EXPERIMENTAL OPTIMIZATION OF STATISTICAL SIMULATION

David L. Eldredge

University of Evansville

ABSTRACT

Although simulation involves solving a mathematical model through experimentation, the literature of simulation does not reflect a broad, sustained interest in the design of simulation experiments. The objective of this paper is to fulfill a part of this need by suggesting a general five-phase experimental approach for determining the optimum for a particular, but common, type of simulation problem. The majority of the suggested procedure makes use of existing design and analysis techniques. However, the problem of multimodal simulation outcomes has resulted in the development of the "random factorial" experimental design. This design is a combination of a complete factorial design and a random balance design. The majority of the paper is devoted to a discussion of the use of this design approach for solving the multimodal problem.

OPTIMIZATION PROBLEM

The result of an evaluation run of a simulation model is the determination of a numerical value for one or more simulation outcome variables. As pointed out by Box and Hunter (2), one of the two possible objectives of a simulation study is to determine the set of values for the decision variables which yields the optimum value for a particular outcome variable. At times the complete enumeration of all possibilities might be used to accomplish this objective. However, oftentimes complete enumeration is not feasible. This latter situation is that considered in this paper. Furthermore, the paper will be concerned with static, statistical simulations. The decision variables are assumed to be quantitative, and the response surface of simulation outcomes to possibly be multimodal.

A simulation involves solving a symbolic model by obtaining numerical outcome values. Thus, it is not inherently optimizing. To solve the optimization problem posed here, one must superimpose optimization upon the model by varying the decision-variable values in search of the optimum outcome. A procedure which specifies the manner for performing such a search is suggested in the following section.

OPTIMIZATION PROCEDURE

The search for the optimum outcome of a response function involves the three distinct, but interrelated, sub-problems of (1) investigating the global properties of the response function, (2) investigating the local properties of the response function, and (3) identifying the optimal solution. Accordingly, the first part of our optimization procedure involves experimentation over the entire solution space in order to establish global properties of the response function. The objective of this experimentation is to transform the original optimization problem into a number of smaller and more manageable problems.

An obvious first step toward the accomplishment of this objective is to determine whether all the decision variables included in the simulation affect the response-function values significantly. If one can identify some decision variables as having an unimportant effect, one can then set the value of these variables equal to some nominal value and treat them as deterministic parameters throughout the remainder of the optimization. Such a reduction in the dimensionality of the optimization problem can greatly reduce the number of evaluation runs required. Thus, Phase I of the optimization procedure might be called "Determination of the Effective Decision Variables."

There are a number of optimization techniques available which are capable of effectively locating the optimum of a response surface which contains either a single peak or a saddlepoint. Consequently, the establishment of a second global property, the existence of a peak or saddlepoint within a solution subspace would be advantageous. Such a solution subspace will be called a "locally explorable" subspace.

Accordingly, the objective of the second phase of the procedure is to divide the total solution space into a number of subspaces such that the response function over each subspace is locally explorable.

Next the second sub-problem of investigating the local properties of each of the subspaces can be considered. The objective here is to establish an approximation to the response function in the vicinity of the local optimum of each subspace. Therefore, one must first estimate the location of the local optimal solution point, Phase III of the optimization procedure, and then determine a suitable approximating function, Phase IV.

Phase V of the procedure is to consider the third sub-problem, that of identifying the global optimal solution point. In this phase, a figure of merit is determined for each of the local optima in order to establish which of them is the preferred operating point. Thus, the optimization problem previously posed will have been completely solved.

For this procedure to be made operational, it is necessary to identify appropriate experimental designs and analysis procedures for satisfying the requirements of each of these five phases. A review of the literature (see (6), Chapter II) resulted in the identification of appropriate designs and analysis procedures for four of the five phases. However, for Phase II, the literature search resulted in the conclusion that the problem of a multimodal response function is a difficult aspect of experimental optimization. The detection of some measure of the multimodality of a response function requires an experiment with four or more levels for each variable, and as pointed out by Cochran and Cox (4, p. 273), "Experiments with all factors at four levels do not appear to be common." This statement appears to be true also for experimental designs for variables involving more than four levels. As a result of this finding, a new form of experimental design, called a random factorial design, is suggested in the following section.

EXPERIMENTAL DESIGNS FOR PHASE II

The concept of random factorial experimental designs is based on the Random Balance Experiment proposed by F. E. Satterthwaite (11) and on a suggestion for its use made by Budne (3, p. 141). The need for this new design approach arises because the commonly used experimental designs such as complete and fractional factorial designs require an unreasonably large number of evaluation runs for Phase II of the optimization procedure. As this implies, the criterion used here in judging the effectiveness of designs for Phase II is the number of evaluation runs required. In addition, a second general requirement imposed upon the Phase II designs is that they must provide unconfounded estimates for at least all the main and two-factor interaction effects. Although justifications can be found for neglecting interactions involving three or more variables (e.g., see (1, p. 313), (7, p. 91 and 138), (10, p. 306), and (12, p. 459)), it is generally recognized that effects involving less than three must be explicitly considered.

To begin this discussion consider the application of complete and fractional factorial designs to Phase II. The first observation one can make for these designs is that no fractional design is available for simulations involving two, three, or four decision variables if one is to obtain unconfounded estimates of all main and two-factor interaction effects. Consequently, the number of evaluation runs required for problems with two, three or four decision variables corresponds to all the possible factorial combinations, that is 16, 64 and 256 respectively.

For problems with more than four decision variables, fractional factorial designs become feasible. Although four-level fractional factorials are not generally discussed in the literature one can derive one-half replicate designs for five and six decision variable problems, and a one-fourth replicate design for seven variable problems. Accordingly, the numbers of runs for these fractional designs for five, six, and seven variable problems are 512, 2048, and 4096, respectively.

Although we have required the minimum number of levels and the minimum number of unconfounded effects, the number of runs required by these designs makes Phase II impractical for many simulations. Of course, if we require a greater number of levels and/or additional unconfounded effects, the numbers become even more demanding of our simulation resources. Therefore, it is desirable to develop an experimental-design approach such as random factorial designs which requires a smaller number of evaluation runs.

A random factorial design is a combination design made up of a complete factorial design and a random balance design. In essence, a random balance design is simply one for which (1) the values of each decision variable are selected through the use of some random process, and (2) the random selection process used for each decision variable is independent of the values selected for all the remaining decision variables.

A "complete random factorial" design is made up of a number of fractional designs, that is, "fractional random factorials," just as a complete factorial design is made up of a number of fractional factorials. For each of the fractional random factorials within a complete design, a subset of the variables form a complete factorial design while each of the remaining variables is maintained at some constant value. Thus, we define two categories of variables for each fractional factorial: the factorial variables and the random balance variables. The former category includes all those variables which form the complete factorial design within a fractional random factorial, and the latter category all the remaining variables whose values are selected such that together they form a random balance design. That is, the single value for each of these random balance variables is selected so as to satisfy the two aspects of the previously specified definition of a random balance design.

The condition, or assumption, upon which the validity of our random factorial approach is based is that all interaction effects involving more than some number of variables, say k, are negligible. To clarify, consider an experimental design problem involving n design variables with L levels each. Suppose for this problem, it is known that the significant interaction effects involve at the most k variables where $0 < k < n$. This means that the problem could be analyzed using the results of experimental designs which yield unconfounded estimates for only those effects involving k factors or less. Thus, a single k-factor complete factorial would provide some of the required estimates, but only those involving the particular set of k factors included in the design. Estimates of any of the effects involving the remaining (n - k) variables would have to be found from the results of additional designs. Accordingly, further k-factor complete factorials might be used to provide these estimates.

Moreover, we observe that, if it is necessary to estimate all the k-factor interaction effects, the number of individual k-factor complete factorials required is equal to the number of all possible combinations of n items taken k at a time, say C(n,k). The experimental results from all these C(n,k) designs would yield one estimate for each of the k-factor interaction effects, and at least one estimate for all the effects involving less than k variables. If, in addition, we require that the (n - k) variables not included in each k-factor complete factorial satisfy the two conditions of random balance, then each of these C(n,k) complete factorials will have associated with it a pure random balance design in (n - k) variables. As previously stated, this combination of a complete factorial design and a pure random balance design is what we term a "fractional random factorial." The totality of all these C(n,k) fractional random factorials is called here a "complete random factorial design."

The number of evaluation runs necessary for this design approach is significantly reduced from the number required by a constant balance designs. For example, the number of runs for a four-level, seven-variable problem with k equal to two is reduced from 4096 to 336.

ANALYSIS OF PHASE II DESIGNS

? Mueller

In order to accomplish the objective of Phase II, that is, isolate locally explorable solution subspaces, one must perform a statistical analysis of the simulation outcome values obtained for a Phase II experimental design. In essence, the objective of this statistical analysis is to test for differences in the simulation outcome at different solution points. As demonstrated by research results reported by Naylor, Wertz, and Wonnacott (9), the analytical tool which is most compatible with such an experimental objective is the F-test of an analysis of variance. Moreover, the specific form of the analysis of variance which is of interest here is called the "single-degree-of-freedom" approach. This analysis technique, which is applicable to problems involving quantitative decision variables with equi-spaced levels, provides one with an indication of the shape of a response function. Consequently, through the application of it to Phase II experimental results, one is able to ascertain statistically whether a function is locally explorable over a specified solution subspace.

The technique for getting the sum of squares necessary to conduct a single-degree-of-freedom analysis of variance is to decompose the sum of squares associated with the complete effect. This can be accomplished by applying a special set of contrasts called orthogonal polynomials to the experimentally derived outcomes (see (4), (5), (7) or (8)). The result one derives from the use of these are sums of squares which correspond to estimates of appropriate single-degree-of-freedom effects of a decision variable. Each main effect is decomposed into (L-1) effects where L represents the number of decision-variable levels. These include a linear effect, a quadratic effect, a cubic effect, and so on up to a final effect which has an order of (L-1). Correspondingly, each two-factor interaction effect is divided into the $(L-1)^2$ individual two-factor effects, and so on.

The number of single-degree-of-freedom effects into which the total effect of a decision variable can be decomposed depends only upon the number of levels one considers for that decision variable. The greater the number of levels selected for a variable, the greater the number of effects which may be tested, and the more assurance one has that all neglected effects are actually negligible. Of course, the number of evaluation runs undergoes a corresponding increase. Thus, the number of levels selected for each variable must be based on two conflicting considerations: (1) accuracy of representation by the analysis model, and (2) resources available for making evaluation runs. Although a general discussion of these considerations cannot be given, there is some justification for recommending that only four, or at the most five, levels should be utilized for each variable.

Having established the general analysis procedure, we may now consider the single-degree-of-freedom analysis of variance specifically for the random factorial designs of the previous section. It will be recalled that each of these designs is made up of C(n,k) fractional random factorial designs, where k equals the number of factorial variables in the design. In turn, each of the fractional designs is made up of a complete factorial design in k variables and a single level for the (n - k) remaining variables.

A derivation of the expected mean square expressions (see (6)) for an n-factor fractional random factorial yields hypothesis tests for all the single-degree-of-freedom main and interaction effects. The experimental outcomes of each fractional random factorial provide statistical tests of all effects involving the k factorial variables, that is, all effects up to and including the k-factor interactions. Each of the n decision variables appears as a factorial variable in C(n-1,k-1) of the fractional parts of a complete random factorial, so the combined analysis of all the fractional parts provides the required statistical tests of all effects involving k or fewer variables.

ITERATIVE PROCEDURE FOR PHASE II

Many procedures could be formulated for iteratively applying Phase II designs in order to define a set of solution subspaces over which the response function of a statistical simulation is locally explorable. The one we suggest is a sequential halving process which resembles Bolzano's method for finding a root of an equation.

To begin the suggested approach, evaluation runs are made over the entire solution space based on an appropriate random factorial design. Next, these experimental results are analyzed by the techniques of the previous section. If this analysis reveals that some of the decision variables exhibit significant cubic or higher-order effects, we continue to the second stage of the procedure. The second stage involves dividing into halves the full ranges of those variables which have shown a significant cubic or higher-order effect. These half-ranges, together with complete ranges of the undivided variables, define a number of smaller subspaces within the original solution space. To continue, we make the necessary additional evaluation runs over each of these subspaces based on the appropriate random factorial design, and again carry out the random factorial analysis. If significant cubic or higher effects are found for any of these subspaces, they are further sub-divided. The process is continued until we obtain an "adequately" fitting, piece-wise quadratic equation over a number of solution subspaces.

At each iteration, or stage, of this procedure, a number of levels are specified for each of the decision variables. In general for the first stage, L levels are specified for each of the n decision variables. As previously discussed, L must be greater than three, and the levels must be equally spaced. At Stage Two, the range of each of the variables having a significant cubic or higher-order effect is divided into half-ranges. For each half-range, L equally spaced levels are specified. One of these levels for each half is coincident with the center of the original full range. Accordingly, we have a total of $(2 \cdot L-1)$ levels defined over the full range of each of the divided variables.

Since the levels are equally spaced, not all of the $(2 \cdot L-1)$ levels are new at the second stage. In fact, only (L-1) represent new values, and the remaining L levels are those specified for the first stage. Furthermore, each of the (L-1) new levels bisects one of the intervals between two adjacent levels from the first stage. It is observed that the iterative procedure we have presented has an additional analytical benefit. Namely, after the division of the range of a variable into two equal parts and the addition of (L-1) new levels to the variable, there exist $(2 \cdot L-1)$ equally-spaced levels over the full, undivided range. Consequently, before the second-stage analysis of the experimental outcomes is performed, it is possible to conduct a $(2 \cdot L-1)$-level analysis over the entire range. This supplemental analysis provides a check on the results one has obtained for the first stage. Moreover, such a $(2 \cdot L-1)$-level analysis is possible, not only between the first and second stages, but also between the second and third, third and fourth, and any other two successive stages. In general, any sub-range of a variable which is divided into two equal parts at a stage s, based on the results of an L-level experiment, can be re-examined after the experimental outcomes for stage (s+1) are obtained.

Additional considerations for the Phase II procedure such as (1) the specification of the solution points which must be evaluated at each stage of the process, (2) the determination of the cost of experimentation, (3) the procedure for combining the analysis results of all the fractional designs of a complete random factorial, and (4) the limitations imposed by the possibility of statistical errors are considered elsewhere (6, Chapter V). However, for this paper it seems appropriate to turn our attention to a consideration of an example problem.

PHASE II EXAMPLE

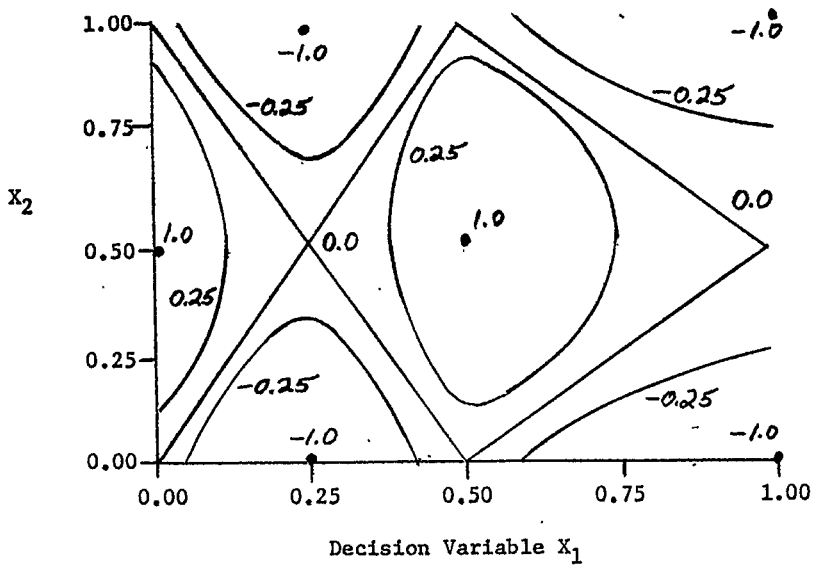
In order to illustrate the experimental design and analysis procedure suggested here, seven multimodal example problems have been examined. These examples treat a number of the common characteristics an experimenter is likely to encounter in practice such as (1) response surfaces composed of ridges, peaks, and/or saddlepoints, (2) response surfaces with non-interacting or interacting independent variables, (3) response surfaces with extreme slopes or gradual slopes, (4) continuous and discontinuous response surfaces, and (5) the effect of Type I errors and of Type II errors.

All seven examples were for the smallest random factorial designs which have been discussed, that is, for four-level, two-decision variable problems. Two reasons account for this selection. First, for problems involving three dimensions (i.e., two decision variables together with the dependent simulation outcome), the true response surface can be displayed pictorially so one can follow the progress of the iterative procedure. Second, the two-variable procedure is that which is used for analyzing the results of all n-variable random factorials with two factorial variables. These four-level, two-factorial-variable designs are probably the most useful form.

In this section we will illustrate the Phase II procedure through the presentation of some of the details of one of these examples. A representation of the contours of constant expected values for this example is shown in Illustration 1. As indicated, this example has two local maxima, one saddlepoint, and a number of local minima. The results one would expect from our Phase II procedure are that two locally explorable subspaces would be identified at the close of a second-stage analysis.

ILLUSTRATION 1

Expected-outcome Contours for Example Seven



Since we are concerned with two, four-level decision variables, the number of first-stage evaluation runs is sixteen. These include all the factorial combinations for the levels 0.0, 1/3, 2/3, and 1.0 for both decision variables. The corresponding experimental outcomes were derived from the response surface equation and randomly selected error values. The analysis of the resulting sixteen outcomes begins with the determination of the sum of squares for the two main and the one interaction effect.

The next step in the procedure is to apply orthogonal polynomials to these data in order to find the single-degree-of-freedom components of the main and interaction effects. The results of this step are shown in Table 1. Since the sum of squares for interaction effect for the results obtained is zero, its single-degree-of-freedom effects are not designated.

TABLE 1

Stage One Single-degree-of-freedom Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	F Statistic
X_{1L}	1	1.422 222	1580.2
X_{1Q}	1	0.197 533	219.5
X_{1C}	1	0.800 000	888.9
X_{2L}	1	0.000 000	0.0
X_{2Q}	1	3.160 573	3511.7
X_{2C}	1	0.000 000	0.0
X_1X_2	9	0.000 000	--

These results indicated that a second stage experimentation is required, and that only X_1 must be divided. In accordance with the procedure of the previous section, the range of X_1 is divided into halves. The twelve additional solution points are added to the first-stage design. These correspond to the new design points created by adding three new levels to the variable X_1. These levels are 1/6, 1/2, and 5/6. When the experimental outcomes for these new design points are added to the sixteen results of stage one, the resulting data are those necessary for a seven-level, first-stage analysis of X_1.

The single-degree-of-freedom effects for the second-stage subspace ($0.0 \leqslant X_1 \leqslant 0.5$, $0.0 \leqslant X_2 \leqslant 1.0$) are given in Table 2. These data indicate that the quadratic components of the two main effects give an adequate empirical representation of the response over this subspace, so a further division of this subspace is not required. Similar results for the remaining second-stage subspace ($0.5 \leqslant X_1 \leqslant 1.0$, $0.0 \leqslant X_2 \leqslant 1.0$), indicate that no further analysis is necessary for it.

TABLE 2

Stage Two Single-degree-of-freedom Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	F Statistic
X_{1L}	1	0.000 000	0.0
X_{1Q}	1	3.160 573	3511.7
X_{1C}	1	0.000 000	0.0
X_{2L}	1	0.000 000	0.0
X_{2Q}	1	3.160 573	3511.7
X_{2C}	1	0.000 000	0.0
X_1X_2	9	0.000 000	--

RESULTS AND CONCLUSIONS

A summary of all the examples examined is shown in Table 3. In this table, we give for each example evaluated (1) the five characteristics which describe the problem situation analyzed, (2) the type of Phase II results which would be considered ideal, and (3) an indication of the actual results obtained through the use of the Phase II procedure.

TABLE 3

SUMMARY OF PHASE II EXAMPLES

Example Number	Composition of Surface	Independent Variables	Slope of Surface	Surface Continuity	Stage One Error	No. of Sub-Spaces Ideal	No. of Sub-Spaces Actual
1	Peaks	Interacting	Very Steep	Continuous	None	2	20
2	Ridges	Non-Interacting	Gradual	Continuous	None	2	2
3	Ridges	Interacting	Gradual	Continuous	None	16	16
4	Ridges	Non-Interacting	Gradual	Continuous	None	2	2
5	Peaks	Interacting	Gradual	Continuous	None	2	2
5	Peaks	Interacting	Gradual	Continuous	Type I	4	4
6	Ridges	Non-Interacting	Gradual	Discontinuous	None	4	4
6	Ridges	Non-Interacting	Gradual	Discontinuous	Type II	4	4
6	Ridges	Non-Interacting	Gradual	Discontinuous	None	2	2
7	Saddlepoint and Peaks	Non-Interacting	Gradual	Continuous	None	2	2

As shown in the last two columns of this table, the procedure performed in an ideal manner for all the examples except Example Number One. The failure to obtain an ideal outcome for this example is the result of the extreme slopes associated with the peaks of this response surface. In essence, the Phase II procedure calls for dividing the total solution subspace into smaller and smaller subspaces until there is no need for cubic and higher order terms in an empirically derived polynomial representation of the response surface. For surfaces with extreme slopes, this condition is not satisfied until very small subspaces have been defined.

Similar results would be expected for response surfaces with very sharp discontinuities. Although Example Number Six showed that the procedure is effective for a particular discontinuous surface, this is not a result which could be generally expected. The discontinuity of this example was such that it was well suited to the Phase II procedure.

In general, the procedure was effective for response surfaces made up of peaks, ridges, or saddlepoints, and for response surfaces with either interacting or non-interacting independent variables. The insertion of a Type I error in the first-stage results for Example Number Five did affect the effectiveness of the procedure. However, a Type I error does have the adverse effect of increasing the number of resulting subspaces. The reaction of the Phase II procedure to the insertion of a Type II error in the first stage results of Example Number Six also was favorable. The fact that a Type II error was inserted was uncovered in the second stage.

In summary, it is felt that the ideas suggested in this paper fulfill a part of the need for experimental design approaches to simulation experiments. In particular, we have suggested (1) a five-phase experimental optimization procedure for computer simulations; (2) a form of experimental design suited to Phase II of the optimization procedure; and (3) an iterative Phase II procedure compatible with computer-controlled simulations. However, there is a need for additional theoretical and experimental research in this area.

BIBLIOGRAPHY

1. Box, G.E.P. and Hunter, J.S., "The 2^{K-P} Fractional Factorial Designs," Technometrics, Vol. 3, No. 3, August, 1961, and Vol. 3, No. 4, November, 1961.
2. Box, G.E.P., and Hunter, William G., "Sequential Design of Experiments for Nonlinear Models," Proceedings of the IBM Scientific Computing Symposium on Statistics. White Plains, New York: International Business Machines Corp., 1965.
3. Budne, T.A., "The Application of Random Balance Designs," Technometrics, Vol. 1, No. 2, May 1959.
4. Cochran, W.G. and Cox, G.M., Experimental Designs, 2nd ed. New York: John Wiley and Sons, Inc., 1957.
5. Davies, O.L. (ed.), The Design and Analysis of Industrial Experiments, 2nd ed., New York: Hafner Publishing Co., 1956.
6. Eldredge, David L., "Experimental Designs for the Optimization of Statistical Simulations," Unpublished Ph.D. Dissertation, The Ohio State University, 1968.
7. Hicks, C.R., Fundamental Concepts in the Design of Experiments. New York: Holt, Rinehart and Winston, 1964.
8. Myers, Raymond H., Response Surface Methodology. Boston: Allyn and Bacon, Inc., 1971.
9. Naylor, T.H., Wertz, K., and Wonnacutt, T., "Some Methods for Analyzing Data Generated by Computer Simulation Experiments," Paper presented at the National Meeting of The Institute of Management Sciences, Boston, Massachusetts, April 5-7, 1967.
10. Plackett, R.L. and Burman, J.P., "The Design of Optimum Multifactor Experiments," Biometrika, Vol. 33, Part 4, June, 1946.
11. Satterthwaite, F.E., "Random Balance Experimentation," Technometrics, Vol. 1, No. 2, May, 1959.
12. Whitwell, J.C. and Morbey, G.K., "Reduced Designs of Resolution Five," Technometrics, Vol. 3, No. 4, November, 1961.

1957 - Keiser -
(SIAM-?)
Hughes Hunter - can't do