Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

A TUTORIAL ON DESIGN OF EXPERIMENTS FOR SIMULATION MODELING

Averill M. Law

Averill M. Law & Associates, Inc.
4729 East Sunrise Drive, #462
Tucson, AZ 85718, USA

ABSTRACT

Simulation models often have many input factors, and determining which ones have a significant impact on performance measures (responses) of interest can be a difficult task. The common approach of changing one factor at a time is statistically inefficient and, more importantly, is very often just *incorrect*, because for many models factors interact to impact on the responses. In this tutorial we present an introduction to design of experiments specifically for simulation modeling, whose major goal is to determine the important factors often with the least amount of simulating. We discuss classical experimental designs such as full factorial, fractional factorial, and central composite followed by a presentation on Latin hypercube designs, which are designed for the complex, nonlinear responses typically associated with simulation models.

1 INTRODUCTION

In this tutorial we discuss the use of statistical experimental design techniques when the “experiment” is the execution of a computer simulation model. In experimental-design terminology, the input parameters and structural assumptions composing a model are called *factors*, and output performance measures are called *responses*. Factors can be *quantitative* or *qualitative* (also called *categorical*). Quantitative factors naturally assume numerical values (e.g., the number of machines in a workstation), while qualitative factors represent structural assumptions that are not naturally quantified (e.g., a queue discipline that can be first-in, first-out or shortest-job first).

The major goal of experimental design in simulation is to determine which factors have the greatest effect on a response, and often to do so with the least amount of simulating. This is called *factor screening* and is typically performed using 2^k factorial designs or 2^{k-p} fractional factorial designs, which are discussed in Sections 2 and 3, respectively.

After we learn which factors are important and how they impact on the response, we are often interested in developing a *metamodel* (a simple model of the simulation model) based on the significant factors. Metamodels, which are discussed in Section 4, are used for the following purposes:

- *Gain further insight* into how changing a factor impacts on the response.
- *Predict* the model response for system configurations that were not simulated, since the setup or execution time for the model might be large, or because answers are needed in real time.
- Find that combination of input-factor values that *optimizes* (i.e., maximizes or minimizes) a response.

Metamodels are usually given in the form of a low-order polynomial regression equation.

978-1-5386-3428-8/17/$31.00 ©2017 IEEE

Finally, in Section 5 we discuss the potential dangers of using experimental designs or analyses that were designed for physical experiments (e.g., analysis of a manufacturing process) in the context of simulation modeling. This paper is based on Chapter 12 of Law (2015) and on a three-day short course on the same topic that the author has given since 2007. A general reference on design and analysis of experiments is Montgomery (2013). Papers and books that discuss experimental design in the context of discrete-event simulation include Ankenman et al. (2010), Barton (2013), Kleijnen (2015), Kleijnen et al. (2005), and Sanchez and Wan (2015).

2 2^k FACTORIAL DESIGNS

Suppose that there are $k(k \geq 2)$ factors and we want to get an initial estimate of how each factor affects the response. We might also want to determine if the factors interact with each other, i.e., whether the effect of one factor on the response depends on the *levels* of the others. One way to measure the effect of a particular factor would be to fix the levels of the *other* $k-1$ factors at some set of values and make simulation runs at each of two levels of the factor of interest to see how the response reacts to changes in this single factor. The whole process is then repeated to examine each of the other factors, one at a time. This strategy, which is called the *one-factor-at-a-time* (OFAT) *approach*, is quite inefficient in terms of the number of simulation runs needed to obtain a specified precision (see Montgomery 2013). More importantly, it does not allow us to measure any interactions; indeed, it assumes that there are no interactions, which is often not the case in simulation applications.

Example 1. Suppose that we have two factors *A* and *B*. Let the baseline levels of these two factors be A^- and B^-. Also, A^+ and B^+ be proposed levels for these factors. Then the OFAT method would specify simulating the following *three* combinations of *A* and *B*:

$$A^-, B^- \text{ (baseline)}$$
$$A^+, B^- \text{ (change } A\text{)}$$
$$A^-, B^+ \text{(change } B\text{)}$$

resulting in the responses $R(A^-,B^-), R(A^+,B^-)$, and $R(A^-,B^+)$. Then the effect on the response of changing factor *A* from A^- to A^+ would be computed as

$$R(A^+,B^-) - R(A^-,B^-) \tag{1}$$

However, this calculation is based *only* on factor *B* being at its B^- level. It could be, though, that the effect on the response of changing factor *A* would be quite different if factor *B* were at its B^+ level (i.e., if the factors interact); see Example 3 for a numerical example. (A similar discussion applies to factor *B*.)

If we had also simulated the combination A^+, B^+ resulting in the response $R(A^+,B^+)$, then the effect on the response of changing factor *A* could also be computed as

$$R(A^+,B^+) - R(A^-,B^+) \tag{2}$$

However, this last calculation would *not* actually be possible under the OFAT strategy, since A^+, B^+ would have not been simulated. For 2^2 factorial designs, which will be discussed next, the average of the differences given by expressions (1) and (2) will be used to estimate the effect on the response of moving factor *A* from its A^- level to its A^+ level.

A much more economical strategy for determining the effects of factors on the response with which we can also measure interactions, called a 2^k *factorial design*, requires that we choose just *two* levels for each factor and then calls for simulation runs at each of the 2^k possible factor-level combinations, which are called *design points*. We associate a minus sign with one level of a factor and a plus sign with the other. The levels, which should be chosen in consultation with subject-matter experts, should be far enough apart that we would expect to see a difference in the response, but not so separated that nonsensical combinations are obtained. Because we are using only two levels for each factor, we assume that the response is approximately linear (or at least monotonic) over the range of the factor. (If the response is nonmonotonic over the range, then we might be misled into thinking that the factor has no effect on the response.) We will discuss a method for testing the linearity assumption in Section 4.

The form of a 2^k factorial design can be compactly represented in tabular form, as in Table 1 for $k = 3$. The variable R_i for $i = 1, 2, \ldots, 8$ is the value of the response when running the simulation with the *i*th combination of the factor levels. For example, R_4 is the response resulting from running the simulation with factors 1 and 2 at their respective "+" levels and factor 3 at its "-" level. We shall see later that writing down this array, called the *design matrix*, facilitates calculation of the factor effects and interactions.

Table 1: Design matrix for a 2^3 factorial design.

Factor combination (design point)	Factor 1	Factor 2	Factor 3	Response
1	-	-	-	R_1
2	+	-	-	R_2
3	-	+	-	R_3
4	+	+	-	R_4
5	-	-	+	R_5
6	+	-	+	R_6
7	-	+	+	R_7
8	+	+	+	R_8

The *main effect* of factor *j*, denoted by e_j, is the *difference* between the average response when factor *j* is at its "+" level and the average response when it is at its "-" level. For the 2^3 design of Table 1, the main effect of factor 1 is thus

$$e_1 = \frac{R_2 + R_4 + R_6 + R_8}{4} - \frac{R_1 + R_3 + R_5 + R_7}{4}$$

which can be rewritten as

$$e_1 = \frac{-R_1 + R_2 - R_3 + R_4 - R_5 + R_6 - R_7 + R_8}{4}$$

Thus, to compute e_1, we simply apply the signs in the "Factor 1" column of Table 1 to the corresponding R_i's, add them up and divide by 4. A geometric interpretation of main effects is given on page 243 in Montgomery (2013).

The main effects measure the average change in the response due to a change in an individual factor. It could be, though, that the effect of factor j_1 depends in some way on the level of some other factor j_2, in which case the factors are said to *interact*. A measure of the interaction, denoted by $e_{j_1 j_2}$, is the difference between the average response when factors j_1 and j_2 are at the same (both "+" or both "-") level and the average response when they are at opposite levels. It is also called the $j_1 \times j_2$ *interaction*. For the 2^3 design, the 1×2 interaction effect is given by

$$e_{12} = \frac{R_1 + R_4 + R_5 + R_8}{4} - \frac{R_2 + R_3 + R_6 + R_7}{4}$$

which can be rewritten as

$$e_{12} = \frac{R_1 - R_2 - R_3 + R_4 + R_5 - R_6 - R_7 + R_8}{4}$$

Thus, if we create a new column labeled "1×2" of eight signs by "multiplying" the *i*th sign in the "Factor 1" column by the *i*th sign in the "Factor 2" column (the product of like signs is a "+" and the product of opposite signs is a "-"), we get a column of signs that gives us precisely the signs of the R_i's used to form e_{12}; as with main effects the divisor is 4.

Although its interpretation becomes more difficult, we can also define the three-way interaction, e_{123}, as follows:

$$e_{123} = \frac{-R_1 + R_2 + R_3 - R_4 + R_5 - R_6 - R_7 + R_8}{4}$$

This expression for e_{123} is obtained by multiplying the *i*th signs from the columns for factors 1, 2, and 3 in Table 1, applying them to the R_i's and summing, and then dividing by 4.

If two- or three-way interactions appear to be present for a 2^3 factorial design, then the main effect of each factor involved in such a significant interaction *cannot* be interpreted as simply the effect in general of moving that factor from its "-" level to its "+" level, since the magnitude and possibly the sign of the change in the response depend on the level of at least one other factor.

Example 2. Customers arrive to a company and want to buy a product. Suppose that the *interarrival times* of customers are exponentially distributed with mean 0.1 month. The *demand size* of an arriving customer is 1, 2, 3, or 4 items with respective probabilities 1/6, 1/3, 1/3, and 1/6. The company uses an (s, S) *inventory policy* to decide when and how much to order from its supplier, where *s* is the *reorder point* and *S* is the *target amount*. In particular, let I_i be the inventory level at the beginning of the *i*th month, where I_i can be positive, zero, or negative. If $I_i < s$, then the company will order $Z = S - I_i$ items and incur an *ordering cost* of $O_i = K + iZ$, where *K* = \$32 is the *setup cost* and *i* = \$3 is the *incremental cost* per item ordered. If $I_i \geq s$, then no order is placed and the company incurs no ordering cost. The *delivery lag*, which is the time from when the company places an order with its supplier until it actually arrives, is assumed to be uniformly distributed between 0.5 and 1 month.

If the demand size for a particular customer is greater than the current inventory level, then the shortage is *backlogged* and satisfied by future orders. (In this case, the new inventory level is the old inventory level minus the demand size, resulting in a negative inventory level.)

The company also incurs a *holding cost*, H_i, in month i that is \$1 times the average number in items *physically* in inventory (nonnegative) in month *i*. The holding cost includes such things as warehouse rental, insurance, taxes, etc. There is also a *shortage cost*, S_i, in month i that is \$5 times the average shortage level (nonnegative) in month *i*. The shortage cost includes the extra record keeping when a backlog exists, as well as loss of customers' goodwill. Then the total cost, C_i, of operating the inventory system in month i is given by

$$C_i = O_i + H_i + S_i$$

The response of interest is the *average total cost per month* over a 120-month planning horizon, and the initial inventory level is assumed to be 60.

For the sake of experimental design, it is convenient to reparameterize the inventory model slightly in terms of the ordering policy. Specifically, we will now take the factors to be the reorder point s and the difference $d = S - s$. (Clearly, $S = s + d$, so the two parameterizations are equivalent.) The "low" and "high" values we chose for these new factors are given in the *coding chart* in Table 2. (If we had used s and S in the coding chart, then we would have obtained nonsensical inventory policies such as $(20,10)$.)

Table 2: Coding chart for *s* and *d* in the inventory model.

Factor	-	+
s	20	60
d	10	50

We simulated each of the four design points, and the relevant design matrix and corresponding response values are given in Table 3, together with an extra column of signs to be applied in computing the $s \times d$ interaction. Each R_i is the average cost from a single 120-month replication. The main effects are

$$e_s = \frac{-144.16 + 144.50 - 119.99 + 147.00}{2} = 13.68$$

and

$$e_d = \frac{-144.16 - 144.50 + 119.99 + 147.00}{2} = -10.84$$

and the $s \times d$ interaction effect is

$$e_{sd} = \frac{144.16 - 144.50 - 119.99 + 147.00}{2} = 13.34$$

Thus, the average effect of raising *s* from 20 to 60 was to increase the monthly cost by 13.68, and raising *d* from 10 to 50 decreased the monthly cost by an average of 10.84. Therefore, it appears that the smaller value of *s* and the larger value of *d* would be preferable, since lower monthly costs are desired. Since the $s \times d$ interaction is positive, there is further indication that lower costs are observed by setting *s* and *d* at opposite levels. However, if this interaction is present in a significant way (a question addressed in Example 3 below), then the effect that *s* has on the response depends on the level of *d*, and vice versa.

Table 3: Design matrix and simulation results for the 2^2 factorial design on *s* and *d*, inventory model.

Factor combination (design point)	s	d	$s \times d$	Response
1	-	-	+	144.16
2	+	-	-	144.50
3	-	+	-	119.99
4	+	+	+	147.00

Since the R_i's are random variables, the observed effects are random also. To determine whether the effects are “real,” as opposed to being explainable by sampling fluctuation, we must determine if the effects are statistically significant. This is often addressed in the experimental-design literature by performing an *analysis of variance* (see Montgomery 2013), which assumes that the response has the same population variance for each design point. However, as we will see in Example 3, this is usually *not* a good assumption in simulation modeling. We will, therefore, take the simple approach of replicating the *whole design n* times to obtain *n* independent and identically distributed (IID) values of each effect. These values can then be used to construct confidence intervals for the *expected* effects. For example, let e_j^i be the observed main effect of factor *j* on replication *i*, for $i = 1, 2, \ldots, n.$ Let

$$\overline{e}_j(n) = \frac{\sum_{i=1}^{n} e_j^i}{n}$$

and

$$S_j^2(n) = \frac{\sum_{i=1}^{n} \left[e_j^i - \overline{e}_j(n) \right]^2}{n-1}$$

Then an (approximate) $100(1-\alpha)$ percent confidence interval for the expected main effect $E(e_j)$ is given by

$$\overline{e}_j(n) \pm t_{n-1,1-\alpha/2} \sqrt{S_j^2(n)/n}$$

where $t_{n-1,1-\alpha/2}$ is the upper $1-\alpha/2$ critical point for a *t* distribution with $n-1$ degrees of freedom. If the confidence interval for $E(e_j)$ does not contain 0, we conclude that the effect is *statistically significant*; otherwise, we have no statistical evidence that it is actually present. We can construct a confidence interval for an expected interaction effect in a similar manner. We must also keep in mind that statistical significance of an effect does not necessarily imply that its magnitude is *practically* significant (i.e., large enough to make a tangible difference).

Example 3. We replicated the entire 2^2 factorial design of the inventory model in Example 2 n = 10 times, and Table 4 gives the sample mean and variance of the responses across the 10 replications for each of the four design points. Note that the largest and smallest sample variances differ by a factor of approximately 14. Based on the 10 IID values of each of the three effects that we obtained, Table 5 gives 95 percent confidence intervals for $E(e_s), E(e_d)$, and $E(e_{sd})$. All effects appear to be real since their confidence intervals do not contain zero. If we could interpret the main effects literally, we would expect the average cost per month to increase by 18.47 when we move *s* from 20 to 60, and to decrease by 6.02 when we move *d* from 10 to 50. However, since there is a significant interaction between *s* and *d*, these main effects actually provide a limited amount of information.

Table 4: Sample means and variances of the responses for the inventory model.

Design point	Sample mean	Sample variance
$s = 20, d = 10$	135.71	22.24
$s = 60, d = 10$	143.94	2.26
$s = 20, d = 50$	119.45	15.07
$s = 60, d = 50$	148.17	1.60

Table 5: 95 percent confidence intervals for the expected effects, inventory model.

Expected effect	95 percent confidence interval
$E(e_s)$	18.47 ± 2.33
$E(e_d)$	-6.02 ± 2.23
$E(e_{sd})$	10.25 ± 2.60

In Figure 1 we give an *interaction plot* for *s* and *d*, where the presence of an interaction is indicated by the *nonparallel lines* (see Equation (8)). In particular, when $d = 10,$ moving *s* from 20 to 60 increases the average cost by 8.23 (see Table 4). However, when $d = 50,$ moving *s* from 20 to 60 increases the average cost by 28.72. Note that the OFAT approach would give 8.23 as the increase in average cost resulting from moving *s* from 20 to 60, whereas we got 18.47 from the 2^2 factorial design.

We conclude from Figure 1 that both *s* and *d* have a significant effect on the average cost per month. However, the actual numerical change in the average cost due to changing *s* depends on the level of *d*, and vice versa; this will be discussed further in Section 4.

It should be mentioned that a factor can be important even if the magnitude of its main effect is small, since it might have a significant interaction with another factor (see pages 647-648 in Law 2015).

3 2^{k-p} FRACTIONAL FACTORIAL DESIGNS

It is clear that 2^k factorial designs may become unaffordable if k is large. For example, k = 11 factors would require $2^{11} = 2048$ design points.

Fractional factorial designs provide a way to get good estimates of the main effects and perhaps two-factor interactions at a fraction of the computational effort required by a full 2^k factorial design. Basically, a 2^{k-p} fractional factorial design is constructed by choosing a certain subset (of size 2^{k-p}) of all the 2^k

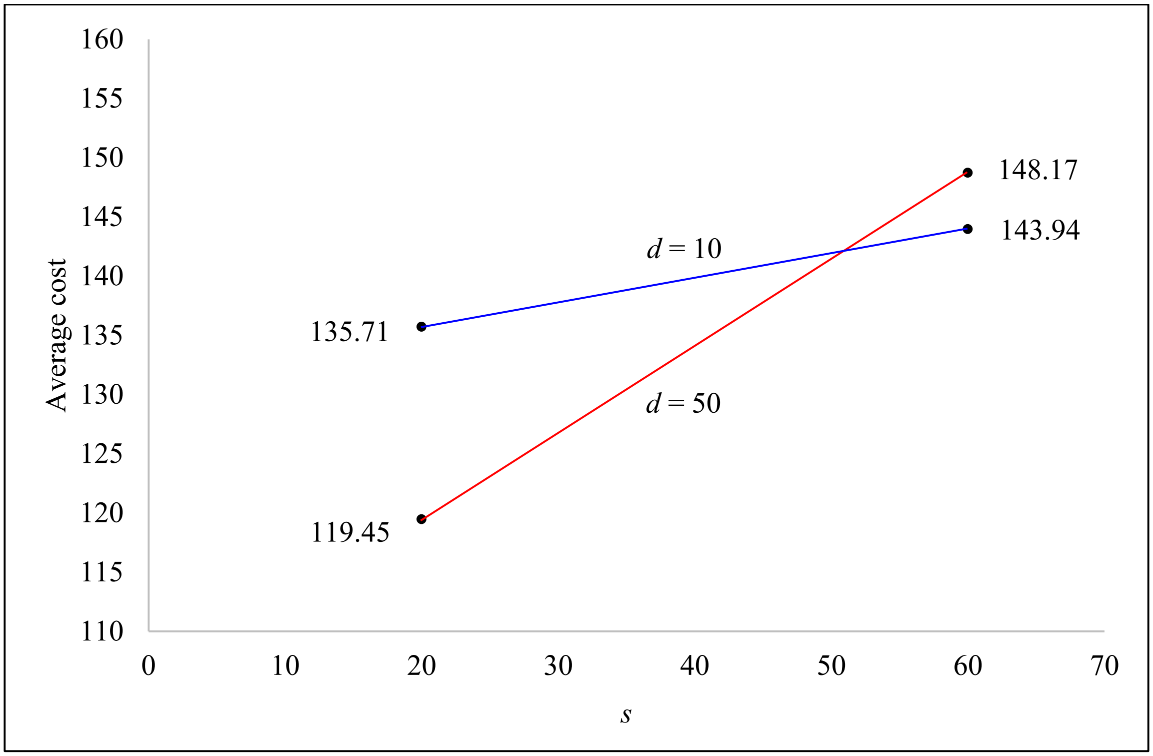

Figure 1: Interaction plot for factors *s* and *d*, inventory model.

possible design points and then running the simulation for only these chosen points. Since only $1/2^p$ of the possible 2^k factor combinations are actually run, we sometimes speak of a "half fraction" if $p=1$, a "quarter fraction" if $p=2$, and so on. Clearly, we would like p to be large from a computational viewpoint, but a larger p may also result in less information from the experiment, as one might expect.

Consider again the k = 11 factor example above. If we are willing to assume that three-way and higher-way interactions are negligible, then we could use what is called a resolution V fractional factorial design (see page 650 in Law 2015) requiring only 128 design points to get "clear" estimates of main effects and two-way interactions

Unfortunately, for some simulation models there are, in fact, significant three-way and four-way interactions, and fractional factorial designs give biased estimates of two-way interactions and even main effects, limiting their usefulness.

4 METAMODELS AND RESPONSE SURFACES

We discuss metamodeling using the inventory model of Examples 2 and 3, where all effects were found to be statistically significant. Let *E*[*R*(*s*, *d*)] denote the expected average cost per month for particular values of the reorder point, *s*, and the difference, *d*. Then in a 2^2 factorial design, we are in fact assuming that *E*[*R*(*s*, *d*)] can be represented by the following *regression model* (see chapter 10 in Montgomery 2013):

$$E[R(s,d)] = \beta_0 + \beta_s x_s + \beta_d x_d + \beta_{sd} x_s x_d \tag{3}$$

where $\beta_0, \beta_s, \beta_d$, and β_{sd} are coefficients, and x_s and x_d are *coded variables* for the factors that we now define. In particular, let $\bar{s}$ and $\bar{d}$ be the average values of *s* and *d* (called the *natural variables* for the factors) in Table 2; that is, $\overline{s} = 40$ and $\overline{d} = 30$. Also, let Δs and Δd be the differences between the "-" and "+" levels for *s* and *d*, respectively, so that $\Delta s = 40$ and $\Delta d = 40$. Then the coded variables for *s* and *d* are defined by

$$x_s = \frac{2(s-\overline{s})}{\Delta s} = \frac{s-40}{20} \tag{4}$$

and

$$x_d = \frac{2(d-\overline{d})}{\Delta d} = \frac{d-30}{20} \tag{5}$$

Note that Equation (4) maps $s = 20$ into $x_s = -1$ and $s = 60$ into $x_s = +1$. Similarly, (5) maps $d = 10$ into $x_d = -1$ and $d = 50$ into $x_d = +1$. Coded variables are commonly used in experimental design because the effect on the response of a change in a factor is always measured relative to the range -1 to +1.

Suppose that $\overline{e}_s(10)$, $\overline{e}_d(10)$, and $\overline{e}_{sd}(10)$ are the effect estimates from the *n* = 10 independent replications of Example 3. Also, let $\overline{R}_F(10)$ be the average response over the four factorial (denoted by *F*) design points and over the 10 replications. Then *least-squares estimators* (see pages 280-282 in Montgomery 2013) are given by

$$\hat{\beta}_0 = \overline{R}_F(10),\ \hat{\beta}_s = \frac{\overline{e}_s(10)}{2},\ \hat{\beta}_d = \frac{\overline{e}_d(10)}{2},\ \hat{\beta}_{sd} = \frac{\overline{e}_{sd}(10)}{2} \tag{6}$$

The reason that a regression coefficient (other than $\hat{\beta}_0$) is one-half of the effect estimate is that a regression coefficient measures the effect of a unit change in *x* on the mean *E*[*R*(*s*, *d*)], while the effect estimate is based on a two-unit change (from -1 to +1).

Substituting the estimated coefficients from Equation (6) into the model (3), we obtained the following *fitted* regression model in the coded variables x_s and x_d:

$$\hat{R}(s,d) = 136.819 + 9.237x_s - 3.009x_d + 5.123x_s x_d \tag{7}$$

Note that the coefficients 9.237, -3.009, and 5.123 are, in fact, one-half of the effect estimates in Table 5 (up to roundoff). Putting x_s and x_d as given by (4) and (5) into the model given by (7), we get the following equivalent regression model in the natural variables *s* and *d*:

$$\hat{R}(s,d) = 138.226 + 0.078s - 0.663d + 0.013sd \tag{8}$$

Equation (8) is a model of how the simulation transforms the input parameters *s* and *d* into the output response $\hat{R}(s,d)$, and it is called a *metamodel* (i.e., a model of the simulation model). We plot $\hat{R}(s,d)$ as given by (8) in Figure 2; this plot was made using the Design-Expert experimental design software (see Stat-Ease 2017). This plot, which is called a *response surface*, is a "twisted plane" because of the interaction term in (8). Note that if the coefficient of *sd* in (8) were 0, then the effect of *s* on the response would not depend on *d*, and vice versa; i.e., there would be no interaction between *s* and *d*.

The metamodel (8) could be regarded as a proxy for the full simulation model's response surface; all we would need is a pocket calculator or spreadsheet to evaluate it for any (*s*, *d*) pair of interest. We must remember, though, that (8) is just an approximation to the actual simulation and may thus be inaccurate, especially far from the values of *s* and *d* that provided the data on which it is based. A

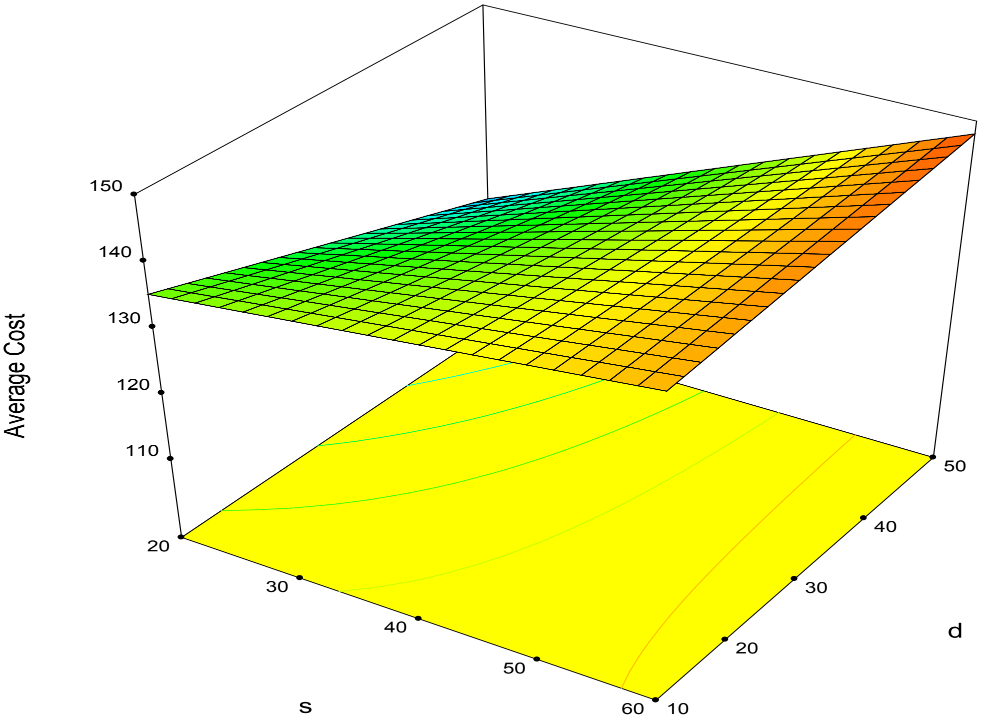

Figure 2: Response-surface plot from the 2^2 factorial design, inventory model.

metamodel, after all, is itself a model, and as such may or may not be valid relative to the simulation model.

The regression model given by Equation (3) assumes that the response is a *linear* function in the coded variables. However, in some cases the simulation model's response might be better represented by the following *quadratic* (or second-order) regression model:

$$E[R(s,d)] = \beta_0 + \beta_s x_s + \beta_d x_d + \beta_{sd} x_s x_d + \beta_{ss} x_s^2 + \beta_{dd} x_d^2 \tag{9}$$

To determine whether the model given by (3) is a good approximation to the simulation model's response surface or whether the second-order model given by (9) is necessary, we made n = 10 independent replications of the simulation at the *center point* (denoted by *C*), $x_s = 0$ and $x_d = 0$ (or, equivalently, s = 40 and d = 30), and we obtained an average response of $\bar{R}_C(10) = 122.95$. Substituting $x_s = 0$ and $x_d = 0$ into (7) gives $\bar{R}_F(10) = 136.82$, which is the *predicted* average response for the model at the center point. Thus, we get a difference of $\bar{R}_F(10) - \bar{R}_C(10) = 13.87$, which turns out to statistically significant (see pages 661-662 in Law 2015 for additional details). Thus, it appears that quadratic (or higher-order) curvature *is* present and the second-order model given by Equation (9) should be considered.

Unfortunately, we cannot uniquely estimate the six required coefficients in (9) because we have only collected data from five independent design points (i.e., four from the factorial design and one at the center point). Therefore, we will augment our five existing points with four *axial points.* The resulting

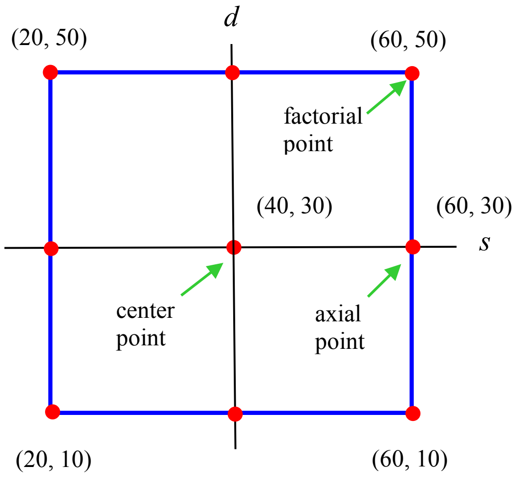

Figure 3: Face-centered central composite design for $k = 2$.

design, called a face-centered *central composite design* (CCD) and shown in Figure 3, will be used to fit the second-order model. We made n = 10 independent replications for each of the four axial points. From the data for all nine design points, we obtained the following fitted second-order model in the coded variables:

$$\hat{R}(s,d) = 122.803 + 9.518x_s - 1.959x_d + 4.627x_s x_d + 8.869x_s^2 + 5.205x_d^2 \qquad (10)$$

The equivalent second-order model in the natural variables is

$$\hat{R}(s,d) = 167.771 - 1.645s - 1.341d + 0.012sd + 0.022s^2 + 0.013d^2 \qquad (11)$$

Substituting $x_s = 0$ and $x_d = 0$ into (10), we get 122.80, which is very close to the average simulation response, $\bar{R}_C(10) = 122.95$, at the center point. In Figure 4 we give the response-surface plot corresponding to (11), and in Figure 5 we give a *contour plot* of the response surface, where all (s, d) points along a particular contour line would give approximately the same average-response value.

In order to validate the quadratic model given by (11), and to compare it more definitively to the first-order model with interaction term given by (8), we considered the four new design points (50, 20), (50, 40), (30, 20), and (30, 40). We made n = 10 independent replications of the simulation at each of these design points and the corresponding average responses are given in Table 6, along with the predicted average responses for each of the two metamodels. Also, given in the table for each design point and each metamodel is the percentage error in the predicted average response relative to the average simulation response. For example, for design point (50, 20), the average simulation response was 134.50 while the quadratic metamodel predicts that the average response will be 130.90, which represents a percentage error of 2.68 percent. It is clear from the table that the second-order metamodel given by (11) provides better predictions than does the first-order metamodel with interaction term given by (8), at least for the four new design points considered. Moreover, since the second-model appears to give “valid” predictions, we could now use it to predict the average response for other design points *within* our area of experimentation.

The Design-Expert software, which we have used to make many of our plots, employs a nonlinear programming algorithm (i.e., the Nelder-Mead simplex method) to try to find the factor levels that give

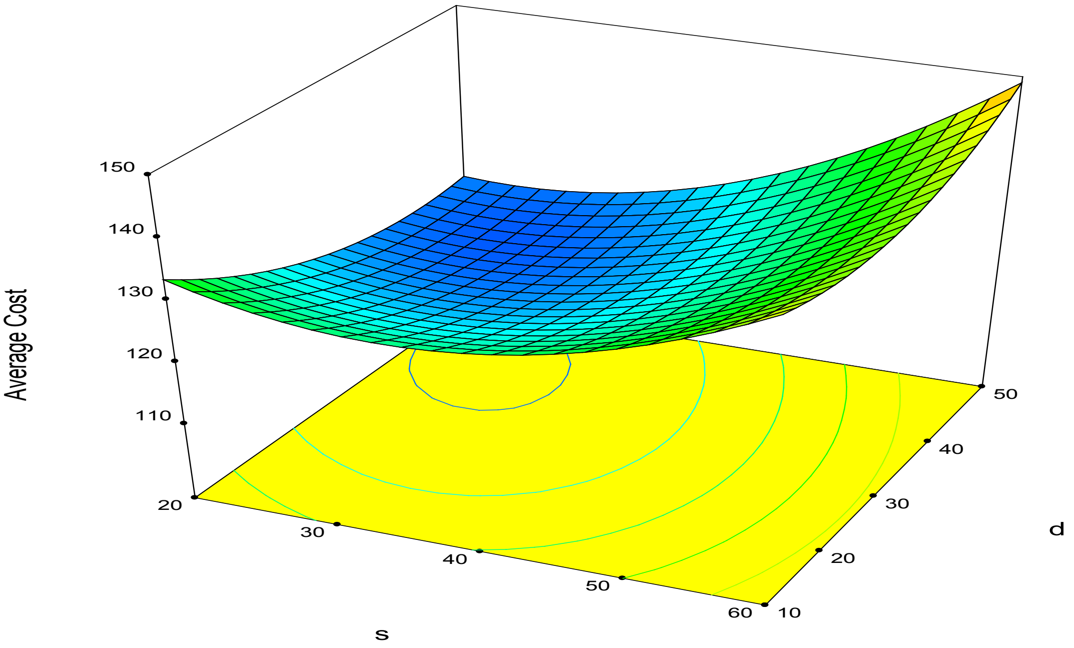

Figure 4: Response-surface plot of the second-order metamodel from the central composite design, inventory model.

the minimum (or maximum) response for a fitted metamodel. For the inventory model, we used Design-Expert to find the values of s and d that minimize the average response $\hat{R}(s,d)$ given by the metamodel (11), subject to the constraints $20 \le s \le 60$ and $10 \le d \le 50$. Design-Expert's optimization algorithm found the "optimal" factor levels to be s = 27 and d = 40, and the corresponding minimum average cost was $119.18 (see Figure 5).

If the response surface for a simulation model is "complex" in the interior of the experimental region (e.g., has nonmonotonic behavior), then a CCD may lead to a metamodel with poor predictive capabilities, because it only dictates sampling at the center point there. In this case we might consider the use of a Latin hypercube design, whose goal is to spread the design points "uniformly" throughout the experimental region. They require the factors to be continuous variables or discrete variables with potentially a large number of different levels.

In a *Latin hypercube design* (LHD) the design matrix has m rows and k columns, where m is the desired number of design points (levels for each factor). For a particular column (factor), the m levels are equally spaced between the lower and upper endpoints of the factor's range. Then each column is randomly permuted independently of every other column (see section 12.4.3 in Law 2015 for additional details). A rule of thumb for choosing the number of design points is $m = 10k$.

For the inventory model discussed above, we used the JMP statistical package (see SAS 2017) to generate a LHD with $m = 10(2) = 20$ design points in the experimental area defined by $20 \le s \le 60$ and $10 \le d \le 50$, and the resulting design points are shown in Figure 6. We made n = 10 independent replications of the simulation at each of the 20 design points and obtained the following fitted quadratic metamodel in the natural variables:

$$\hat{R}(s,d) = 141.197 - 0.834s - 0.774d + 0.006sd + 0.016s^2 + 0.009d^2 \qquad (12)$$

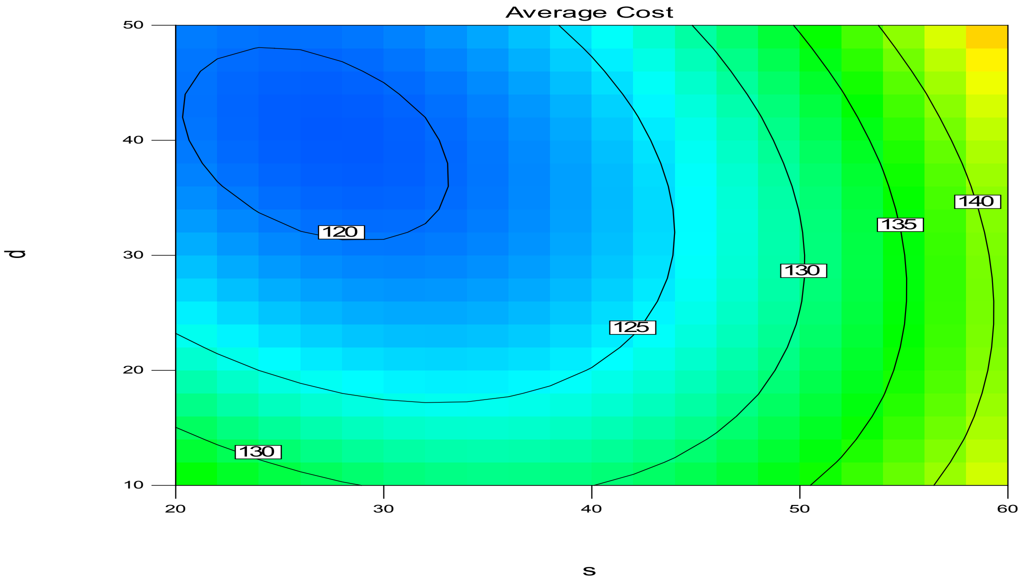

Figure 5: Contour plot of the second-order metamodel from the central composite design, inventory model.

We used the metamodel given by (12) to predict the average response at the each of the four design points given in Table 6 and we obtained values of 133.82, 134.51, 122.23, and 120.66, respectively, corresponding to metamodel errors of 0.51, 0.33, 0.28, and 0.79 percent. This is an average error of 0.48 percent across the four design points, whereas the average errors for the metamodels given by (11) and (8) were 1.79 and 7.20 percent, respectively.

Note that for some simulation models, the quality of the predictions provided by a LHD will be dramatically better than those given by a CCD (see page 678 in Law 2015). A LHD will also require fewer design points than a CCD for $k \geq 6$. For example, in the case of k = 10 a LHD requires 100 = 10(10) design points whereas a CCD requires $1045 = 2^{10} + 2(10) + 1$ design points.

Table 6: Comparison of the metamodels given by Equations (8) and (11) for four new design points, inventory model.

Design point	Average simulation response	Predicted average response from (8) and percentage error	Predicted average response from (11) and percentage error
(50, 20)	134.50	141.66 (5.32%)	130.90 (2.68%)
(50, 40)	134.97	141.21 (4.62%)	131.26 (2.75%)
(30, 20)	121.89	134.99 (10.75%)	123.70 (1.48%)
(30, 40)	119.71	129.42 (8.11%)	119.43 (0.23%)

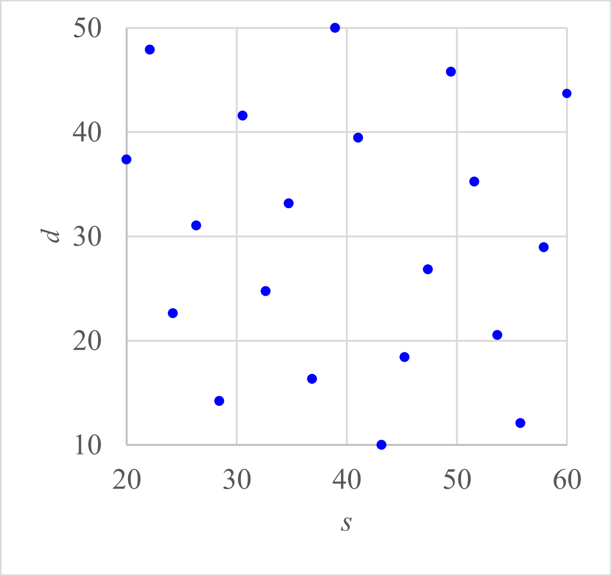

Figure 6: Latin hypercube design with m = 20 for the inventory model.

5 CONCLUSIONS AND SUMMARY

Design of experiments is a computationally efficient methodology for determining which model factors have an important impact on a response, taking into account interactions that may occur between factors. Note that in practice simulation models will typically have multiple responses, as illustrated in, for example, Sanchez et al. (2012). Metamodels based on the important factors may then be developed and are useful for predicting the model response for factor configurations that were not actually simulated, or because answers are needed in real time.

However, experimental designs and analyses developed for physical experiments are often blithely applied to simulation experiments, resulting in the following potential pitfalls:

- Two-level factorial and fractional factorial designs may produce misleading results, since simulation responses are often nonmonotonic functions of the factor levels.
- Fractional factorial designs may give significantly biased estimates of main effects and two-factor interactions because of the presence of large three-factor and even four-factor interactions for some simulation models.
- Analysis of variance, which is typically used to determine the statistical significance of factor effects, assumes constant variances and normally distributed error terms (differences between observed and predicted average responses), which are generally *not* valid assumptions for simulation models.
- Metamodels based on CCDs may provide poor predictions for simulation models with “complex” response surfaces. For some models Latin hypercube designs will give much better predictions.
- In physical experiments, there is a strong emphasis on designs that require a “small” number of design points. However, this need not be the case for many simulation models, because of computer speeds, multi-core processors, and cloud computing.

Note that some simulation analysts might forego factor screening using a two-level design (e.g., 2^k) and move immediately to constructing a metamodel where each factor is simulated at three or more levels.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling." *Operations Research* 58: 371-382.

Barton, R. R. 2013. "Designing Simulation Experiments." In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 342-353. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kleijnen, J. P. C. 2015. *Design and Analysis of Simulation Experiments*, 2nd ed. New York: Springer.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. "A User's Guide to the Brave New World of Designing Simulation Experiments." *INFORMS Journal on Computing* 17: 263-289.

Law, A. M. 2015. *Simulation Modeling & Analysis.* 5th ed. New York: McGraw-Hill Education.

Montgomery, D. C. 2013. *Design and Analysis of Experiments*. 8th ed. New York: John Wiley.

Sanchez, S. M., T. W. Lucas, P. J. Sanchez, C. J. Nannini, and H. Wang. "Designs for Large-Scale Simulation Experiments, with Applications to Defense and Homeland Security." In *Design and Analysis of Experiments: Special Designs and Applications*, edited by K. Hinklemann, Volume 3, Chapter 12, 413-441. New York: John Wiley.

Sanchez, S.M. and H. Wan. 2015. "Work Smarter, Not Harder: A Tutorial on Designing and Conducting Simulation Experiments." In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 1795-1809. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

SAS Institute Inc. 2017. JMP Statistical Software, www.jmp.com.

Stat-Ease, Inc. 2017. Design-Expert Software, www.statease.com.

AUTHOR BIOGRAPHY

AVERILL M. LAW is President of Averill M. Law & Associates, a company specializing in simulation seminars, simulation consulting, and software. He has presented more than 575 simulation and statistics short courses in 20 countries, including onsite seminars for AT&T, Boeing, Caterpillar, Coca-Cola, Defence Research and Development Canada, Defence Science and Technology Group (Australia), GE, GM, IBM, Intel, Lockheed Martin, Los Alamos National Lab, NASA, NATO (Netherlands), Norwegian Defence Research Establishment, NSA, Sasol Technology (South Africa), 3M, UPS, U.S. Air Force, U.S. Army, and U.S. Navy. Dr. Law has been a simulation consultant to more than 50 organizations including Booz Allen & Hamilton, Conoco/Phillips, Defense Modeling and Simulation Office, Kimberly-Clark, M&M/Mars, Oak Ridge National Lab, U.S. Air Force, U.S. Army, U.S. Marine Corps, and U.S. Navy. He has written or coauthored numerous papers and books on simulation, operations research, statistics, manufacturing, and communications networks, including the book *Simulation Modeling and Analysis* that has more than 165,000 copies in print and 17,500 citations. He developed the ExpertFit® distribution-fitting software and also several videotapes on simulation modeling. He was awarded the INFORMS Simulation Society Lifetime Professional Achievement Award in 2009. Dr. Law wrote a regular column on simulation for *Industrial Engineering* magazine. He has been a tenured faculty member at the University of Wisconsin-Madison and the University of Arizona. He has a Ph.D. in industrial engineering and operations research from the University of California at Berkeley. His e-mail address is <averill@simulation.ws> and his website is <www.averill-law.com>.