ANALYSIS AND DESIGN OF SIMULATION EXPERIMENTS FOR THE APPROXIMATION OF MODELS

V. Fedorov

July 1983
WP-83-71

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

PREFACE

Much of IIASA's work is concerned with modeling large and complex systems. However, the resulting mathematical models tend to become very complicated and unwieldy, making it very difficult to identify the key relationships between the variables. To overcome this problem, it is often necessary to approximate these "primary models" by more transparent "secondary models".

In this paper, Valeri Fedorov discusses these problems and presents a statistical procedure for secondary model construction. The use of the method is illustrated by application to one of the IIASA energy models.

Andrzej Wierzbicki
Chairman
System and Decision Sciences

ANALYSIS AND DESIGN OF SIMULATION EXPERIMENTS FOR THE APPROXIMATION OF MODELS

V. Fedorov

1. INTRODUCTION

Modern computers make it possible to construct and run complicated mathematical models of complex systems (e.g., economic systems, ecological systems) which involve hundreds of inputs and equations. The links between the different variables (inputs and outputs) and equations in these models are usually very difficult to follow, and this is complicated by the fact that the models are continuously being updated and improved by the incorporation of new mathematical features. Sometimes models consist of modules (elements) prepared by different scholars, and this is one reason why mathematical models (or, more accurately, their computerized counterparts) occasionally become "mysterious" even to their authors. Analytical techniques prove to be useless in analyzing the properties of these models. Since it is not possible to obtain the required results in this way, it is natural to try another approach: one possibility is to carry out experiments on the mathematical models themselves. We shall call these *simulation experiments*.

The question of the effectiveness of the experiments and whether the chosen model adequately describes the empirical data arises at the very beginning of such investigations. To study this, "models" of the models are often constructed. In what follows the terms *secondary model* and *primary model* will be used in an attempt to avoid confusion.

The construction of secondary models can also be stimulated by the fact that the primary models are frequently too detailed for the specific investigations that the researcher wishes to perform. For instance, to describe the behavior of a primary model over a relatively small range of input values it might be sufficient to use a polynomial approximation of the model. One attractive feature of this approximation is that it then becomes possible to develop fast real-time interactive software. This type of software can be extremely useful to decision makers because it allows them to scan a lot of variants in a relatively short time.

The secondary model should reflect the structure of both the primary model and the experiment. As in other experimental situations, it is possible to construct a number of primary models of one system which are all based on different principles and suited to a different type of experiment. Everybody can recall cases in which the same system has been described by either a stochastic model or a deterministic model, depending on the experiment planned.

For the sake of simplicity we shall restrict ourselves to deterministic primary models. Assume that the primary model connects three sets of variables $\boldsymbol{x}$, $\boldsymbol{w}$, and γ:

$$\boldsymbol{w} = \psi(\boldsymbol{x},\gamma) \tag{1}$$

where $\boldsymbol{x} \in R^k$, $\boldsymbol{w} \in R^l$ and $\gamma \in R^q$. Usually vector $\boldsymbol{x}$ is composed of control variables and ill-defined variables, while γ comprises variables whose values are known with relatively high precision. The way in which these groups are defined

will, of course, depend on the researcher.

We shall now explain some of the terms and notation used in the following sections. The result obtained by evaluating function (1) for given x_i and γ_i,

$$y_i = \psi(x_i, \gamma_i)$$

will be called the (*simulation*) *measurement*. The set $\xi_N = \{x_i\}_1^N$ will be taken to represent the *design* of an experiment, while the set $\Xi_N = \{y_i, \gamma_i, x_i\}_1^N$ will be defined as a (*simulation*) *experiment*. (We shall generally omit the word "simulation" in what follows.)

In most cases, the dimension of the output or response vector w is small while the dimensions of vectors x and γ can be as large as several hundreds. But (and this is one of the main assumptions) it is assumed that the responses y depend "strongly" upon only a few "significant" components of vector x.

The goal of simulation experiments is to identify these significant components (variables) and to construct some approximation $\eta(x)$ of the response function $\psi(x, \gamma)$. It should be emphasized that the function $\eta(x)$ does not depend on variables γ because the latter are assumed to be known relatively precisely.

2. STATISTICAL BACKGROUND

2.1. Screening Experiments

The aim of screening experiments is to detect the truly significant factors in a large collection of possibly significant factors (see, for instance, Li, 1962; Meshalkin, 1970; Satterthwhaite, 1959). To get an idea of the methods used, we shall consider one of the simplest approaches.

Suppose that we have

$$y_i = \vartheta_0 + \sum_{\alpha=1}^{m} \vartheta_\alpha x_{\alpha i} + \varepsilon_i, \qquad i = \overline{1,N} \quad ,$$

where the y_i are measurements, the ϑ_α are unknown parameters, $b_\alpha \le x_\alpha \le c_\alpha$, $\alpha=\overline{1,m}$, and the ε_i are independent random errors with zero mean and variance σ^2. It is assumed that s parameters ($s \le m$) are nonzero, where the value of s is known.

Consider a random design ξ_N constructed in the following way. Each measurement is carried out under random conditions x such that $x_{\alpha i} = b_\alpha$ or $x_{\alpha i} = c_\alpha$ with probability 0.5. Assume that $\hat{\vartheta}_{\alpha_1}, \ldots, \hat{\vartheta}_{\alpha_s}$ is the solution of the following extremal problem:

$$\text{Arg} \min_{\vartheta} \sum_{i=1}^{N} \left[y_i - \sum_{\alpha=1}^{m} \vartheta_\alpha q_\alpha x_\alpha \right]^2 , \Sigma q_\alpha = 1 , \tag{2}$$

where q_α can be zero or one. Let $P_N(\sigma^2)=1-\delta_N$ be the probability that the nonzero parameters have been identified correctly. Then

$$\lim_{\sigma^2 \to 0} \delta_N(\sigma^2) = \delta_N \le 2^{-N+s+\log_2(m-s+1)} . \tag{3}$$

For $N \ge m$ there are regular deterministic designs with the property

$$\lim_{\sigma^2 \to 0} \delta(\sigma^2) = 1 . \tag{4}$$

In other words, the use of such random designs makes it possible to reduce the number of measurements (which is usually essential) although the researcher pays for it with the resulting value of $P_N = 1-\delta_N < 1$ (compare (3) and (4)). Most of the efforts in the theory of screening experiments have been directed towards minimizing the number of observations N necessary under given δ_N.

2.2. Design of Regression Experiments

Consider the regression model

$$y_i = \eta(x_i,\vartheta) + \varepsilon_i = \vartheta^T f(x_i) + \varepsilon_i , \tag{5}$$

where ϑ is the vector of unknown parameters, $f(x)$ is the vector of basis

functions, the ε_i, $i = \overline{1,N}$ are random errors with zero means and variances of $E[\varepsilon_i^2] = \lambda^{-1}(x_i)$, and $x_i \in X$ where X is the operability region. The precision of the best linear unbiased estimator $\hat{\vartheta}$ of the parameters ϑ is determined by the information matrix:

$$M(\xi_N) = N\sum_{i=1}^{N} p_i \lambda(x_i) f(x_i) f^T(x_i) \quad ,$$

where $p_i = \tau_i / N$ and τ_i is the number of measurements necessary under condition x_i, such that $\sum_{i=1}^{N} \tau_i = N$. Recall that the covariance matrix $D(\hat{\vartheta})$ of the estimator $\hat{\vartheta}$ equals $M^{-1}(\xi_N)$. Then

$$\xi_N^* = \text{Arg} \min_{\xi_N} \Psi[M(\xi_N)]$$

is the optimal design for regression model (5) under criterion $\Psi(M)$. $\Psi(M)$ is usually a monotonic convex function of a positive semidefinite matrix M, for example, $\Psi(M) = \ln \det M^{-1}$ or $\Psi(M) = \text{tr } M^{-1}$.

2.3. Regression Analysis

This branch of statistics is composed of two main areas. The first is concerned with pure numerical problems, for instance, the extremal problem (2) or the following extremal problem:

$$\hat{\vartheta} = \text{Arg} \min_{\vartheta} \sum_{i=1}^{N} \lambda(x_i) \psi[\,|y_i - \eta(x_i, \vartheta)|\,], \tag{6}$$

where ψ is usually motonically increasing. The second area deals with the statistical properties of the estimators obtained using methods similar to (2) or (6).

In conclusion we should note that the areas of statistics described above are well developed in terms of both theory and available software.

3. STATISTICAL METHOD FOR SECONDARY MODEL CONSTRUCTION

It has already been pointed out that, although the values of variables γ are known more precisely than those of variables x, we still never know the exact values of the γ. If the problem is approached deterministically, then our knowledge of γ takes the form of a set of ranges $b_\alpha \leq \gamma_\alpha \leq c_\alpha$, $\alpha = \overline{1,q}$. Under the probabilistic approach, on the other hand, this information is given in the form of a distribution function $F(\gamma)$, which assigns a confidence level to each possible value of γ.

Because the exact values of γ_α are not available the researcher should really calculate the function $\psi(x,\gamma)$ using different sets of values $\gamma_{1j}, \ldots, \gamma_{qj}$ to see how it fluctuates. Obviously for high-dimensional γ much effort can be wasted in trying to consider every possible $\psi(x,\gamma_i)$; in practice it is generally sufficient to take the averaged behavior together with some confidence interval.

If the researcher is only interested in specific aspects of this averaged behavior (for instance, extreme points), stochastic optimization techniques can be used (see, for instance, Ermoliev, 1976). In cases where more detailed description is necessary, an approach based on the methods described briefly in Section 2 would be more appropriate.

The main steps in building a secondary model are summarized below.

(1) The variables included in the primary model are divided into two vectors, γ and x. The permissible sets for γ and x are ascertained ($\gamma \in \Gamma$, $x \in X$), and the possibility of calculating $\psi(x,\gamma)$ for points from $X \times \Gamma$ is considered.

(2) For each calculation (measurement), the values of γ_i are assigned using a random number generator with density function $F(\gamma)$, $\gamma \in \Gamma$. The structure of $F(\gamma)$ is usually chosen in accordance with the Bayesian approach. In

the simplest case a variable γ_α can take values b_α or c_α with equal probability. It should now be obvious that any measurement y_{ij} can be described by the following regression model:

$$y_{ij} = \eta(x_i) + \varepsilon_{ij} \quad , \tag{7}$$

where

$$\eta(x_i) = E[\psi(x_i,\gamma_j)] \, , \qquad E[\varepsilon_{ij}] = 0 \quad ,$$
$$E[\varepsilon_{ij}^2] = E\left\{[\psi(x_i,\gamma_j) - \eta(x_i)]^2\right\} = \sigma_{ij}^2 \quad ,$$
$$E[\cdots] = \int \cdots \mathrm{d}F(\gamma) \quad .$$

(3) It is, of course, unrealistic to hope to find $\eta(x)$ analytically in practice, but it may be possible to obtain a suitable approximation of $\eta(x)$. Very simple approximations are usually employed at this stage, e.g.,

$$\eta(x) \simeq \vartheta_0 + \sum_{\alpha=1}^{m} \vartheta_\alpha x_\alpha \tag{8}$$

where m can be several hundred. It is obvious that this approximation will be very rough. But then the goal is very modest: we only wish to identify the significant variables. In the simplest case the necessary calculations can be performed by two standard statistical programs: a procedure for generating random designs and another for stepwise regression (the latter should be present in any modern package of statistical programs). This yields the numbers $\alpha_1, \ldots, \alpha_s$ and the estimates $\hat{\vartheta}_\alpha$ of significant parameters. Here "significance" has its usual statistical meaning (broadly speaking, a parameter is significant if its estimated value is larger than its standard deviation). As pointed out elsewhere (see, for instance, Devyatkina and Tereokhin, 1981), the classical statistical methods for testing the significance of parameters (for instance, the F-test) are not appropriate to stepwise procedures; simple permutation tests should be used instead.

(4) It is now assumed that a comparatively small number (10-20) of significant variables are known from previous steps. Let $\tilde{x}^T = (x_{a_1}, \ldots, x_{a_s}) \in \tilde{X} \subset R^s$. In region $\tilde{X}$ we can use the more sophisticated approximation (compare with (8)):

$$\eta(x,\vartheta) \simeq \vartheta^T f(x) \quad .$$

The basis functions $f(x)$ are selected using some *a priori* information on the behavior of $\eta(x,\vartheta)$; a multidimensional second-degree polynomial $f^T(x) = (1, \tilde{x}_1, \ldots, \tilde{x}_s, \tilde{x}_1\tilde{x}_2, \ldots, \tilde{x}_s^2)$ often gives satisfactory results. If all variables are of equal interest then the simulation experiments should be carried out in accordance with D-optimal design, $\Psi = \ln \det M^{-1}$ (see, for instance, Fedorov, 1972). If there is some need for interpolation or extrapolation, then the design should minimize

$$\Psi = \max_{x \in X} w(x) d[\eta(\tilde{x},\hat{\vartheta})]$$

or

$$\Psi = \int_X w(x) d[\eta(\tilde{x},\hat{\vartheta})] dx$$

where $d[\eta(\tilde{x},\hat{\vartheta})] = f^T(\tilde{x}) M^{-1} f(\tilde{x})$ is the variance of the estimator $\hat{\vartheta}^T f(\tilde{x})$, and $w(x)$ is a weight function reflecting the researcher's interests. If the systematic discrepancy between $\eta(\tilde{x},\vartheta)$ and $\vartheta^T f(\tilde{x})$ is negligible then the function $(\lambda^{-1}(x) + f^T(\tilde{x}) M^{-1} f(\tilde{x}))^{1/2}$ represents the standard deviation of the forecast.

The problem of optimal design has received much attention and there are now catalogues of optimal designs for certain standard situations (see, for instance, Brodsky et al., 1982) as well as some rather advanced software.

(5) The final step involves the use of secondary models to analyze the system. The type of problems that can be solved are indicated in Figures 1-3. Figure 1 illustrates the possibility of testing or refining the primary model by

applying the secondary model to initial data. Figure 2 explains the possibility of combined analysis of two systems. Figure 3 illustrates how primary models can be compared through approximation by the same secondary model.

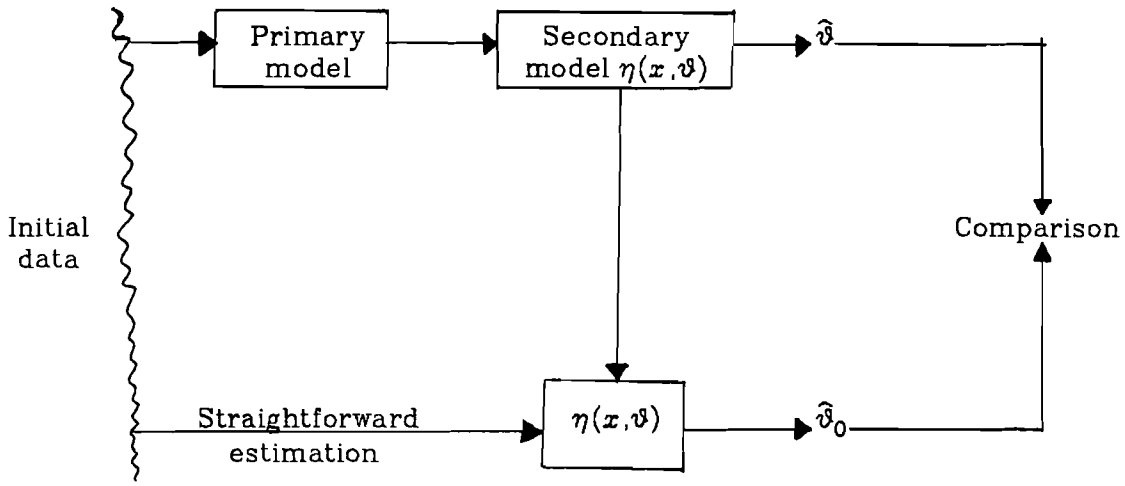

Figure 1. Testing a primary model by applying a secondary model to initial data.

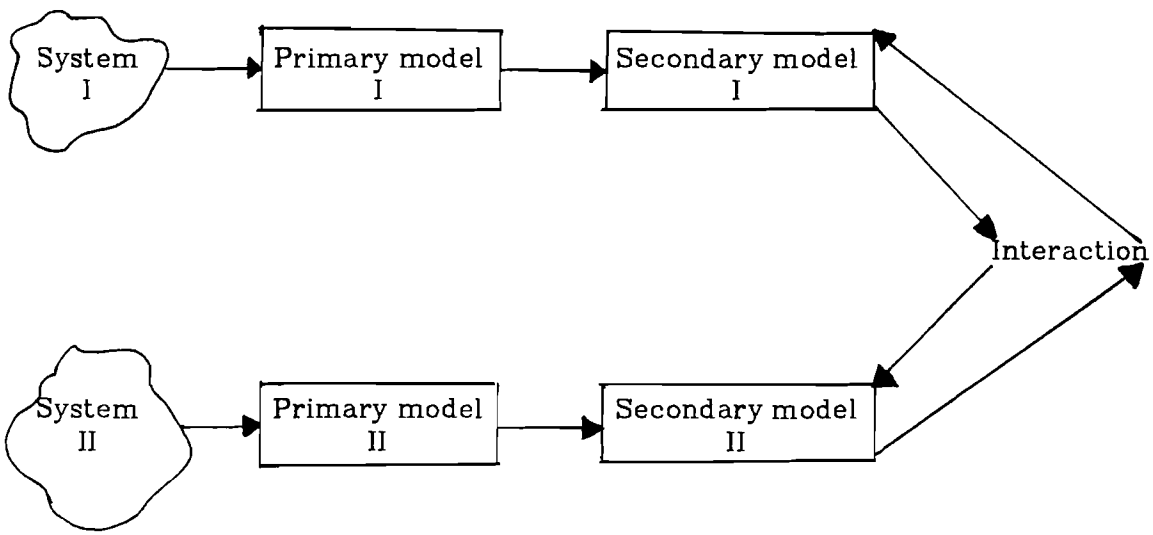

Figure 2. The combined analysis of two systems.

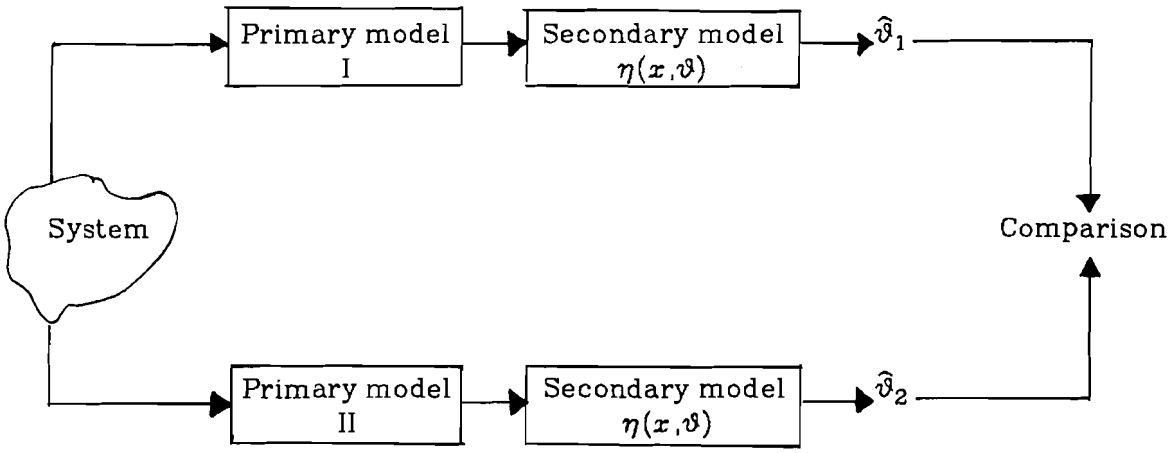

Figure 3. Comparison of primary models through approximation by the same secondary model.

4. EXAMPLE ,

To illustrate steps 1-4 in the procedure described above we shall consider a simple numerical example. The simplified version of the IIASA energy supply model MESSAGE-II was chosen as a primary model. This is a linear programming model:

$$\min C^T U$$
$$Au \leq b \ , u \geq 0 \ .$$

Twenty elements of matrix A were chosen as components of vector x. These are the first twenty entries in Table 1. Analysis of *a priori* information showed that all of these elements could vary in a 20% range. Ten other elements of matrix A were selected as components of vector γ and assumed to have a variation of 1%. The response function was $\psi(x,\gamma)=\min C^T u$.

The experiment was conducted using a two-level random design for both x and γ (see step 3 from the previous section). To improve the statistical properties of this design the randomization was carried out under the constraint that

Table 1. Components of x and γ.

1	uehh...a	te....1a	↑		
2	uehh...a	te....2a			
3	uehh...a	te....3a			
4	uehh...a	te....4a			
5	uehh...a	te....5a			
6	uehh...b	te....1b			
7	uehh...b	te....2b			
8	uehh...b	te....3b		Direct electric heating	The matrix elements represent the amount of electricity used
9	uehh...b	te....4b			
10	uehh...b	te....5b	x		
11	uehh...c	te....1c			
12	uehh...c	te....2c			
13	uehh...c	te....3c			
14	uehh...c	te....4c			
15	uehh...c	te....5c			
16	u0ii...a	t0....1a			
17	u0ii...a	t0....2a		Production of process heat in industry by an electric furnace	The matrix elements represent the amount of electricity used
18	u0ii...a	t0....3a			
19	u0ii...a	t0....4a			
20	u0ii...a	t0....5a	↓		
21	uohh...a	uh.....a	↑	Oil heating system	The matrix elements represent the amount of heat produced per unit of input
22	ughh...a	uh.....a		Gas heating system	
23	uenh...a	uh.....a		Electric night storage	
24	u2ii...a	ui.....a			
25	u2ii...b	ui.....b	γ	Oil furnaces for industrial process heat	
26	u2ii...c	ui.....c			
27	u8ii...a	ui.....a			
28	u8ii...b	ui.....b		Gas furnaces for industrial process heat	
29	u8ii...c	ui.....c			
30	yu.si..a	mu.si..b	↓	Market penetration constraint on industrial solar heat production systems	

the correlation between vectors x_j and x_k should be less than 0.1 for all j and k.

The number of measurements was determined from (3) with $m=20$, $s=6$, and $\delta=0.03$, and turned out to be 15. Of course, this number is only a rough estimate because the assumptions under which (3) holds are not completely fulfilled.

The results of the stepwise regression analysis are represented by the solid points in Figure 4, where

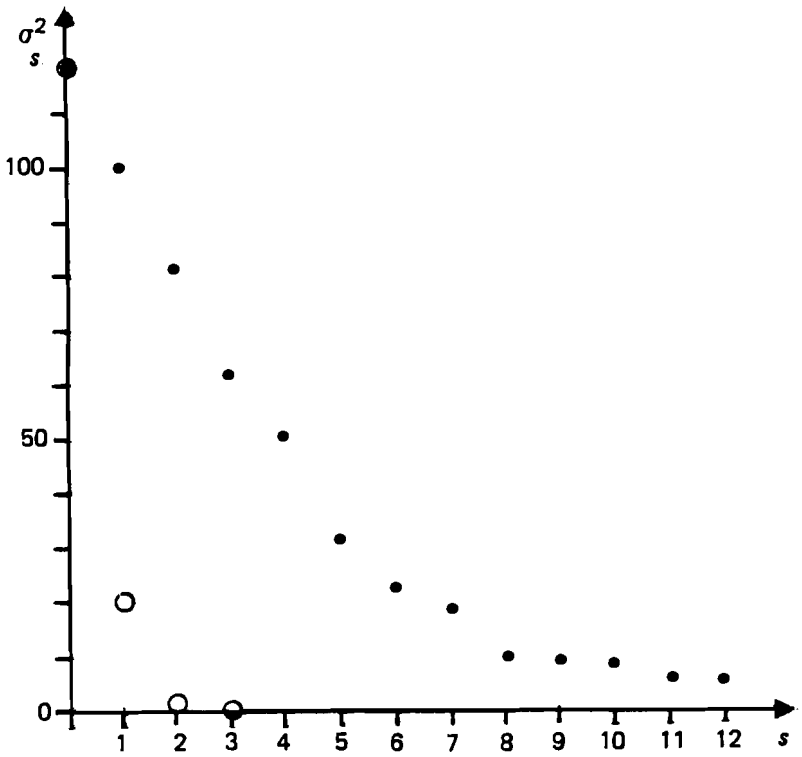

Figure 4. Results of the stepwise regression analysis.

$$\sigma_s^2 = \sum_{i=1}^{N} \left[y_i - \vartheta_0 - \sum_{\alpha=1}^{s} \vartheta_\alpha x_{i\alpha} \right]^2 / (m - s - 1)$$

is the simplest measure of the discrepancy between a response function $\eta(x)$ and its approximation. The value σ_0^2 can be considered as an estimate of the variability of the response function $\eta(x)$ (see (7)) within the operability region X.

The smooth decrease in σ_s^2 indicates that the contribution of every component of the vector x is comparable to the "noise" arising from the variation of γ. To avoid this, the stepwise regression analysis was repeated for both x and γ simultaneously. The open circles in Figure 4 illustrate the results of this analysis. The final estimates of the regression coefficients were ordered by

their absolute value: the largest are given in Table 2. These coefficients correspond to the scale $-1 \leq \mathrm{var}\, x_\alpha \leq 1$, $-1 \leq \mathrm{var}\, \gamma_\beta \leq 1$.

Table 2. Estimates of regression coefficients, ordered by their absolute value.

Variable	γ_2	γ_1	γ_3	x_{11}	x_{13}	x_7
Coefficient	-10.95	-4.11	-1.43	0.025	-0.022	0.020

It is clear that the most significant variables are γ_1, γ_2, and γ_3, despite the fact that they have a variation of only 1%. In other words, comparatively small changes in these variables provide the greatest contribution to the variability of the function $\psi(x,\gamma)$. This indicates that the main priority should be to evaluate these variables.

ACKNOWLEDGEMENTS

I am very grateful to E. Nurminski, S. Scherbor and M. Strubegger for their help in obtaining the numerical results, and also to H. Gasking for editing the paper.

REFERENCES

Brodsky, V.Z., L.I. Brodsky, T.I. Golikova, E.D. Nikitina, and L.A. Panchenko. (1982) *Tables of Optimal Designs*. Metallurgy, Moscow (in Russian).

Devyatkina, G.N., and A.T. Tereokhin (1981) Statistical hypothesis tests in stepwise regression analysis based on the permutation method. In V. Fedorov and V. Nalimov (Eds.), *Linear and Nonlinear Parametrization in Experimental Design Problems*, pp. 111-121. Part of the series *Problems in Cybernetics*, Moscow (in Russian).

Ermoliev, Y.M. (1976) *Stochastic Programming Methods*. Nauka, Moscow (in Russian).

Fedorov, V.V. (1972) *Optimal Design of Experiments*. Academic Press, New York.

Li, C.H. (1962) Sequential method for screening experiments. *Journal of the American Statistical Association*, Vol. 57, pp. 455-477.

Meshalkin, L.D. (1970) Validity of the random balance method. *Industrial Laboratory*, Vol. 36, pp. 316-318.

Satterthwhaite, F.E. (1959) Random balance experimentation. *Technometrics*, Vol. 1, pp. 111-138.