TECHNOMETRICS, MAY 1991, VOL. 33, NO. 2

Factorial Sampling Plans for Preliminary Computational Experiments

Max D. Morris

Engineering Physics and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6367

A *computational model* is a representation of some physical or other system of interest, first expressed mathematically and then implemented in the form of a computer program; it may be viewed as a function of *inputs* that, when evaluated, produces *outputs*. Motivation for this article comes from computational models that are deterministic, complicated enough to make classical mathematical analysis impractical and that have a moderate-to-large number of inputs. The problem of designing computational experiments to determine which inputs have important effects on an output is considered. The proposed experimental plans are composed of individually randomized one-factor-at-a-time designs, and data analysis is based on the resulting random sample of observed *elementary effects*, those changes in an output due solely to changes in a particular input. Advantages of this approach include a lack of reliance on assumptions of relative sparsity of important inputs, monotonicity of outputs with respect to inputs, or adequacy of a low-order polynomial as an approximation to the computational model.

KEY WORDS: Computational model; Factor screening; Latin hypercube sampling; Sensitivity analysis.

In recent years the sciences and other disciplines have come to rely on computational models as important tools in many types of investigations. Here, the phrase *computational models* refers to representations of physical or other systems of interest that are first expressed mathematically and then implemented in the form of computer programs. Meteorological phenomena, heat transfer in engineered structures, and global economic activity, as examples, are modeled mathematically, often in the form of large or complicated systems of differential equations, and computer programs are written to evaluate the *outputs* of interest that result from specified *inputs*.

For purposes of this article, a computational model is viewed as a representation of a function that produces unique values of outputs when executed for specific values of inputs. Although the function is actually known—for example, a computer program has been written that includes a complete implicit or explicit definition of it—it is complicated enough to defy classical mathematical analysis. In practice, the behavior of the model is investigated empirically through a *computational experiment*, in which a number of evaluations—that is, runs of the program at different values of inputs—are made and some analysis is carried out on the results. The phrase "computational experiment" emphasizes parallels with the classical "physical experiment." The latter activity is an investigation of the "real world," whereas the former uses an invented substitute, but apart from this, experimental goals may be quite similar. For example, Currin, Mitchell, Morris, and Ylvisaker (in press) and Sacks, Schiller, and Welch (1989) discussed statistically motivated procedures for predicting the output of a computational model for runs not executed.

Often, computational models have as many as several hundred input variables. Additionally, for complicated models, runs of the computer program may be time-consuming and hence expensive. It follows that in the early stages of exploring a computational model a useful experimental activity is the discovery of which inputs are important—that is, which have a substantial influence on the outputs. "Factor screening" and "sensitivity analysis" are phrases commonly used to describe this kind of experiment.

The subjects of this article are the kind of information that might be useful in a preliminary computational experiment and a class of experimental plans for collecting that information. Experimental plans are proposed that are composed of individually randomized one-factor-at-a-time designs in the input variables. For these plans, data analysis can be based

on examination of changes in an output that are unambiguously attributed to changes in individual inputs. Since model outputs do not contain random error, the well-known inefficiencies of estimation based on linear-models analyses of noisy data from one-factor-at-a-time plans are not an issue here. Furthermore, this approach has the advantage of a lack of reliance on tacit or explicit assumptions of relatively few inputs having important effects, monotonicity of outputs with respect to inputs, or adequacy of a low-order polynomial model as an approximation to the computational model.

In Section 1, a brief review of some basic approaches that have been taken to designing preliminary computational experiments is presented. The basic idea on which the proposed approach is based is outlined in Section 2 and experimental plans, including two examples, are presented in Sections 3 and 5. In Section 4, comparisons between these designs and Latin hypercube designs are discussed.

1. NOTATION AND PREVIOUS APPROACHES

Consider a computational model for which an output y is a deterministic function of k inputs denoted by $x_1, x_2, x_3, \ldots, x_k$ of collectively by the k-element row vector $\mathbf{x}$. For given $\mathbf{x}$, the model can be used to evaluate y without error; that is, two evaluations of y at the same $\mathbf{x}$ are identical. In this article, y will be scalar valued, although most real computational models actually produce several output variables of interest. Alternatively, y may be thought of as a scalar-valued function of several model outputs, such as the average of some quantity that changes over simulated time. A computational experiment will consist of a collection of n *runs*, or evaluations, of the model. For a given experimental design, the ith row of the n-by-k design matrix $\mathbf{X}$ is the set of input values $\mathbf{x}$ for the ith run.

One common goal in a preliminary computational experiment is the determination of which inputs are important and so cannot reasonably be ignored in future investigations. It is often necessary that this be done using a limited number of runs of the model. In both physical and computational experiments, fractional factorial designs of resolution III or IV and Plackett and Burman (1946) designs have been used for this purpose. An approximating first-order linear model can be fit to the resulting data and the importance of each input assessed by the size of the associated model coefficient, either in absolute terms or in comparison to the size of the residuals. Since there are no physical random quantities in this setting, it is generally difficult to justify such screening rules statistically. If the approximating model is adequate, however, this approach often works well in practice anyway. Of course, nonlinear effects and individual interactions cannot be detected using these designs. The most obvious remedy to this problem is to use a more extensive design, such as a central composite design or a fractional factorial design of higher resolution, so that all coefficients that may be important in the approximating model can be individually estimated. Baker and Bargman (1985) used composite designs that allow fitting of cubic polynomials to computational models in a few inputs. When k is large, however, even a two-level resolution V fractional factorial experiment may be too expensive to perform in a preliminary study.

The need to estimate a potentially large number of linear-model coefficients from a small physical experiment led Satterthwaite (1959) to introduce random balance designs. The basic strategy of these designs is to specify the values of each x_i randomly and then to investigate the marginal effect of each input individually, ignoring the values actually selected for the other inputs. He pointed out that under a first-order model the least squares estimator of each coefficient, ignoring the rest, is unbiased; the bias that would be present were the entire design considered to be fixed is absorbed into the variance of the estimator. Satterthwaite's arguments were not fully accepted for several reasons, however; see the discussion of Youden, Kempthorne, Tukey, Box, and Hunter (1959) following Satterthwaite's article. A major argument against their use was their lack of efficiency in estimating individual coefficients.

A randomzied experimental plan often used for input screening with deterministic computational models is the Latin hypercube design (Iman and Conover 1980). Latin hypercube designs were originally introduced in the context of a different problem, the estimation of the distribution of an output that propagates as the result of a known joint distribution of inputs (McKay, Beckman, and Conover 1979). As in the case of random balance designs, the values of each x_i are independently generated using a restricted random procedure. Each x_i takes on n unique values selected from n nonoverlapping intervals in its domain of interest.

A primary appeal of these randomized plans is that, unlike central composite designs and fractional factorials of given resolution, k does not necessarily play a role in determining n. Satterthwaite apparently felt that random balance designs were not an attractive option for cases in which $n > k$; see Satterthwaite and Budne (1959). Latin hypercube designs have been applied, however, in situations in which n is much greater than k; Iman and Conover (1980) presented an example when $k = 8$ and $n = 200$ in which the identification of influential inputs is based on partial correlation coefficients computed on the ranks of the data.

In Section 2, a different approach to identifying the important inputs in a computational model is proposed, intended specifically for those cases in which n is to be of order k and no simplifying assumptions about the form of the model are to be used.

2. ELEMENTARY EFFECTS

For purposes of this article, each x_i will be assumed to be scaled to take on values in the interval [0, 1], and the *region of interest*, Ω, will be assumed to be the k-dimensional unit hypercube. If y is at least once differentiable with respect to each input, $\partial_i(\mathbf{x}) = \partial y/\partial x_i|_{\mathbf{x}}$ is a functional index of the influence of x_i on y. For example, $\partial_i(\mathbf{x})$ may be exactly or approximately (a) zero over all values of $\mathbf{x}$, (b) a nonzero constant over all values of $\mathbf{x}$, (c) a nonconstant function of x_i, or (d) a nonconstant function of one or more $x_j (j \neq i)$. These correspond, respectively, to situations in which the effect of x_i on y is said to be (a) negligible, (b) linear and additive, (c) nonlinear, or (d) involved in interactions with other inputs. The guiding philosophy in this work is that a major role of a preliminary experiment is to determine, within reasonable uncertainty, which inputs may be considered to have effects which are (a) negligible, (b) linear and additive, and (c or d) other. The reason for combining (c) and (d) here is that, where important nonlinearity or interaction exists, an experiment of any design that is small relative to the number of inputs will generally not produce enough information to resolve the nature of these effects; a subsequent detailed experiment in only those inputs whose effects are complicated is the appropriate stage for this study.

Here, a "discretized" approach following the preceding idea is proposed. Attention will be restricted to a *region of experimentation*, ω, which is a regular k-dimensional p-level grid, where each x_i may take on values from $\{0, 1/(p-1), 2/(p-1), \ldots, 1\}$. The following discussion will be based on what will be called the *elementary effects* attributable to each input. For a given value of $\mathbf{x}$, define the elementary effect of the ith input as

$$d_i(\mathbf{x}) = [y(x_1, x_2, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_k) - y(\mathbf{x})]/\Delta, \quad (1)$$

where $\mathbf{x} \in \omega$, except that $x_i \leq 1 - \Delta$ and Δ is a predetermined multiple of $1/(p-1)$. (The development to follow could easily be generalized to allow assignment of potentially different values of p and Δ to each input. Although some prior knowledge about the function y might make this desirable, single values of p and Δ have been used here to simplify the presentation.)

The intent here is to discover useful information about which inputs are important using a design in which the number of runs is proportional to k (rather than k^2, etc.). Therefore, it seems reasonable that the approach should be based on trying to discover a fixed number of properties about the influence of each input on y (rather than individual interactions in an approximating polynomial model whose number grows with k). Here it is proposed that the finite distribution of $p^{k-1}[p - \Delta(p-1)]$ elementary effects associated with each input be estimated; for input i, this distribution will be denoted by F_i. A large (absolute) measure of central tendency for F_i indicates an input with an important "overall" influence on the output. A large measure of spread indicates an input whose influence is highly dependent on the values of the inputs—that is, one involved in interactions or whose effect is nonlinear. In particular, estimates of the means and standard deviations of these distributions will be used as indicators of which inputs should be considered important. The sampling plans to be discussed provide random samples from each of these distributions on which such estimates may be based.

3. PLANS FOR INDEPENDENT RANDOM SAMPLING

In simplest form, randomly selecting a value from F_i requires random selection of a value of each $x_j (j = 1, 2, 3, \ldots, k)$ and evaluation of y twice, once at the selected values and again after increasing x_i by the quantity Δ; these two runs would then yield one elementary effect. This could be repeated r times to produce a random sample of r elementary effects from F_i. If the procedure were performed for each input, the result would be a random sample of r values from each F_i at a total cost of $n = 2rk$ runs. In this article, the *economy* of a design will be defined to be the number of elementary effects it produces divided by the number of experimental runs, so this sampling scheme has an economy of $\frac{1}{2}$. (I have intentionally avoided use of the word "efficiency" here. Although economy is closely related to statistical efficiency in the discussion of this section, the relationship is less direct for the plans of Sec. 5.) Note that under this plan all rk observed elementary effects are independently drawn.

More economical designs can be constructed if some runs are used in computing more than one elementary effect. In this development, it will be convenient to restrict attention to the case in which p is even and $\Delta = p/[2(p-1)]$; the reason for this will be clear shortly. To generate such designs, the first step will be the selection of an m-by-k *sampling matrix*, $\mathbf{B}$, with elements that are 0's and 1's, which has the key property that for every column $i = 1, 2, 3,$

. . . , k there are two rows of $\mathbf{B}$ that differ only in their ith entries; for example,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}. \quad (2)$$

(In this section, m will always equal $k + 1$; this is generalized in Sec. 5.) $\Delta\mathbf{B}$ could be used as a design matrix (i.e., each row a value for $\mathbf{x}$) for which the corresponding experiment would provide k elementary effects, one for each input, based on only $k + 1$ runs. These would not be random selections from the distributions $F_1, F_2, \ldots, F_k$, however. To obtain a random selection from each distribution, a randomized version of the sampling matrix is employed as follows:

1. Let $\mathbf{D}^*$ be a k-dimensional diagonal matrix in which each diagonal element is either $+1$ or -1 with equal probability. Letting $\mathbf{J}_{m,k}$ be the m-by-k matrix of 1's, note that $(1/2)[(2\mathbf{B} - \mathbf{J}_{m,k})\mathbf{D}^* + \mathbf{J}_{m,k}]$ is an m-by-k matrix in which each column is either set equal to its corresponding column in $\mathbf{B}$ or is determined by replacing 1's for 0's and 0's for 1's in the corresponding column of $\mathbf{B}$.
2. Let $\mathbf{x}^*$ be a randomly chosen "base value" of $\mathbf{x}$ for which each element is randomly assigned a value from $\{0, 1/(p-1), 2/(p-1), \ldots, 1 - \Delta\}$, each with equal probability.
3. Let $\mathbf{P}^*$ be a k-by-k random permutation matrix in which each column contains one element equal to 1 and all others equal to 0 and no two columns have 1's in the same position, where each such matrix has an equal probability of selection.

Attention is restricted to the case in which each decision in the randomization process—that is, selection of values for each element of $\mathbf{D}^*$ and $\mathbf{x}^*$ and selection of $\mathbf{P}^*$—is made independently of all others. Then $\mathbf{B}^* = (\mathbf{J}_{m,1}\mathbf{x}^* + (\Delta/2)[(2\mathbf{B} - \mathbf{J}_{m,k})\mathbf{D}^* + \mathbf{J}_{m,k}])\mathbf{P}^*$ is called a random *orientation* of $\mathbf{B}$. Like $\Delta\mathbf{B}$, $\mathbf{B}^*$ also provides one elementary effect per input, but one which is randomly selected. [If it is desired to assign values of p_i and $\Delta_i = p_i/[2(p_i - 1)]$ to each input, step 2 of the randomization procedure would be modified so that each element of $\mathbf{x}^*$ takes a randomly chosen value from the appropriate set and step 3 would be modified to permute columns only within sets corresponding to inputs with equal values of p_i and Δ_i.] Finally, if a sample of r effects is required from each F_i, r independent random orientations of $\mathbf{B}$ can be concatenated to form the design matrix for the entire experiment:

$$\mathbf{X} = \begin{bmatrix} \mathbf{B}_1^* \\ \mathbf{B}_2^* \\ \cdots \\ \mathbf{B}_r^* \end{bmatrix}. \quad (3)$$

The preceding randomization scheme defines the probabilities with which elementary effects are selected with each orientation of $\mathbf{B}$. To characterize these probabilities, consider the two rows of $\mathbf{B}$ which differ only in their ith elements,

$$\mathbf{B}(i) = \begin{bmatrix} x_1\, x_2 \cdots x_{i-1}\, x_{i,1}\, x_{i+1} \cdots x_k \\ x_1\, x_2 \cdots x_{i-1}\, x_{i,2}\, x_{i+1} \cdots x_k \end{bmatrix},$$

and the result of only the first two stages of the randomization process on these rows; that is, $\mathbf{J}_{2,1}\mathbf{x}^* + (\Delta/2)[(2\mathbf{B}(i) - \mathbf{J}_{2,k})\mathbf{D}^* + \mathbf{J}_{2,k}]$. It is clear that in any column except the ith, the two elements will be equal and have one of the values

$$0 + 0,\quad 1/(p-1) + 0,\quad 2/(p-1) + 0, \ldots, (1-\Delta) + 0,$$
$$0 + \Delta,\quad 1/(p-1) + \Delta,\quad 2/(p-1) + \Delta, \ldots, (1-\Delta) + \Delta,$$

each with equal probability. If p is even and Δ has been chosen to be $p/[2(p-1)]$, this list of values reduces to $0, 1/(p-1), 2/(p-1), \ldots, (p-2)/(p-1)$, and 1. Similarly, in column i the two unequal values will be 0 and Δ, $1/(p-1)$ and $1/(p-1) + \Delta$, $2/(p-1)$ and $2/(p-2) + \Delta, \ldots$, or $1 - \Delta$ and 1, each with equal probability. Since randomization is applied independently in each column, each of the $p^{k-1}[p - \Delta(p-1)] = p^k/2$ elementary effects for input i has an equal probability of selection following the first two randomization steps. It is easy to see that the third step in the process does not change this, since it is simply an independent shuffling of input labels.

Although the third step of this randomization procedure is not necessary to ensure equal-probability sampling from each F_i, it results in a certain symmetric treatment of inputs which may be desirable. For example, using only the first two steps in the randomization process, inputs represented by the central columns in $\mathbf{B}$ from Equation (2) are likely to have a more even distribution of values in $\mathbf{X}$ than those represented by the first and last columns. Adding column permutation to the randomization procedure averages this potential "lumping" of values across all columns and decreases the probability of a severely imbalanced sample in any column.

Over the entire design $\mathbf{X}$, r elementary effects are produced for each input at a total cost of $n = rm$

runs, so the economy of the sampling plan is $rk/rm = k/m$, or $k/(k + 1)$ for the sampling matrix of Equation (2). The price of the increased economy is nonindependence *among* samples from different F_i; within a single orientation of **B**, the elementary effects observed from the different F_i have a restricted random relationship. Since only one elementary effect is observed from each F_i with each randomization of **B**, however, the r observations within the sample from each F_i remain independent.

Since **X** produces a simple random sample, with replacement, from each F_i, the sample mean ($\bar{d}_i$) and variance (S_i^2) of the observed elementary effects for input i are unbiased estimators of the mean and variance of F_i, and the standard error of the mean can be estimated as $\text{SEM}_i = S_i/\sqrt{r}$. Since the sample is taken with replacement, two orientations of **B** can produce the same elementary effect for a given input, but the probability of this happening is reduced rapidly as k increases relative to r; in the range of most practical applications, it may occur very infrequently. When it does occur, it can be treated as in any other sampling-with-replacement setting; for example, both occurrences of the (same) elementary effect are included in summary statistics. As an aside, duplication of an elementary effect in a sample implies that at least two pairs of rows in **X** are duplicates; in practice, some computational expense would be spared because one run of the deterministic model will be required for each *distinct* row in **X**.

3.1 Example

To demonstrate the basic idea of this article, an artificial computational model with 20 inputs was constructed as follows:

$$y = \beta_0 + \sum_{i}^{20} \beta_i w_i + \sum_{i<j}^{20} \beta_{i,j} w_i w_j + \sum_{i<j<l}^{20} \beta_{i,j,l} w_i w_j w_l + \sum_{i<j<l<s}^{20} \beta_{i,j,l,s} w_i w_j w_l w_s,$$

where $w_i = 2(x_i - \frac{1}{2})$ except for $i = 3$, 5, and 7, where $w_i = 2(1.1x_i/(x_i + .1) - \frac{1}{2})$. Coefficients of relatively large value were assigned as

$$\beta_i = +20, \qquad i = 1, \ldots, 10,$$

$$\beta_{i,j} = -15, \qquad i, j = 1, \ldots, 6,$$

$$\beta_{i,j,l} = -10, \qquad i, j, l = 1, \ldots, 5,$$

and

$$\beta_{i,j,l,s} = +5, \qquad i, j, l, s = 1, \ldots, 4.$$

The remainder of first- and second-order coefficients were generated independently from a normal distribution with zero mean and unit standard deviation, and the remainder of third- and fourth-order coefficients were set to 0.

The 21-run, 20-factor sampling matrix of Equation (2) was used, and $r = 4$ random orientations were generated using $p = 4$ ($\Delta = \frac{2}{3}$) to produce an 84-run design. Based on the 84 computed values of y, a random sample of four elementary effects was observed for each of the 20 inputs. In Figure 1, the mean and standard deviation of each sample of elementary effects are displayed; numbers on the graph identify the inputs. Two lines are also graphed, corresponding to $\bar{d}_i = \pm 2\,\text{SEM}_i$; if the coordinates for input i lie outside of the wedge formed by these two lines, one might interpret this, approximately, as significant evidence that the expectation of F_i is nonzero.

From a practical standpoint, judgment about the importance of particular values of $\bar{d}_i$ and S_i will usually be context dependent. For example, $\bar{d}_1 = 76$ would be declared significantly different from 0 by comparison to SEM_1. But in the context of a real problem, it is important to decide whether an average change estimated to be 76 units in y due to a change of $\frac{2}{3}$ is scaled x_1 is physically meaningful and whether a standard deviation estimated to be $S_1 = 24$ about this value, as one chooses d_1 values randomly from over ω, constitutes physically meaningful variation. In the absence of such knowledge and for purposes of this example, we may examine the plotted values of Figure 1 relative to each other to see which appear to be most important.

Inputs 1–10 are clearly separated from the cluster of remaining outputs, which have means and standard deviations close to 0. In particular, inputs 8, 9, and 10 have mean elementary effects that are substantially different from 0 while having small standard deviations. Considering both means and standard deviations together, one might (correctly in this case) conclude that the first 10 inputs are important, and that of these the first seven appear to have effects that involve either curvature or interactions.

For comparison with what might be called a "standard method," an 84-run Latin hypercube design was also constructed and the corresponding experiment performed using this model. (Actually, a popular variant of the original Latin hypercube design was used; the randomized values included in each column of the design matrix were .5/84, 1.5/84, 2.5/84, . . . , 83.5/84, rather than being random quantities taken from 84 subintervals.) The ranks of the observed values of y were then fit to a first-order regression model in the inputs; the significance levels for the model coefficients are listed in Table 1. Six inputs had effects significant at the .05 level. Not

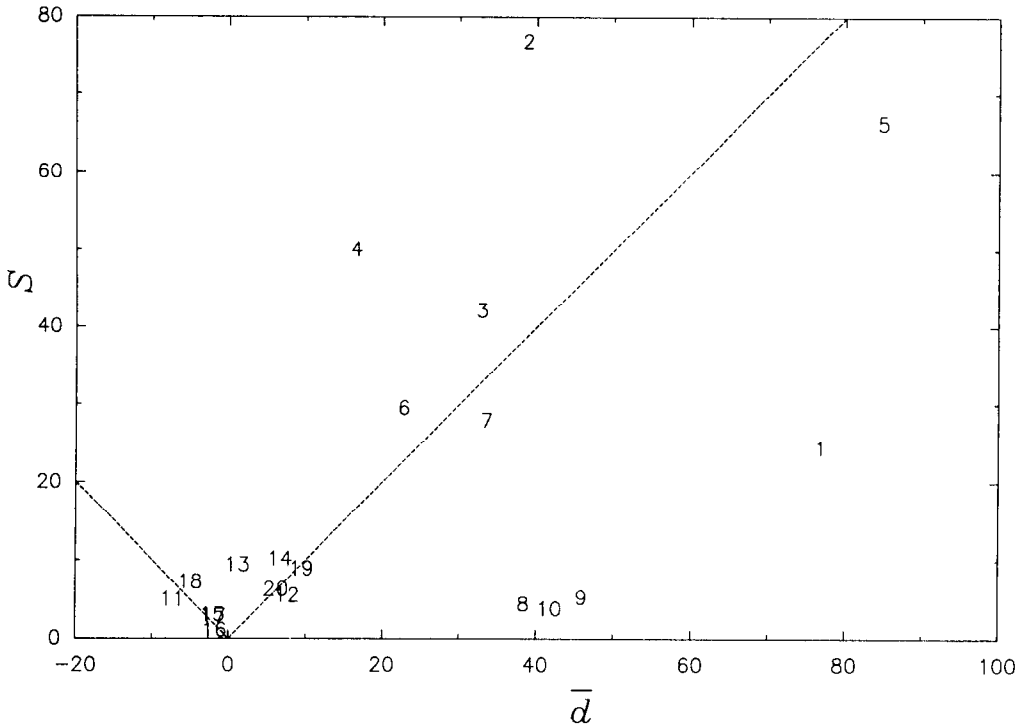

Figure 1. Estimated Means ($\bar{d}_i$) and Standard Deviations (S_i) of the Distributions of Elementary Effects in the Example of Section 3. Lines correspond to $\bar{d}_i = \pm 2\ SEM_i$.

surprisingly, these were all among the truly active factors and in four cases, were inputs not involved in interactions. No coefficients had significance levels between .05 and .10. Input screening based on these values alone would result in missing some active inputs (4, 2, and perhaps 5 and 1), or including some that are inactive (perhaps 20, 13, and 15), depending on the significance value selected for screening. In Section 4, some of the differences between Latin hypercube designs and those described here are discussed in more detail.

3.2 Selection of a Sampling Matrix

The sampling matrix **B** of Equation (2) when viewed as a design matrix is an example of what statisticians often call one-factor-at-a-time (o.a.t.) designs; Daniel (1973) offered an interesting discussion and system of classification of o.a.t. designs. They have historically received less attention than other main-effects plans in the statistical literature because they lead to least squares estimators of model coefficients that are relatively imprecise in settings in which the response variable contains random error.

Table 1. Significance Levels for Latin Hypercube Design of Section 3

Input	p value	Input	p value
1	.12	11	.69
2	.31	12	.67
3	.01	13	.23
4	.41	14	.75
5	.20	15	.23
6	.01	16	.32
7	.0002	17	.96
8	.0002	18	.87
9	.0001	19	.81
10	.0001	20	.15

Other $(k + 1)$-row sampling matrices exist that allow calculation of one elementary effect per input. Any such sampling matrix can, by reordering columns and exchanging 0's and 1's within some columns, be put in a standard form:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ * & 1 & 0 & \cdots & 0 \\ * & * & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ * & * & * & \cdots & 1 \end{bmatrix}, \quad (4)$$

where each asterisk can be either 0 or 1, but such that

the ith row of **B** differs from some row above it in only the $(i - 1)$st element. (5)

For a proof of this, see the Appendix. As an aside, Webb (1968) showed that under the usual first-order linear model assumptions including random noise, in the class of designs with design matrices of form (4),

a necessary and sufficient condition for minimum variance estimation of the model coefficients is (5).

All experimental plans generated by sampling matrices satisfying (4) and (5) have the same economy, $k/(k + 1)$. The following criterion is suggested as a possible aid for selecting a sampling matrix from this class. Let the random variable Z denote the number of runs in $\mathbf{X}$ for which a selected input appears at a selected value. It is easy to show that, with respect to the three-stage randomization described previously,

$$E(Z) = n/p \tag{6}$$

and

$$V(Z) = (2r/pk) \sum_{j=1}^{m-1} [j - (m/2)]^2 C(j) + (m^2 r/p)[(1/2) - (1/p)], \tag{7}$$

where $C(j)$ is the number of columns of $\mathbf{B}$ containing j 1's. Since Equation (6) is independent of the sampling matrix used, it describes a kind of expected balance, which is a property of each design generated by the sampling matrices of form (4) and (5). On the other hand, $V(Z)$ depends on $\mathbf{B}$ through the function C, and it seems desirable to minimize this quantity with respect to selection of $\mathbf{B}$. For each value of k up to 10, $V(Z)$ was computed for each sampling matrix in this class, and in each case the matrix which minimizes $V(Z)$ is the one described in Equation (2). [For a given value of k, $(k - 1)!$ matrices can be constructed that satisfy Equations (4) and (5), but this includes large groups of matrices that are equivalent in that one may be transformed to another by reversing symbols in some columns and reordering rows. In the calculation just described, *all* sampling matrices that minimize $V(Z)$ are either of the form displayed in Equation (2) or are equivalent to it in this sense.] I conjecture that this is also true for larger k, but I have not proven it.

4. CONTRASTS WITH LATIN HYPERCUBE SAMPLING

As noted previously, Latin hypercube sampling (LHS) has become a popular technique for generating designs for computational experiments. These designs have several appealing characteristics, including the following:

LHS1. Latin hypercube designs tend to be spread fairly uniformly throughout Ω.

LHS2. The projection of a Latin hypercube design into any subspace of Ω formed by ignoring some inputs is also a Latin hypercube design with the same number of unique runs.

Property LHS1 is certainly appealing in the sense that an exploratory experiment should intuitively cover the region of interest as completely as possible in the event that some subregion turns out to be particularly interesting. Moreover, if y is a relatively simple function of the inputs, property LHS1 may imply that a useful approximation to y may be constructed from Latin hypercube data. The primary attraction of property LHS2 is that if it turns out that only a few inputs have important effects, the design will likely be well distributed in the reduced input space.

The designs described in Section 3 do not have the properties just listed, but they do have the following, which are not properties of Latin hypercubes:

OAT1. Randomized sampling matrices allow direct observation of elementary effects.

OAT2. The projection of each randomized sampling matrix into any subspace of Ω formed by ignoring some inputs is also a randomized sampling matrix with one unique run lost for each dimension ignored.

Property OAT2 is much weaker than its counterpart for Latin hypercubes; reduction to a lower dimension is costly in terms of the loss of unique design runs. If it turns out that several of the inputs have important effects, however, property OAT1 guarantees that information can be extracted for each input, and that this information (a) is meaningful regardless of the complexity of the input's effect and (b) cannot possibly be mistakenly attributed to that input through partial aliasing or confounding with effects actually associated with other inputs.

The following artificial example illustrates a situation in which the designs introduced here may have an advantage over Latin hypercube designs in the input screening context. Define y as follows to have 100 inputs, where only the first 30 actually influence the output:

$$y = \sum_{i=1}^{30} \exp(5.5x_i - 1.5\bar{x}),$$

$$\text{where } \bar{x} = (1/30) \sum_{i=1}^{30} x_i.$$

The heavy curve in Figure 2 shows the cumulative distribution of $\partial_1(\mathbf{x})$, uniformly sampled over Ω. (Since the first 30 inputs play symmetric roles in y, derivatives taken with respect to inputs 2–30 would be distributed identically; of course, derivatives with respect to inputs 31–100 are 0 throughout Ω.) Although y is not monotonic in x_1, $\partial_1(\mathbf{x})$ is positive whenever its absolute value is relatively large. Since y is a function of only 30% of the inputs, some degree of "effect sparsity" (Box and Meyer 1986) is clearly present,

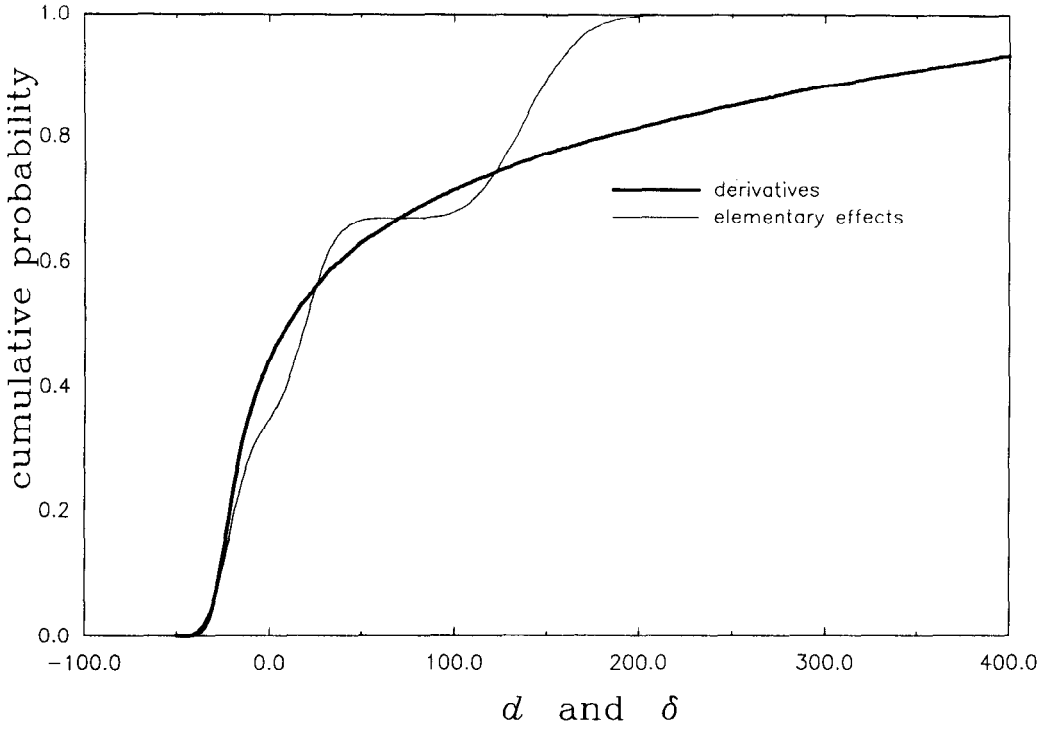

Figure 2. Cumulative Distribution Functions of Elementary Effects (d_1) and Derivatives (∂_1) for Input 1 in the Example of Section 4.

although perhaps not so much as may often be implied by that phrase. The term involving $\bar{x}$ introduces some interaction among the inputs, and the exponential form results in strong curvature in the effects.

One hundred Latin hypercube designs in $n = 303$ runs were generated, and for each of these the rank-transformed y values were fit to all inputs by multiple linear least squares. (As in the example of Sec. 3, each x_i took on 303 equally spaced values, randomly assigned to runs; regression on these input values is therefore equivalent to regression on their ranks.) The importance of each input was assessed by the usual t statistic associated with the corresponding regression coefficient; those with values of 2 or greater were declared to be important. Using this rule, the overall proportion of tests that indicated that one of the first 30 inputs was important was .634, and the corresponding proportion for inputs 31–100 was .048. Hence on average one might expect to correctly identify 19 or 20 of the 30 important inputs and incorrectly identify 3 or 4 of the unimportant inputs with this procedure.

For comparison, consider any design of the form discussed in Section 3 in $p = 6$ levels ($\Delta = 3/5$) and $n = 303$ runs. For each of inputs 1–30, a simple random sample of three elementary effects would be drawn from a population with cumulative distribution graphed as the lighter curve in Figure 2. In Figure 3, contours of the joint cdf of $|\bar{d}_i|$ and S_i are graphed. Note that in less than 10% of such samples would both statistics have values less than 20; 75% of the time either the absolute value of the mean would be greater than 25 or the standard deviation would be greater than 35. Of course, all samples of elementary effects associated with inputs 31–100 would contain only zeros.

Admittedly, this example function is much simpler than computational codes encountered in practice; for example, all important inputs have the same effect, and all others have absolutely no effect at all. Yet it presents one situation in which an o.a.t. approach seems much more likely to succeed in screening inputs than the use of rank analysis of data from an LHS. It should be stressed, however, that the designs described here are not suggested as competitors for Latin hypercube designs. Indeed, they are unattractive alternatives in the cases for which Latin hypercube designs may be best suited—for example, relatively few important inputs with effects that are not highly nonlinear or interactive. An important distinction seems to be that, using a Latin hypercube design, the effect of any one input can only be assessed either by ignoring the other inputs or by adjusting for them *using some assumed approximating model*, at least in the rank data. The designs discussed here are intended for situations in which several inputs may have complex effects, and the type

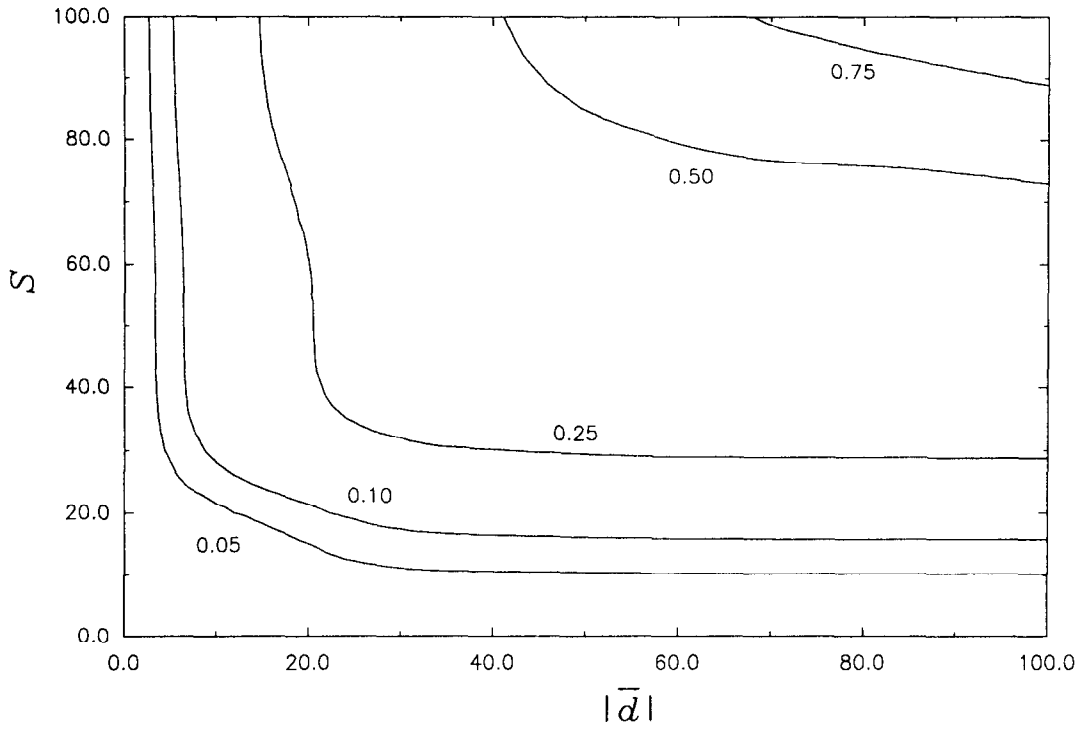

Figure 3. Contours of the Joint Cumulative Distribution Function of the Absolute Value of the Mean ($|\bar{d}_1|$) and Standard Deviation (S_1) of a Sample of Three Elementary Effects for Input 1 in the Example of Section 4.

of analysis required by the Latin hypercube design seems risky.

5. PLANS FOR CLUSTER SAMPLING

In the designs discussed in Section 3, the sampling matrix **B** was constructed to yield one elementary effect for each input. In this section, sampling matrices with $m > k + 1$ rows, which yield more than one elementary effect for each input, are considered. These can be constructed for two different reasons, to increase the economy of the resulting design or to ensure some degree of separation (in Ω) of the sampled elementary effects for each input.

5.1 Cluster Sampling for Increased Economy

Consider first the sampling matrix

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{8}$$

One random orientation of **B** yields two elementary effects for each input at a cost of less than two runs per input, so a design constructed from this sampling matrix has higher economy then the designs discussed in Section 3. The result is not a simple random sample of size 2 from each F_i, however, because pairs of elementary effects for the same input are selected together. As will be discussed, this may be thought of as a *cluster sample* (e.g., Cochran 1977) of cluster size 2. So a sampling matrix that provides more than one elementary effect per input per orientation will be referred to as a *cluster sampling matrix* and the resulting design as a cluster design.

As usually defined, cluster sampling actually refers to situations in which the sampling unit is a subset of the population of interest and the subsets available for sampling constitute a partition of that population. The collection of elementary effects for a given input, sampled as a result of a randomized orientation of **B**, does not constitute a cluster from the population of interest in this sense. Valid inferences about F_i can, however, be obtained by treating this collection as if it were a cluster sample. To see this, let the number of elementary effects observed for each input from one orientation of **B** be c; in the preceding example $c = 2$. Moreover, suppose that these sampled elementary effects are arbitrarily labeled by the order in which they can be calculated by examining $\mathbf{B}^*$ row by row from top to bottom as $d_i^1, d_i^2, \ldots, d_i^c$ for input i. Following the arguments of Section 3, each of the d_i^j can be any one of the $p^k/2$ possible elementary effects attributable to input i with equal probability. This means that, of the $p^k k!$ equally

probable random orientations of **B**, $2k!$ of them lead to any particular realization for each d_i^j. But no two of $d_i^1, d_i^2, \ldots, d_i^c$ can yield the same realization on a given orientation, unless **B** contains duplicate rows. Hence $2ck!$ of the possible orientations of **B** lead to selection of any given elementary effect. Therefore, our sampling plan is equivalent to cluster sampling from an artificial population containing $2ck!$ complete copies of the actual population of interest. But since the *distribution* of this population is the same as that of the population of interest—that is, F_i—inferences about that distribution can be drawn using the standard tools of cluster sampling.

One series of designs that yields more than one elementary effect per run can be generated by sampling matrices of form

$$\mathbf{B} = \begin{bmatrix} \mathbf{o} & \mathbf{o} & \mathbf{o} & \ldots & \mathbf{o} \\ \mathbf{C} & \mathbf{O} & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{J} & \mathbf{C} & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{J} & \mathbf{J} & \mathbf{C} & \ldots & \mathbf{O} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbf{J} & \mathbf{J} & \mathbf{J} & \ldots & \mathbf{C} \end{bmatrix}, \quad (9)$$

where **o** is a q-element row vector of 0's, **O** and **J** are $p \times q$ matrices of 0's and 1's, respectively, and **C** is a $p \times q$ matrix of both 0's and 1's. In this investigation, selection of **C** has been limited as follows. Once a value of q has been determined, allowable rows for **C** are grouped by the number of 1's they contain; group 1 contains all q-element vectors with a single 1 and the remainder 0's, group 2 contains all vectors with two 1's and the rest 0's, and so forth through group q. (Group 0, the single vector of q 0's is not included here because it is redundant.) A matrix **C** is then constructed of entire groups of rows; for example, either all of group 1 appears or no row with a single 1 appears. **C** is thus a special case of a balanced array of strength q. Selection of these groups determines the value of p and hence the row dimension of **B**.

The number of elementary effects provided (per input, per orientation) by the sampling matrix of equation (9) is

$$I(1)I(q) + \sum_{i=2}^{q} I(i-1)I(i)\binom{q-1}{i-1},$$

where $I(i)$ is 1 if group i is included and 0 otherwise. In Table 2, sampling matrices from this series are listed that yield the indicated number of elementary effects per orientation with a minimum number of runs, where the number of rows in **B** is less than $4k$.

The sampling matrix of Equation (9) has not been shown to be optimal with respect to economy. This form is tried here primarily because of its similarity to that of Equation (2). More economical plans may well exist.

Table 2. Approximate (large k) Economies for Some Designs Generated by Equation (9)

q	Groups	$(m-1)/k$	Elementary effects per input	Approximate economy
2	1, 2	1.50	2	1.33
3	1, 2, 3	2.33	4	1.71
5	1, 2, 5	3.20	5	1.56
5	3, 4, 5	3.20	5	1.56
4	1, 2, 3	3.50	6	1.71
4	1, 2, 3, 4	3.75	8	2.13

5.2 Cluster Sampling for Separation of Elementary Effects

A second reason for using cluster sampling is to ensure that the elementary effects sampled for each input will be separated, to some degree, in Ω. For example, each orientation of

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & \cdots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \quad (10)$$

yields $c = 2$ elementary effects from each F_i at a cost of $m = 2k$ runs. The economy of designs constructed with this sampling matrix is roughly that of those discussed in Section 3, about one elementary effect per run. Within each sampled pair, however, the two observed elementary effects represent response differences along opposite edges of a hypercube of edge length Δ [rather than, for example, adjacent edges as in the sampling matrix of Eq. (8)]. In some cases, this may increase our ability to detect spread in the distribution of each F_i. [An alternative justification might be based on the fact that since **B** is a foldover (Box and Wilson 1951) of a resolution III o.a.t. design matrix, each random orientation is a complete resolution IV design.]

5.3 Parameter Estimation

The mean of F_i may be estimated from a cluster sample using the same estimator as would be used with an independent random sample—for example, the sample average. As discussed by Cochran (1977), an unbiased estimator of the variance of F_i is

$$S_i^2 = [(p^k k! - 1)S_{i,a}^2 + p^k k!(c-1)S_{i,w}^2] \div (cp^k k! - 1),$$

where $S^2_{i,a}$ and $S^2_{i,w}$ are the among- and within-cluster mean squares, respectively, from an analysis of variance of the observed elementary effects for input i. This is nearly identical to the approximation $S_i^2 \doteq [S^2_{i,a} + (c - 1)S^2_{i,w}]/c$. The standard error of the mean of elementary effects for input i can be estimated as

$$\text{SEM}_i = S_{i,a}/\sqrt{rc}. \tag{11}$$

5.4 Example

Next, an experiment is described that was based on a cluster design, performed to investigate properties of the computational model TWOLAYER, which was developed by Alan Solomon and his colleagues at the Oak Ridge National Laboratory. TWOLAYER models heat transfer into, out of, and through a wall containing two layers of possibly different phase-change materials. Heat is applied to the wall during a 10-hour charge cycle, during which time some of the phase-change material melts. During the following 14 hours (the discharge cycle), heat is released from the wall as the phase-change material solidifies.

In this experiment, effects of $k = 20$ inputs were investigated; they are material properties and physical dimensions of the phase-change materials and are listed in Table 3. Each input was scaled so that 0 and 1 represent 90% and 110%, respectively, of a nominal value. The output investigated was a "utility index," which takes on values in the interval [0, 1] and is a measure of the effectiveness of the wall as a heat-storage device. The experiment was performed on a Cray X/MP computer, and each run of the model required approximately 20 seconds of machine time.

The sampling matrix used to generate the design is

$$\mathbf{B} = \begin{bmatrix} \mathbf{o} & \mathbf{o} & \mathbf{o} & \mathbf{o} & \mathbf{o} \\ \mathbf{B}_1 & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{J} & \mathbf{B}_1 & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{J} & \mathbf{J} & \mathbf{B}_1 & \mathbf{O} & \mathbf{O} \\ \mathbf{J} & \mathbf{J} & \mathbf{J} & \mathbf{B}_1 & \mathbf{O} \\ \mathbf{J} & \mathbf{J} & \mathbf{J} & \mathbf{J} & \mathbf{B}_1 \end{bmatrix},$$

where

$$\mathbf{B}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

O and **J** are 7×4 matrices of 0's and 1's respectively, and $\mathbf{o} = (0, 0, 0, 0)$. This sampling matrix generates a design with economy 1.11 and yields two elementary effects for each input, $d_i(\mathbf{x}^1)$ and $d_i(\mathbf{x}^2)$, for which $\mathbf{x}^1$ and $\mathbf{x}^2$ differ by an amount Δ in three elements. Note that this sampling matrix is not in the class discussed in Section 5.1; even though **B** is of the form described in Equation (9), $\mathbf{B}_1$ is not a balanced array. This particular sampling matrix was chosen as a kind of compromise between the two types of cluster plans described in Sections 5.1 and 5.2 in that it generates a design with slightly better economy than the plans of Section 3 while ensuring some separation between the elementary effects within each cluster for each input. $r = 3$ orientations of **B** were used, each specifying $m = 36$ experimental runs, for a total of 108 runs in the experiment.

Table 3. Inputs Used in the TWOLAYER Experiment of Section 5

Input number	Input description
1	Liquid conductivity of left-layer material
2	Solid conductivity of left-layer material
3	Liquid conductivity of right-layer material
4	Solid conductivity of right-layer material
5	Liquid specific heat of left-layer material
6	Solid specific heat of left-layer material
7	Liquid specific heat of right-layer material
8	Solid specific heat of right-layer material
9	Liquid density of left-layer material
10	Solid density of left-layer material
11	Liquid density of right-layer material
12	Solid density of right-layer material
13	Latent heat of left-layer material
14	Melt temperature of left-layer material
15	Latent heat of right-layer material
16	Melt temperature of right-layer material
17	Width of left layer
18	Width of right layer
19	Charge film coefficient
20	Discharge film coefficient

Results of the experiment are displayed in Figure 4, where $(\bar{d}_i, S_i)$ is plotted for each input. Points for which $|\bar{d}_i| > 2\ \text{SEM}_i$ as computed using Equation (11) are marked in boxes. For 11 of the 20 inputs, estimated means and standard deviations were both exactly 0. This is probably due, at least in part, to the fact that TWOLAYER is a finite element code in which the wall thickness is represented by (in this case) 10 discrete nodes; at a particular point in time, if three nodes represent the liquid state and seven represent the solid state, the modeled wall is said to be 30% liquid. As a result, there were only 23 distinct values of y produced in the 108 runs of the model. Although accuracy of the simulation would undoubtedly be improved by increasing the number of nodes and consequently the amount of computer time needed for each run, numerical curiosities of this type are not an uncommon result of computational experiments.

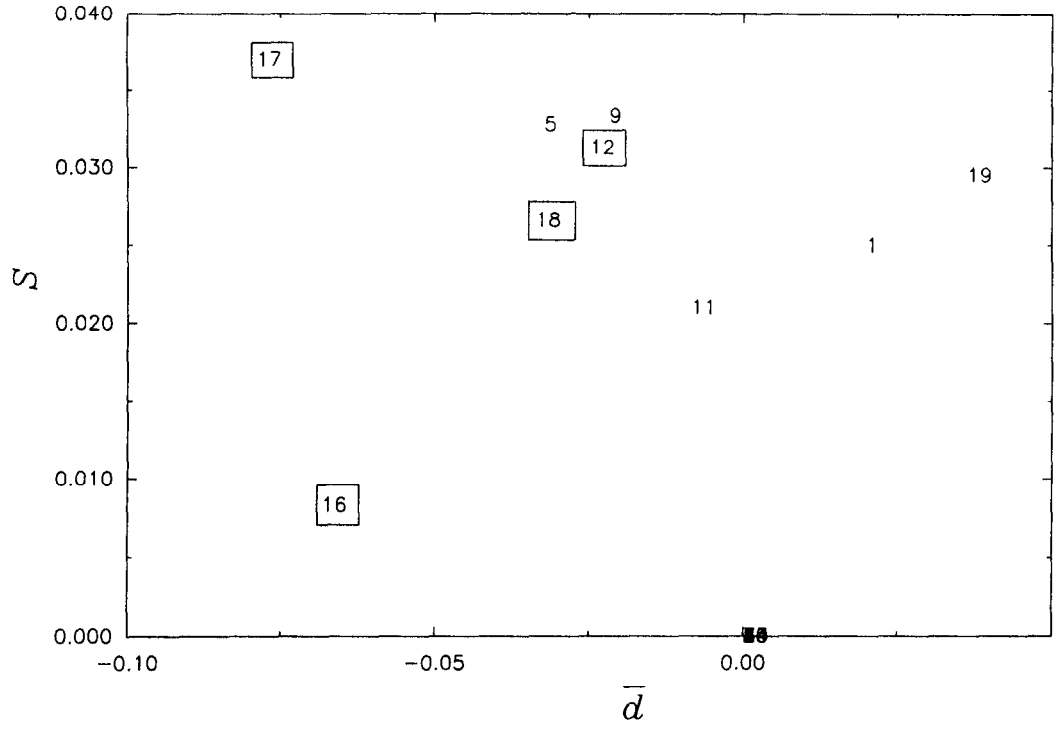

Figure 4. Estimated Means ($\bar{d}_i$) and Standard Deviations (S_i) of the Distributions of Elementary Effects in the Example of Section 5. Inputs for which $|\bar{d}_i| > 2\ SEM_i$ are marked in boxes.

Of the remaining inputs, numbers 16 and 17 (melt temperature of the right-layer material and width of the left layer) have consistently large negative elementary effects. Inputs 1, 5, 9, 11, 12, 18, and 19 also have nonzero means, but each has a substantial standard deviation relative to its mean, indicating potentially extensive patterns of interaction or curvature in the effects. This is at least partially caused by the discreteness of the code noted previously.

For comparison, a second experiment was also run using a Latin hypercube design in 108 runs. Regression of the rank-transformed response values on the 20 inputs yielded an R^2 value of .91 and in this case led to conclusions similar to those just described. The inputs associated with model coefficients significant at the .05 level are nearly the same as the nine previously indicated as important; the differences are that input 18 is not significant at the .05 level and input 15 is. It is interesting to note that input 18 was the single input for which both positive and negative elementary effects were observed in the first analysis. Inputs 4, 8, and 18 would also be selected as important if a significance level of .10 were used, although if about 10 inputs really are unimportant, the less stringent significance level could easily lead to at least one "false positive." The ranges of response values observed in the two experiments were similar, about .2 to .4. It is interesting to note, however, that all 108 response values in the Latin hypercube experiment were distinct. On the other hand, if only the first two significant digits of y are considered, only 10 values are distinct in this data set compared to 13 in the first experiment.

6. SUMMARY AND DISCUSSION

In this article, experimental plans have been described for the purpose of collecting random samples from the distributions of elementary effects associated with each input. These plans require about one model evaluation per elementary effect sampled for independent random sampling and can be constructed with fewer evaluations per elementary effect for cluster sampling. For a given fixed sample size to be obtained from each F_i, these plans then require a total number of runs, which is a linear function of k, as opposed, for example, to resolution V fractional factorial plans for which the number of runs required is of order k^2.

There are certainly situations, however, in which even this number of evaluations is considered too large for an exploratory study. When model evaluations are extremely expensive or k is very large, it may be important to learn something based on an experiment in which n may be considerably less than k, even though it must be recognized that this information will not be as complete or conclusive as what might be desired. In the framework of more

conventional fractional factorial experiments, Watson (1961) and others developed the idea of group screening for instances in which it is believed that most factors have little or no effect on the response. The basic idea is to intentionally confound factors in groups for the purpose of screening. For example, if k factors are completely confounded within groups of size 3, a screening design for the $k/3$ "group factors" may be used to isolate the relatively few important factors, which are investigated more thoroughly in subsequent experimentation. Typically, when the effect due to a group factor is negligible, it is concluded that the individual constituent factors are negligible. Although this may be an incorrect conclusion in some cases, it often effectively isolates most of the important factors quickly when interactions are absent, grouping of factors is randomized, and the effect sparsity assumption is correct. A similar approach to group screening could be based on the designs presented here. If inputs are confounded in groups of size 3, this leads to three-at-a-time sampling matrices, which can be constructed in approximately $\frac{1}{3}$ of the original number of runs. Then with some chance of error the list of potentially important inputs may be shortened earlier.

It must be admitted that there is something intuitively unappealing about an analysis that selectively "ignores" information. Once $\mathbf{x}$ has been selected and the resulting y observed, the modeler cannot be blamed for wanting to use all of the $\mathbf{x}$ vector in the analysis. The suggested practice of looking at sample statistics from each F_i is clearly not a fully efficient analysis in the sense that it ignores the "location" of the sampled elementary effects in ω. When one or more of the samples from the F_i contain unusual values, their location can be observed after sampling and some conclusions may be drawn about the nature of interaction. For example, if the observed d_4 values are always unusually high when $x_8 = 1$ and $x_{11} = 0$, this hints at the existence of an important three-factor interaction. This line of thinking is more akin to the analysis proposed by Satterthwaite, however, and shares its weakness of trying to make fine distinctions about the specific nature of inputs given (relatively) very little data.

It should be reiterated for clarity that the problem treated here is fundamentally different from the problem addressed by McKay et al. (1979) in their introduction of the LHS, in which the primary objective was the efficient empirical determination of the distribution of y that follows from a known multivariate distribution of $\mathbf{x}$. This is often appropriate when inputs represent uncontrolled quantities that vary over time or space, in which case the distribution attached to the inputs may have a frequentist interpretation. Alternatively, the inputs may represent fixed but unknown physical constants, and the input distribution may reflect subjective uncertainty about their values. Our interest here is in the more fundamental issue of which inputs influence the output in an important way, leaving aside questions of how inputs, or our assessments of them, might change with time or improved measurements. This philosophy is consistent with applications in which codes like TWOLAYER are used; most inputs represent physical characteristics that can be controlled more or less exactly by the experimenter, and the goal is to identify which of them can be used to adjust important characteristics (outputs of the code) of the modeled system.

An important aspect of real computational models not discussed here is the typically large number of output variables produced by a single model run. These may represent different modeled characteristics of the system of interest or values of the same characteristic varying over simulated time. In fact the utility index of the TWOLAYER example is a scalar function of a multivariate output representing the phase of material at different points in space over a length of time. Such summary functions can often be used to reduce the cumbersome dimensionality of model output while preserving information about the most important aspects of the model. Alternatively, an approach of the type suggested here could be used to individually analyze several output variables recorded in the same computational experiment. More sophisticated approaches, which could take advantage of any relationships existing among output variables, would certainly be useful.

The underlying challenge in the input screening problem is that of identifying as many of the interesting effects of $\mathbf{x}$ on y as possible using as few runs as possible. When it is reasonable to assume some simplicity in the structure of y—for example, effect sparsity and monotonicity—screening methods based on Latin hypercube designs or fractional factorials are effective and economical. The approach suggested here leads to reliable information about how each element of $\mathbf{x}$ influences y, whether properties of simplicity are present or not, and is particularly useful when an assumption that these properties hold cannot be justified.

ACKNOWLEDGMENTS

This research was sponsored by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy, through Contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. I thank three referees whose comments on an earlier version of this article have led to major improvements in the presentation.

I am also grateful to Dr. Alan Solomon for providing the TWOLAYER model used in the example of Section 5.

APPENDIX: A PROPERTY OF MATRICES DESCRIBED BY EQUATIONS (4) AND (5)

Theorem. A sampling matrix $\mathbf{B}$ of $m = k + 1$ rows can provide one elementary effect for each input iff it can be rewritten in the form of (4) and (5) by reordering columns and exchanging 0's and 1's in some columns.

Proof. The proof is based on the following observation: If any sampling matrix of $m = k + 1$ runs is divided into two groups of runs, there must be at least one "split" pair of runs (i.e., one in each group) that differ only in the value of a single input. To see this, suppose that it is not so and that such a sampling matrix has been partitioned into two such groups without a split pair. Since there are no split pairs, it is possible to exchange 0's and 1's for some of the inputs within one group without changing the number of elementary effects which can be obtained from each F_i. Furthermore, this can clearly be done in such a way that some run in the first group becomes identical to some run in the second group. But then one of these two runs could be deleted without diminishing the number of elementary effects that can be observed. This is impossible, however, because the reduced sampling matrix could be used as a resolution III design matrix for k factors in fewer than $k + 1$ runs.

Based on this observation, the theorem can be proven easily by construction. Given any sampling matrix of $m = k + 1$ runs, pick any run and exchange 0's and 1's for some inputs (throughout the design) so as to transform this run to a vector of 0's; call this the "first run." The preceding observation implies that there must be at least one run in the remainder of the design that differs from the first run by only the values of a single input. Find such a run, reorder the columns of the design so that input that differs is the first input, and call this the "second run." Now the observation implies that there must be at least one run in the remainder of the design that differs from either the first or second run by only the values of a single input. Find such a run, reorder the columns 2–k of the design so that input which differs is the second input, and call this the "third run." This process is continued through identification of a "(k + 1)st run"; the resulting design will have the properties described.

[*Received September 1989. Revised November 1990.*]

REFERENCES

Baker, F. D., and Bargmann, R. E. (1985), "Orthogonal Central Composite Designs of the Third Order in the Evaluation of Sensitivity and Plant Growth Models," *Journal of the American Statistical Association*, 80, 574–579.

Box, G. E. P., and Meyer, R. D. (1986), "An Analysis for Unreplicated Fractional Factorials," *Technometrics*, 28, 11–18.

Box, G. E. P., and Wilson, K. B. (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society*, Ser. B, 13, 1–45.

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

Currin, C., Mitchell, T. J., Morris, M. D., and Ylvisaker, D. (in press), "Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86.

Daniel, C. (1973), "One-at-a-Time Plans," *Journal of the American Statistical Association*, 68, 353–360.

Iman, R. L., and Conover, W. J. (1980), "Small Sample Sensitivity Analysis Techniques for Computer Models, With an Application to Risk Assessment" (with discussion), *Communications in Statistics, Part A—Theory and Methods*, 9, 1749–1842.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code," *Technometrics*, 21, 239–245.

Plackett, R. L., and Burman, J. P. (1946), "The Design of Optimal Multifactorial Experiments," *Biometrika*, 33, 305–325.

Sacks, J., Schiller, S. B., and Welch, W. J. (1989), "Designs for Computer Experiments," *Technometrics*, 31, 41–47.

Satterthwaite, F. E. (1959), "Random Balance Experimentation," *Technometrics*, 1, 111–137.

Satterthwaite, F. E., and Budne, T. A. (1959), "Author's Response to Discussions," *Technometrics*, 1, 184–193.

Watson, G. S. (1961), "A Study of the Group Screening Method," *Technometrics*, 3, 371–388.

Webb, S. R. (1968), "Saturated Sequential Factorial Designs," *Technometrics*, 10, 535–550.

Youden, W. J., Kempthorne, O., Tukey, J. W., Box, G. E. P., and Hunter, J. S. (1959), "Discussion of the Papers of Messrs. Satterthwaite and Budne," *Technometrics*, 1, 157–184.