Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

DESIGN OF SIMULATION EXPERIMENTS

William E. Biles
Industrial Engineering Department
Louisiana State University
Baton Rouge, LA 70803

ABSTRACT

This paper describes the application of experimental design techniques to computer simulation. Three principal areas of experimental design are considered: (1) factor screening experiments; (2) experiments of comparison; and (3) response surface methodology.

INTRODUCTION

Simulation can be defined as the establishment of a mathematical-logical model of a system and the experimental manipulation of that model on a digital computer. This definition emphasizes two principal activities in computer simulation; (1) model development, and (2) experimentation. This paper concentrates on the second of these activities, and assumes that the simulationist has already developed a valid model of the system under study.

The simulationist attempts to utilize the simulation model to gain an understanding of the relationships between a set of system responses η_j, $j = 1, \ldots, m$ and a set of controllable factors x_i, $i = 1, \ldots, n$. These relationships take the form

$$n_j = g_j(x_1,\ldots,x_n),\ j = 1,\ldots,m \tag{1}$$

which are unknown to the simulationist. But by conducting a simulation trial at a point X^h, using a set of random number streams S^h, where X^h is the n-vector of values $(x_1^h, x_2^h, \ldots, x_n^h)$ and S^h is the p-vector of seeds $(S_1^h, S_2^h, \ldots, S_p^h)$, the simulationist is able to observe a set of time series $\{y_j^h(t)\}$, $j = 1, \ldots, m$, where $y_j^h(t)$ represents the measured value of the j-th response variable η_j at time t for the h-th simulation trial. Unlike physical experimentation, which typically involves setting the values of the controllable factors at X^h and directly observing an m-vector of physical values Y^h, simulation requires a judicious selection of the initial seeds S^h for the random number streams that are used to generate the various random processes embedded in the model, as well as a choice of the duration of the trial. The duration is typically either (a) a fixed number of realizations N of a given response η_j, (b) a fixed period of simulated time T, or (c) the achievement of a specified state of the system. The experimental design procedures discussed in this paper are generally applicable to any of these three approaches for choosing the duration of the simulation.

Now the observed value y_j of a given system response η_j as a result of a simulation trial has the form

$$y_j = g_j(x_1,x_2,\ldots,x_n) + \varepsilon_j,\ j = 1,\ldots,m \tag{2}$$

where ε_j has mean $E(\varepsilon_j) = 0$ and variance $\mathrm{Var}(\varepsilon_j) = \sigma_j^2$. That is,

$$y_j = \eta_j + \varepsilon_j,\ j = 1,\ldots,m \tag{2a}$$

Now y_j actually represents the mean of a time series of realizations $\xi_{j\ell}$, $\ell = 1,\ldots, r_j$, where r_j is the number of such realizations recorded during the simulation. That is,

$$y_j = \frac{1}{r_j}\sum_{\ell=1}^{r_j} \xi_{j\ell} \qquad j = 1,\ldots,m \tag{3}$$

The variance of this time series can be estimated by the relation

$$s_j^2 = \frac{1}{(r_j - 1)}\sum_{\ell=1}^{r_j} \xi_{j\ell}^2 - r_j y_j^2 \tag{4}$$

The estimates y_j and s_j^2 are unbiased estimates of η_j and σ_j^2, respectively, where σ_j^2 is the true variance of the response η_j. If the succession of realizations $\xi_{j\ell}$, $\ell = 1,\ldots,r_j$ are not independent, it is necessary to employ other formulae to compute the variance of this time series. Fishman [4] discusses techniques for doing this.

In the following sections, we shall restrict our attention to a single system response η as a function of the n-vector of controllable factors x_i, $i = 1, \ldots, n$.

FACTOR SCREENING EXPERIMENTS

Of the n controllable factors in a computer simulation model, $k \leq n$ of these are also controllable in the real-world system. In addition to these, there is also a set of n-k controllable factors in the model that represent uncontrollable parameters in the real environment, but the simulationist is also interested in determining the response of the system to changes in these uncontrollable factors. For instance, in a model of a naval engagement, ship speed and rate of antimissile fire might be factors that are controllable by the commander in the real system, whereas weather effects and rate of enemy missile fire are factors beyond the commander's direct control. But the simulationist would attempt to measure the effects of each of these factors on the system response, which might be "probability of victory." In the simulation model, all of these factors would be controllable.

In general, not all of the n factors are equally important with respect to their effect on the response η. In factor screening, we attempt to isolate those factors which are highly important from those which are negligible. If $g < n$ factors exert important effects on η, we seek to have an experimental design indicate which factors these are. Jacoby and Harrison [5] discuss these concepts. Montgomery [8] has produced an up-to-date treatment of this subject.

In factor screening, it is generally assumed that the relative importance of a set of n factors can be established by examining the coefficients β_i in the linear model

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \qquad (5)$$

To perform an experiment with the simulation model, we perform simulation trials at each of a set of settings of X which involve one or more levels of each of the n controllable variables x_i, $i = 1,\ldots,n$. The method of least-squares is then employed to estimate the main effects and interactions. From this analysis, the g most important factors are identified.

Some of the experimental designs employed in factor screening include the following:

- 2^n factorial experiments, involving a simulation trial at each of the $N = 2^n$ design points.
- 2^{n-p} fractional factorial designs, where n is large and 2^n simulations represent a very costly investment.
- Supersaturated plans, in which each of the n factors appears at high and low levels N/2 times, $N \leq n$.
- Groups screening designs, in which h groups of the n factors are identified, each such group is considered a single factor, and a 2^h factorial or 2^{h-p} fractional factorial design is employed to evaluate these group effects.

An important consideration in factor screening is that of variance reduction. Because simulation produces a times series of realizations ξ_ℓ, $\ell = 1, \ldots, r$ for the response η, where the time series is induced by a series of pseudorandom numbers, it is possible to reduce the variance of the time series by judicious selection of these pseudorandom numbers. Two well-known variance reduction techniques are as follows:

- Common pseudorandom numbers, where the same set of initial random number seeds S are employed for each simulation trial in the designed experiment.
- Antithetic pseudorandom numbers, where the series of random numbers for one stream R' is the complement of another stream R; that is, $r' = 1 - r$ for each successive pseudorandom number.

Fishman [4] discusses variance reduction in simulation experiments. Schruben and Margolin [10] describe a very useful techniques for pseudorandom number assignment in simulation experiments.

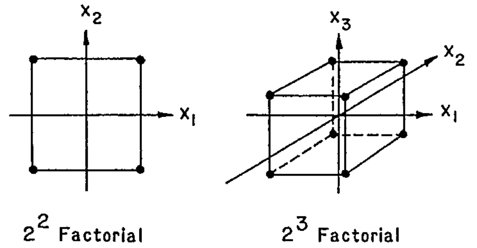
EXPERIMENTS OF COMPARISONS

Some of the n controllable factors are such that they assume quantitative levels in the experimental design. For example, ship speed and rate of antimissile fire are quantitative variables which can be set at selected levels over a continuum of values $a_i \leq x_i \leq b_i$. Other controllable factors are definitely qualitative in nature. For example, the sea state could be calm, high seas, or stormy. In many simulation situations, the simulationist seeks to compare the response η at one level of a controllable factor to that at a different level. Such evaluations are called "experiments of comparison." The controllable variables in such experiments are called <u>factors</u>, and the different levels of each factor are called <u>treatments</u>.

The principal experimental designs employed with experiments of comparison are as follows:

- 2^n factorial designs (Biles and Swain [2] or Montgomery [8]).
- 2^{n-p} fractional factorial designs (Box and Hunter [3]).

These designs are illustrated in Figure 1.



2^2 Factorial 2^3 Factorial

Figure 1: 2^n Factorial Designs

Biles and Swain [2] and Montgomery [8] discuss analysis of variance procedures by which the simulation results obtained from factorial designs are evaluated. These techniques enable the simulationist to test the null hypotheses that the individual factors exert no influence on the behavior of the system response η, or that two-factor interactions exert no effects. As with the factor screening experiment, it is necessary to adopt either common pseudorandom numbers or antithetic pseudorandom numbers to minimize the variance of the estimates.

RESPONSE SURFACE METHODOLOGY

Factor screening and experiments of comparison are not the only objectives the simulationist might have with respect to simulation experimentation. Often it is necessary to utilize the simulation model to attempt to find the optimum conditions for operating the system. These optimum conditions are here denoted as X^* and η^*.

The body of techniques by which one experimentally seeks an optimum set of system conditions is called response surface methodology. The following sections describe first and second order response surface methods as they relate to simulation experimentation.

First-Order Response Surface Methods

First-order response surface methods attempt to accomplish experimentally what the "method of steepest ascent" accomplishes computationally. From a current point X^k, a designed experiment is conducted (with a simulation trial at each design point) to estimate the gradient direction $\nabla g(X^k)$. Simulation trials are then conducted at points along this direction to a new point X^{k+1} which represents the best solution obtained along the direction $\nabla g(X^k)$. This process is an experimental approximation of the relation

$$X^{k+1} = X^k + \lambda^k[\nabla g(X^k)] \tag{6}$$

The step length λ^k can be estimated by a line search or by a regression procedure as described by Biles and Swain [1,2].

The gradient direction $\nabla g(X^k)$ is estimated by placing an appropriate first-order experimental design, such as a w^n factorial, w^{n-p} fractional factorial, or n-dimensional simplex design (Biles and Swain [2]) around the current point X^k. A simulation trial is performed at each point in the selected experimental design. From these N observations the multiple linear regression model

$$\hat{y} = b_o + \sum_{i=1}^{n} b_i x_i \tag{7}$$

can be estimated. Since the gradient direction $\nabla g(X^k)$ is mathematically defined as the n-vector of first partial derivatives of g(X) evaluated at X^k, it is clear that $\nabla g(X^k)$ is simply the n-vector of regression coefficients, exclusive of the b_o term; that is,

$$\nabla g(X^k) = (b_1, \ldots, b_n)' \tag{8}$$

In the multiple-response simulation problem, a simulation trial is conducted at each design point in the selected first-order design and the m observations y_j^ℓ, j = 1,...,m are recorded at each design point. Multiple linear regression is applied separately to each set of observations (assuming independence among the m responses), producing the m models

$$\hat{y}_j = b_{j,0} + \sum_{i=1}^{n} b_{j,i}x_i, \quad j = 1,\ldots,m \tag{9}$$

and hence the m gradient vectors

$$\nabla g_j(X^k) = (b_{j,1}, \ldots, b_{j,n})', \quad j = 1,\ldots,m \tag{10}$$

These estimates can then be employed in any one of several optimization schemes to produce an improved solution X^{k+1}. A generalized procedure for accomplishing this improved solution, and an estimated "optimum," will be described later. But first it is necessary to give attention to the experimental designs employed to estimate the gradient vectors $\nabla g_j(X^k)$, j = 1,...,m.

In selecting a first-order response surface design, it is usually desirable to minimize the variances of the regression coefficients b_i, i = 1,...,n. To accomplish this the first-order experimental design should be orthogonal. An orthogonal first-order experimental design is constructed as follows: The placement of the N experimental points (in our case, simulation trials) is described by the N by n design matrix D, where

$$D = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_{1N} & x_{2N} & \cdots & x_{nN} \end{bmatrix} \tag{11}$$

An N by n+1 matrix X is then formed by placing a unit vector to the left of D. Thus

$$D = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{n1} \\ 1 & x_{12} & x_{22} & \cdots & x_{n2} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & x_{1N} & x_{2N} & \cdots & x_{nN} \end{bmatrix} \tag{12}$$

It is usually convenient to code the design levels, so that the following conditions are achieved:

$$\left.\begin{aligned} \sum_{u=1}^{N} x_{iu}^2 &= N \\ \sum_{u=1}^{N} x_{iu} &= 0 \end{aligned}\right\} \quad i = 1,\ldots,n \tag{13}$$

If the actual value of the u-th level of the i-th variable is z_{iu}, then the corresponding coded value is

$$x_{iu} = \frac{z_{iu} - \bar{z}_i}{s_i} \tag{14}$$

where

$$\bar{z}_i = \sum_{u=1}^{N} z_{iu} / N \tag{15}$$

and

$$S_i = \sum_{u=1}^{N} (z_{iu} - \bar{z}_i)^2 / N \tag{16}$$

Then

$$0\ X'X = \begin{bmatrix} N & 0 & 0 & \cdots & 0 \\ 0 & N & 0 & \cdots & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & 0 & \cdots & N \end{bmatrix} \tag{17}$$

Since the (n+1) - vector of regression coefficients $\bar{b}$ is estimated by the least squares relation

$$\bar{b} = (X'X)^{-1}\ X'\ \bar{y} \tag{18}$$

the condition scheme in (14) simplifies to $\bar{b} = N^{-1}X'\bar{y}$, where $\bar{y}$ is the N-vector of response estimates obtained from N simulation trials. The variance of the regression coefficients b_i, $i = 1,\ldots,n$ is given by

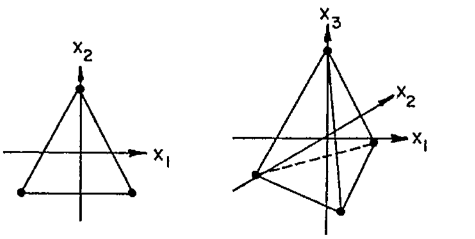
$$\mathrm{Var}(b_i) = \sigma^2/N, \quad i = 1,\ldots,n \tag{19}$$

where σ^2 is the variance of the error term ε. Since we are interested in m separate system response y_j, $j = 1,\ldots,m$, equations (18) and (19) can be generalized to

$$\bar{b}_j = (X'X)^{-1}\ X'\ \bar{y}_j, \quad j = 1,\ldots,m \tag{20}$$

$$\mathrm{Var}(b_{ji}) = \sigma_j^2/N, \quad i = 1,\ldots,n; \quad j = 1,\ldots,m \tag{21}$$

Again, with the coding scheme in (14), equation (20) simplifies to $\bar{b}_j = N^{-1}X'\bar{y}_j$. For an orthogonal first-order design, the results in (17)-(21) hold, giving a so-called "minimum-variance" design. The 2^n factorial and 2^{n-p} fractional factorial designs are orthogonal and hence minimum variance. Orthogonal n-simplex designs can be easily constructed (see Biles and Swain [2]). Since n-simplex designs provide the minimum number of design points needed to estimate the multiple-linear regression models in (7) or (9), and are hence the most "economical" of the first-order response surface designs, they are especially attractive for the purpose proposed here. Figure 2 illustrates n-simplex designs.



Two-Dimensional Simplex Three-Dimensional Simplex

Figure 2: n-Simplex Designs

Biles and Swain [1,2] have described a first-order response surface procedure for approaching the constrained formulation of the multiple-response simulation problem. This procedure involves performing a first-order design around a current point X^k to estimate the gradient direction $\nabla g(X^k)$ according to relation (8). A line search is then performed along $\nabla g(X^k)$ to estimate an optimal step λ in (6). As long as the search remains interior to the region bounded by the constraints, the procedure is basically a gradient search. If one or more constraints are encountered, however, Biles and Swain [1,2] propose that the gradient projection direction be followed. The procedure for estimating the gradient projection direction is as follows.

Suppose that at an estimated boundary point X^k, q constraints are satisfied as equalities. Let B_q be the n x q matrix of first partial derivatives of these active constraints. Thus B_q consists of the q gradient vectors $\nabla g_j(X^k)$, $j = 1,\ldots,q$. That is

$$B_q = \begin{bmatrix} \partial g_1/\partial x_1 & \cdots & \partial g_q/\partial x_1 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \partial g_1/\partial x_n & & \partial g_q/\partial x_n \end{bmatrix} \tag{22}$$

Since $g_j(X)$, $j = 1,\ldots,q$ denotes the set of binding constraint functions, for the moment let f(X) represent the objective function. Then $\nabla f(X^k)$ and $\nabla g_j(X^k)$, $j = 1,\ldots,q$ represent the gradient vectors of the objective and constraint functions, respectively, evaluated at the boundary point X^k.

Performing a first-order response surface experiment about the boundary point X^k yields estimates of the gradient vectors $\nabla f(X^k)$ and $\nabla g_j(X^k)$, $j = 1,\ldots,q$ in the form of the vectors of regression coefficients. The gradient projection direction is then given by

$$S^k = [\nabla f(X^k)] - B_q(B'_qB_q)^{-1}B'_q\ [\nabla f(X^k)] \tag{23}$$

A line search is performed along direction S^k until either (a) a local "optimum" is found, or (b) other constraints are encountered. This new point is denoted X^{k+1}. This procedure is repeated until the gradient projection direction S^k is approximately zero. This point X* is taken as a "constrained optimal" solution. Figure 3 illustrates the application of the gradient projection procedure to a constrained optimization problem.

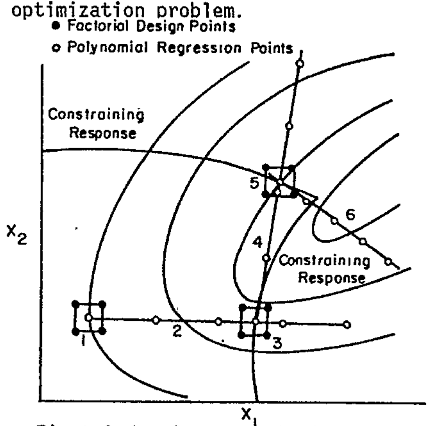


Figure 3: First-Order Response Surface Optimization

The following generalized procedure is followed in employing a first-order response surface approach to the multiple-response simulation problem. The particular problem formulation and optimization procedure will govern the precise sequence of steps in implementing this procedure.

1. Identify the known experimental region $a_i \le x_i \le c_i$, $i = 1,\dots,n$. Select a starting point X^0 within this region. With X^0 as its center, array an orthogonal first-order response surface design within a selected design radius. Place $n_c = n/2 \ge 2$ points at the design center X^0 (coded as a $\bar{0}$ - vector).

2. Perform simulation trials at each of the N experimental design points and record the responses y_j^ℓ, $j = 1,\dots,m$; $\ell = 1,\dots,N$. Using multiple linear regression, fit linear models of the form (9).

3. Apply the appropriate mathematical programming technique to locate the next center point in the search.

4. Repeat steps 1-3 until an "optimum" solution is located. It may be appropriate to add design points to complete a second-order response surface design to test this optimum solution. The procedure for accomplishing this is described in the next section.

Second-Order Response Surface Methods

A second-order response surface approach to the multiple-response simulation problem consists of one ore more repetitions of a two-stage procedure: (a) the execution of a computer simulation trial at each point in a second-order response surface experimental design covering the known experimental region, and the use of multiple linear regression to fit second-order regression models to the resulting data; and (b) the application of a suitable mathematical programming procedure to obtain a solution to the resulting optimization problem. In contrast to the first-order methods, in which the optimization procedure was part and parcel with the experimental procedure, these procedures are distinct and sequential in the proposed second-order approaches.

The first step in the second-order approach is to identify the range of each input variable. A safe strategy is to cover the entire known region $a_i \le x_i \le c_i$, $i = 1,\dots,n$ with the first (and possibly only) experimental design. If we let α_i denote the radius of the n-dimensional hypersphere within which the design points are contained, then

$$\alpha_i = (c_i - a_i)/2, \quad i = 1,\dots,n \tag{24}$$

is effectively the maximum radius we could construct. It is convenient to adopt the coding convention expressed in (14)-(16), but choosing x_{iu} in such a way that α_i satisfied (24). Biles and Swain [2] describe this coding process.

The second-order fitted response surface has the form

$$\hat{y} = b_o + \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n} b_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} x_i x_j \qquad i \neq j \tag{25}$$

where $\hat{y}$ is the estimate of the true response η at a given value $X = (x_1,\dots,x_n)$ and the b_i and b_{ij} are regression coefficients in the fitted model. Since we must estimate $\underline{m}$ separate response relationships, equation (25) is modivied to

$$\hat{y}_k = b_{k,o} + \sum_{i=1}^{n} b_{k,i} x_i + \sum_{i=1}^{n} b_{k,ii} x_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} b_{k,ij}\, x_i x_j \qquad i \neq j \qquad k = 1,\dots,m \tag{26}$$

Given the independence of the $\underline{m}$ responses, these $\underline{m}$ regression equations can be estimated independently from a set of $N \ge (n+1)\,(n+2)/2$ data points obtained by performing a simulation trial at each point in a second-order response surface design.

An experimental design employed for the purpose of estimating the regression coefficients in (26) must contain at least as many design points as there are coefficients b_i and b_{ij} in the fitted model, of which there are $(n+1)\,(n+2)/2$. Because of the nonlinearity of (26), the experimental design must also have at least three levels of each controllable variable x_i, $1 = 1,\dots,n$. It is also desirable to have a design which is rotatable; that is, the predicted response y at some point X is a function only of the distance from the design center to X and not a function of the direction.

The most widely used design for fitting a second-order model is the central composite design, shown in Figure 4 for n=2 and n=3. These designs consist of a 2^n factorial (or suitable fraction thereof), augmented by wn axial points and k center points. A central composite design can be made rotatable by proper choice of α, the distance of the axial point from the design center. With the proper choice of the number of center points k, the central composite design can be made either orthogonal or uniform precision.

Having estimated the $\underline{m}$ second-order regression equations (26) and formulated the appropriate optimization problem, it remains to apply mathematical programming to obtain a solution. For the constrained formulation, any of the following procedures could be employed: (a) Box's complex search; (b) Rosen's gradient projection method; or (c) one of Zoutendijk's methods of feasible directions. These are described in Biles and Swain [2].

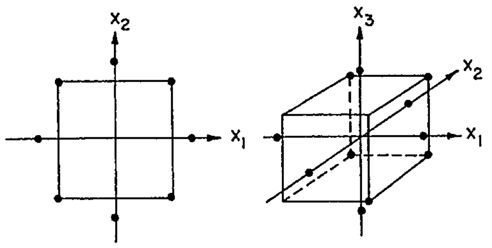


Figure 4: Central Composite Designs

BIBLIOGRAPHY

1. Biles, W. E., and J. J. Swain, "Mathematical Programming and the Optimization of Computer Simulations," in Mathematical Programming Studies on Engineering Optimization (R. S. Dembo and M. Avriel, Editors), North-Holland Publishing Company, Amsterdam, 1979.

2. Biles, W. E., and J. J. Swain, Optimization and Industrial Experimentation, Wiley-Interscience, New York, 1980.

3. Box, G. E. P., and J. S. Hunter, "The 2^{k-p} Fractional Factorial Designs," Parts I and II, Technometrics, Vol. 3, 1961, pp. 311-352 and pp. 449-458.

4. Fishman, G. S., Concepts and Methods in Discrete Event Digital Simulation, John Wiley and Sons, New York, 1973.

5. Jacoby, J. E., and S. Harrison, "Multivariable Experimentation and Simulation Models," Naval Research Logistics Quarterly, Vol. 9, 1962, pp. 121-136.

6. Kleijnen, J. P. C., Statistical Techniques in Simulation, Parts I and II, Marcel Dekker, New York, 1975.

7. Law, A. M., and W. D. Kelton, Simulation Modeling and Analysis, McGraw-Hill, New York, 1982.

8. Montgomery, D. C., Design and Analysis of Experiments, John Wiley and Sons, New York, 1976.

9. Montgomery, D. C., "Methods for Factor Screening in Computer Simulation Experiments," Final Report, Office of Naval Research, Contract N0014-78-C-0312, Georgia Institute of Technology, Atlanta, GA, 1979.

10. Schruben, L. W., and B. H. Margolin, "Pseudorandom Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments," Journal of the American Statistical Association, Vol. 73, No. 363, 1978, pp. 504-520.

11. Shannon, R. E., Systems Simulation: The Art and Science, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.