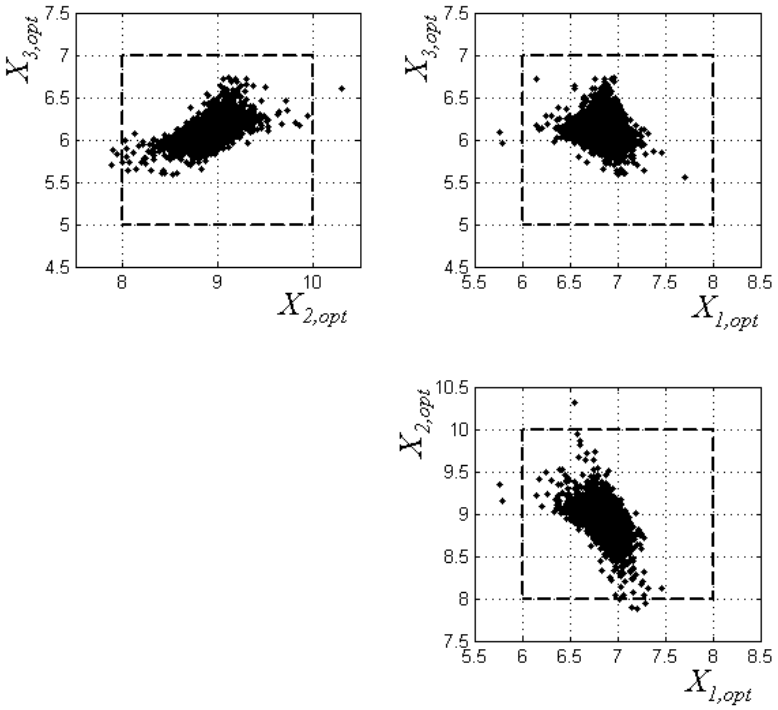

Figure 2. Simulated scattering of the optimum response point estimations at factor manipulation ranges $\Delta X_1 = \Delta X_2 = \Delta X_3 = 2$g/l

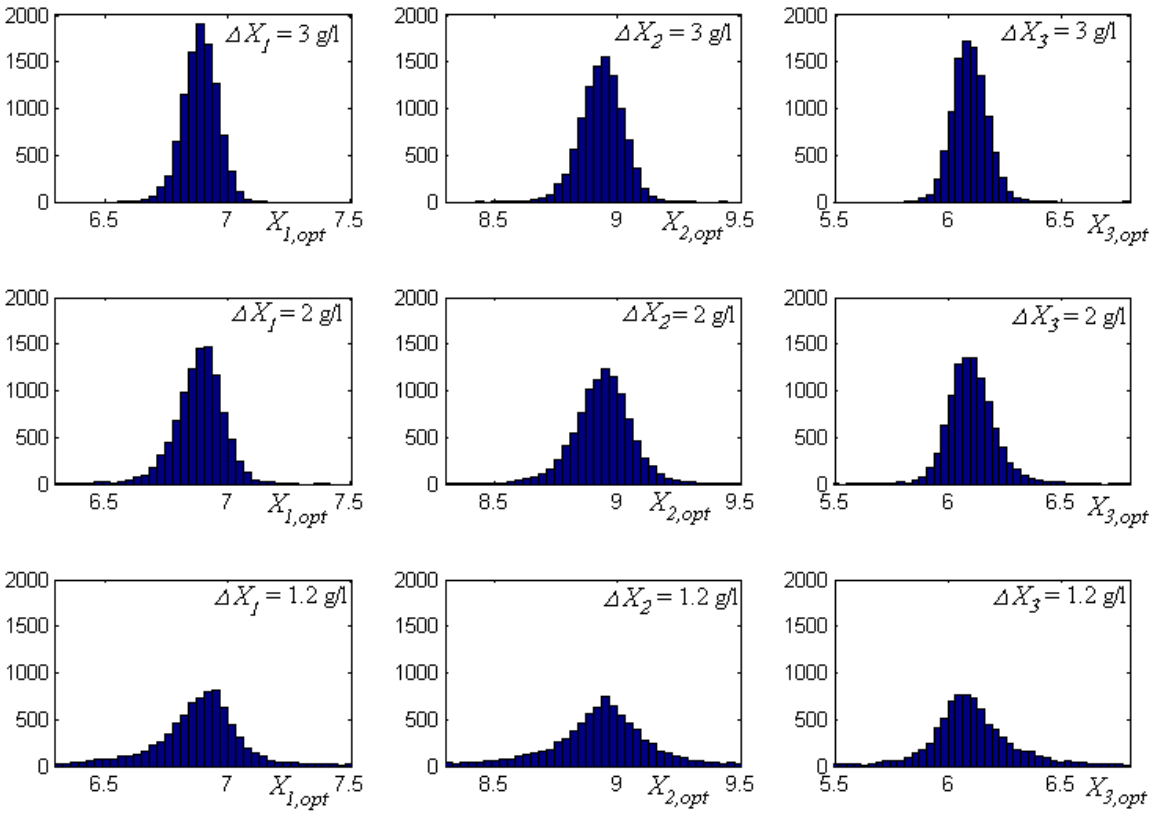

Figure 3. Histograms of simulated optimum point distributions at factor manipulation ranges ΔX_i = 1.2; 2.0, 3.0 [g/l]

Table 2. 95 % confidence intervals and dispersion of the statistical model-based optimum point estimates at various factor variation ranges

Factor variation ranges, g/l	Lower and upper values of the 95 % confidential intervals											
	$X_{1,opt}$, g/l				$X_{2,opt}$, g/l				$X_{3,opt}$, g/l			
	L_1	U_1	U_1-L_1	σ_{x1}	L_2	U_2	U_2-L_2	σ_{x2}	L_3	U_3	U_3-L_3	σ_{x3}
ΔX_i = 1.2	5.13	8.76	3.63	0.68	6.86	11.08	4.22	0.82	4.46	7.93	3.47	0.69
ΔX_i = 2.0	6.64	7.07	0.43	0.09	8.64	9.19	0.55	0.12	5.90	6.35	0.45	0.10
ΔX_i = 3.0	6.74	7.02	0.28	0.07	8.76	9.11	0.35	0.08	5.95	6.25	0.30	0.07

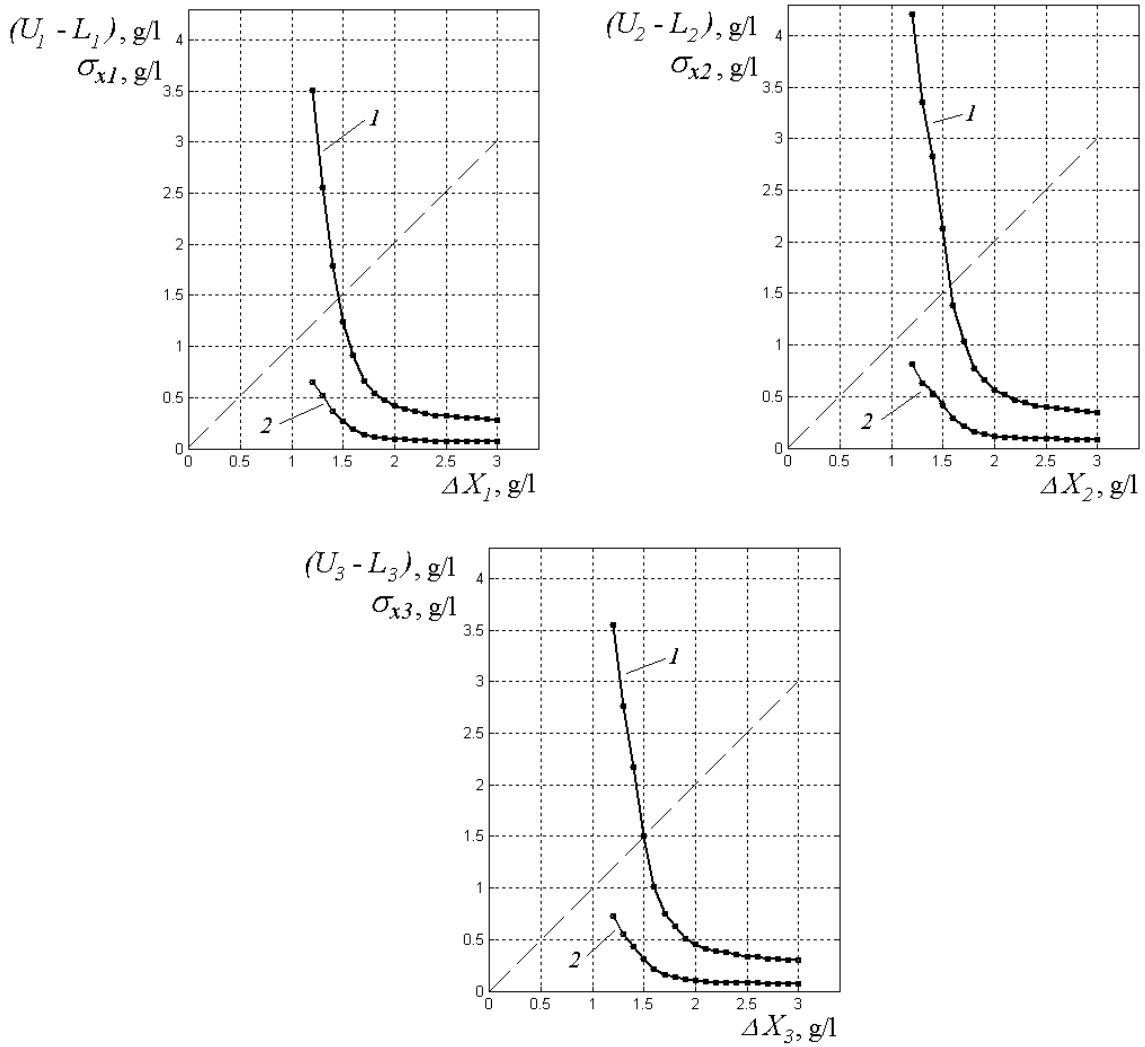

Figure 4. Relationships between the ranges of factor variation ΔX_i and the 95 % confidence interval $U_i - L_i$ (1) and standard deviation σ_{Xi} (2) of the optimal point estimations

As it follows from the results presented in Table 2, the confidence intervals under fixed statistical parameters of experimental errors decrease with increasing the ranges of the factor variations in factorial experiments.

The calculated relationships between the factor manipulation ranges and the statistical characteristics of optimum point estimations are depicted in Figure 4.

The graphs in Figure 4 demonstrate that at factor variation ranges in factorial experiments of less than 2 g/l the confidence interval and the standard deviation of the optimum point estimations substantially increases. At factor variation ranges less than 1.5 g/l, the 95 % confidence intervals of estimated optimal values of factors exceed the factor variation ranges. Therefore, narrowing the search area of the optimum point location below the indicated margin does not provide with higher accuracy of the optimum point estimation in the investigated nutrient medium optimization problem. The calculated relationships between

the range of factor variations and the confidence interval of the optimal point estimate allow us to evaluate achievable accuracy of the statistical model-based optimization and to choose the reasonable area of experimental design.

5. Conclusions

The RSM based optimization procedure is investigated by computer simulation in order to determine influence of experimental errors at different experimental design ranges on the accuracy of optimum point estimation. In the simulation experiments, the problem of optimization the nutrient media for micro-organisms' cultivation process under realistic experimental errors and factorial experiment design conditions has been investigated.

The simulation results demonstrate that the existing errors of factorial experiments noticeably influence accuracy of the statistical model-based optimization results. Narrowing of the factors' manipulation ranges at factorial experiments around the optimum point increases the confidence interval of the optimum point estimates. In the investigated problem, at the factor variation ranges under 2.0 g/l standard deviation of the optimum point estimates noticeably increases; at the ranges under 1.5 g/l the margins of 95% confidence interval of the optimum point estimates outreach the experimental design area.

The quantitative results and conclusions of the presented investigation depend on a shape of response surface and statistical characteristics of experimental errors; therefore they are valid for particular problem only. However, the presented approach can be universally applied for investigation of the RSM procedure based optimization problems in order to evaluate achievable optimization accuracy and the reasonable factor variation ranges under experimental errors.

Acknowledgement

This research is supported by the Agency for International Science and Technology Development Programmes in Lithuania under grant No 31V-22.

References

[1] **R.H. Myers, D.C.Montgomery.** Response Surface Methodology. *John Wiley & Sons, Inc*., 2002.

[2] **D.C. Montgomery.** Design and Analysis of Experiments. *John Wiley & Sons, Inc*., 2001.

[3] **K. Hartmann, E. Lezki, W. Schafer.** Statistische Versuchsplanung und Auswertung in der Stoffwirtschaft. *VEB Deutscher Verlag fur Grundstoffindustrie, Leipzig*, 1974.

[4] **W. Huang, H. Niu, G.H. Gong, Y.R. Lu, Z.S. Li, H. Li.** Individual and combined effects of physicochemical parameters on ellagitannin acyl hydrolase and ellagic acid production from ellagitannin by *Aspergillus oryzae*. *Bioprocess & Biosystem Engineering*. 2007, *Vol*. 30, 281-288.

[5] **D. Levišauskas, T. Tekorius, V. Čipinytė, S. Grigiškis.** Experimental optimization of nutrient media for cultivation of *Arthrobacter* bacteria. *Latvian Journal of Chemistry*, 2004, *No*.1, 75-80.

[6] **D. Levišauskas, V. Galvanauskas, V. Čipinytė, S. Grigiškis.** Optimization of feed-rate in fed-batch culture *Enterobacter aerogenes* 17 E13 for maximization of biomass productivity. *Information Technology and Control*, 2009, *Vol*.38, 102-107.

[7] **D.A. Altman** (Editor). Statistics with Confidence. *London, GBR: BMJ Publishing Group*, 2000.

Received September 2008.