Efficient Modeling & Simulation of Biological Warfare Using Innovative Design of Experiments Methods

Thomas A. Donnelly, Erin E. Shelly and Daniel P. Cinotti

Efficient Modeling & Simulation of Biological Warfare Using Innovative Design of Experiments Methods (U)

Thomas A. Donnelly, Ph.D.
US Army Edgewood CB Center
5183 Blackhawk Road
ATTN: AMSRD-ECB-RT-IM
Aberdeen Proving Ground, MD 21010-5424
410-436-2571
FAX 410- 436-2165
thomas.a.donnelly@us.army.mil

Erin E. Shelly
US Army Edgewood CB Center
5183 Blackhawk Road
ATTN: AMSRD-ECB-RT-IM
Aberdeen Proving Ground, MD 21010-5424
410-436-1937
FAX 410- 436-2165
erin.shelly@us.army.mil

Daniel P. Cinotti
Science Applications International Corporation
4875 Eisenhower Avenue, Suite 210
Alexandria, VA 22304
703-212-2421
FAX 703-683-6249
cinottid@saic.com

Originally presented at:
75th MORS Symposium
Working Group 25, Test & Evaluation
16 November 2007

ABSTRACT

Innovative Design of Experiments (DOE) methods are used to significantly reduce the number of simulations required to model biological warfare (BW) attacks. The methods illustrated are applicable to almost any modeling and simulation (M&S) study involving large numbers of variables with many levels. The goal is to create a fast (seconds) surrogate metamodel of the simulation with which to make suitably accurate predictions of the response output for time consuming (hours) simulation trials not yet run. An example is shown that employs DOE methods that can be used when control variables are categorical and/or continuous. The analysis is based on a re-evaluation of data for all 648 possible combinations of settings of six variables in a BW attack case matrix completed for the military. Analyzing the data showed that as few as 5.6% of all possible trials are required to yield a metamodel that in 324 checkpoints not used

in fitting the model has a worst case prediction of Probability of Casualty (PCAS) that is off by 2.5%. Efficient modeling of simulation experiments opens the door to the inclusion of even more variables to broaden the applicability of the results.

INTRODUCTION

Increasingly, modeling and simulation (M&S) is being used to explore the broad variable design spaces for technologies and processes. In many stages of the Department of Defense (DOD) acquisition cycle computer experiments are used because they are less expensive than actual testing. In the area of biological warfare agents (BWA) the focus is on detection and protection. Testing new technologies with live agents in open conditions is restricted by treaty. As a result M&S has become an increasingly necessary substitute for live BWA testing.

Although M&S offers real savings over actual testing, it can still be a time consuming tool to employ and therefore still moderately expensive to use. One reason for this is that many simulations run in "real time" and so if the simulated process such as a BW attack takes many hours, then so does the simulation. In other cases such as with high-fidelity Lagrangian-particle-based simulations of transport and dispersion of BWA aerosols, the computer code is so computationally intensive that the available computing power may be the reason the trial takes hours to run.

When the intent is to study the effects of many variables on a process, the numbers of possible combinations of variable settings can easily be hundreds, thousands, or even millions. Faced with the prospect of many thousands of hours of simulation run time, too frequently the approach is to scale back the problem space by holding some of the variables constant. Although this reduces the workload it does so at the expense of the breadth of applicability of the analysis and at the risk of not learning about key interactions among the variables.

Mee (2004) discusses using Design of Experiments (DOE) methods to efficiently study main effects and 2-factor interaction effects for 47 two-level factors in simulation experiments for ballistic missile defense systems. Base designs of 512 (resolution IV) and 4,096 (resolution V) simulations were run, with the 512-trial design having 352 simulations added to it to resolve confounding of two-factor interactions. Xu (2007, unpublished paper) discusses algorithmic construction of even more efficient fractional factorial designs with large run sizes including a 2,048-trial (resolution V) design.

DOE methods can be employed to maximize the scope of the problem that can be studied for a given set of resources. The problem for which we demonstrate an example solution here is how to efficiently obtain fast surrogate metamodels (i.e. models of our time consuming simulation models). The metamodels give us an interactive "what if?" tool for querying the problem space and identifying the more interesting conditions to be examined by running the full simulation.

Running the fewest experimental trials to get the most information is why modern DOE methods originated with Fisher (1935) in long-running experiments (a growing season) at the Rothamsted Agricultural Experimental Station in the United Kingdom. Sequentially running blocks of experimental trials was found to be a way to get the main effects early and continue to add higher order model terms to build toward the final suitably accurately predictive model. The same approach generally works for computer experiments with some modification based on different assumptions about real experiments and computer experiments.

IMPORTANCE OF THE DOE SOLUTION

The DOE solutions shown are important because developing fast surrogate metamodels of long-running simulations can drastically cut the time required to understand the impact of the variables being studied. If after running a small fraction of the number of simulations for all possible variable combinations one can predict with 95% accuracy the results of the hundreds or thousands of simulations required to fill out a case matrix, then there is a good chance that whatever strategic, tactical or guidance decisions would be made after running the full set, could very well have been made after running the first small fraction of simulations. Even if the accuracy is only 80% after the first DOE has been run, follow on experimental designs – still small compared to the full case matrix – can be added to the original DOE to improve the accuracy of predictions.

Furthermore, with the increased efficiency of using DOE methods, simulationists and analysts will be less inclined to eliminate from study other variables in an effort to shrink the number of simulations. This will in all likelihood lead to the identification of more interactions among the larger variable set. The increased breadth of the problem space studied will lead to wider applicability of the results.

TRADITIONAL DOE METHODS

Many simulations produce non-stochastic (non-random or deterministic) results - i.e. the exact same result is obtained every time the identical variable settings are run. In virtually all real experiments there is a certain amount of random error. The relative size of the effects (signals) of interest as compared to the random error (noise) drive how many experiments need to be run to achieve a certain level of statistical significance in the results (Wheeler, 1973). Looking for big signals in small noise is easy. Looking for small signals in big noise is hard – i.e. you need to replicate the experiments many times to build up good averages for each unique trial in order to detect the small effects with any statistical confidence. When the simulation models are non-stochastic there is no need to replicate trials.

For most real experiments that are run over suitably narrow ranges of the control variables it is assumed that a single physical mechanism can be well modeled by a simple polynomial model. When all variables are continuous the most frequently assumed response-surface model is the quadratic which besides a constant term has terms that

support (i) linear or main effects, (ii) 2-way interaction effects among the variables, and (iii) curvature effects (via the squared terms).

An example of a continuous (or quantitative) variable would be the mass of a particular BWA. An example of a categorical (or qualitative) variable would be the type of agent. The order of the levels for the type of agent makes no difference. The order of the levels of the mass makes a difference and it makes sense to interpolate predictions for values of mass between the levels for which experiments are run.

The most frequently used traditional experimental designs for obtaining the main effects are called fractional factorial designs. Full factorial designs are able to support analysis for interactions (2-way and higher) among the variables. In the case of all continuous variables, fractional and full factorial designs generally have two levels at the extremes of the ranges of interest. Adding a sufficient number of mid-point levels for the variables enables the support of a quadratic response-surface model.

It is shown in Figure 1 how these three types of designs can be built in a sequential fashion for the case of three control variables x_1, x_2, and x_3. Beneath each block is the increasingly complex polynomial model that can be supported by the cumulative total of trials. From left to right the polynomials shown are the *linear*, *interaction* and *quadratic* models. The cumulative total of trials after each stage yields (i) fractional factorial, (ii) full factorial and (iii) central-composite-in-a-cube designs.
The reader is referred to textbooks on traditional experimental design techniques by Box et. al. (2005), Montgomery (2005) and Wu and Hamada (2000).

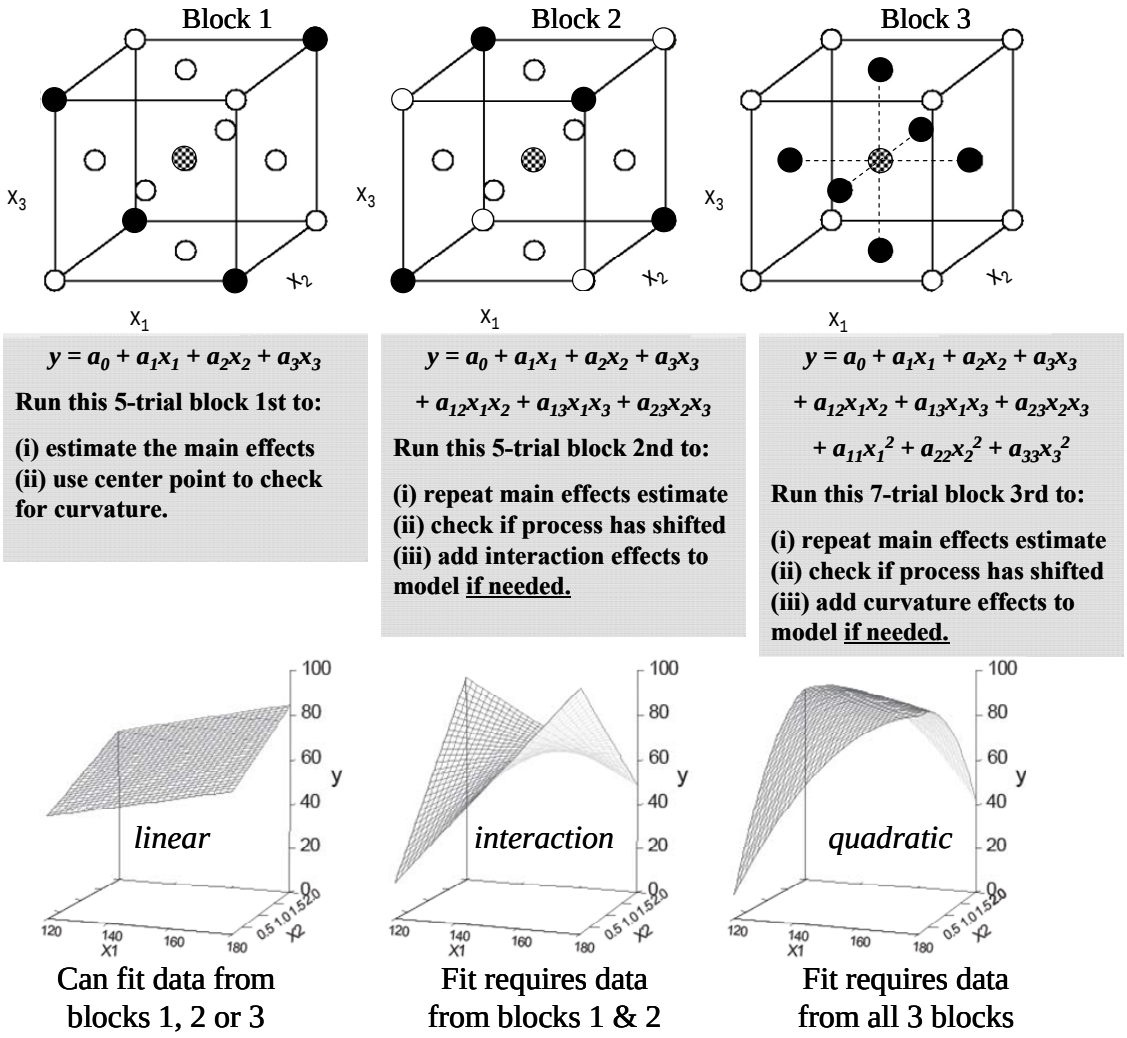

Figure 1: Three blocks of trials (combinations of settings at which to experiment) are shown that when run sequentially support increasingly complex polynomial models. Block 1 is a half-fractional factorial design with a center point added. Block 2 is another half-fractional factorial design (complementary to Block 1) with a center point. The combination of the trials in Blocks 1 and 2 make up a full factorial design in the three variables (plus a center point). Block 3 contains six "star" points at the centers of the six faces of the cube and a center point that when added to the trials in Blocks 1 and 2 make up a central-composite-in-a-cube response-surface design. Response surface plots of y versus x_1 and x_2 (with x_3 held at a constant value) are shown left to right for fits of the same set of data with the *linear*, *interaction* and *quadratic* models and shows their increasing flexibility.

This same approach of sequentially creating blocks of trials to support increasingly complex models can be applied when all variables are categorical. However, the graphical display of the increase in complexity is not as easy to show when all variables are categorical because interpolation of model predictions between the specific combinations of settings has no meaning. Predictions in this case are more generally presented as values in a table where each cell is associated with a unique

combination of variable settings and the order of the levels for a variable has no practical meaning.

COMPUTER AIDED DESIGN OF EXPERIMENTS

The military example will use a sequential approach similar to that illustrated in Figure 1. In this example the six control variables in the simulated process will primarily be analyzed as all categorical. It will however be shown that the identical predictions can be obtained using a model that treats three variables as categorical and three as continuous.

The sequence of innovative designs in this example will use computer algorithms to create designs that cannot be found in textbooks. Although used here with computer experimentation as the source of the response values, these particular algebraic and algorithmic design tools are equally practical with real experimentation.

Future work will treat a case where all 10 variables are treated as continuous but will not use a traditional response-surface experimental design nor will it use a polynomial model to analyze the data. These changes are the result of significant differences in the assumptions about computer experiments as compared to real experiments. This future case is discussed here because it introduces some very different assumptions from those applied to real experiments.

The first big difference which has already been discussed is that many simulations are non-stochastic yielding the same result each time they are run. The second big difference is that instead of assuming that a single mechanism is acting in the process over a narrow range of the variables, it is assumed that the model consists of a series of smaller physical models that each feed their output into the next as input. This multi-mechanism simulation model will not likely be approximated by a simple polynomial model and may very well have multiple local maxima and minima within its multi-dimensional design space.

The innovative designs to be used in this future case are frequently called “space-filling” designs. They are so named because rather than emphasizing trials at the extreme combinations of the variables, these designs instead have their points spread throughout the interior of the space. Figure 2 shows a space-filling design for 3 variables next to the traditional response surface design (central-composite-in-a-cube) for 3 variables. This latter design is the same as the combination of all three blocks of trials shown in Figure 1.

The seminal paper in the field of design, analysis and modeling of simulation and computer experiments is that of Sacks et. al. (1989). Books in this field that address the issues of space-filling designs and kriging analysis with all continuous variables are those of Notz et. al. (2003), Fang et. al. (2005) and Kleijnen (2008). A good review article is that by Kleijnen et. al. (2005).

Space-filling designs are available from many sources including the commercially available software package JMP®. Some of the newer orthogonal and Nearly Orthogonal Latin Hypercube (NOLH) designs are available for download from the Naval Postgraduate School website http://harvest.nps.edu/. Recently developed designs were made available to the authors in a preprint of "Orthogonal-Maximin Latin Hypercube Designs" by Joseph and Hung (2007, an unpublished paper to appear in *Statistica Sinica*) of the School of Industrial and Systems Engineering, Georgia Institute of Technology.

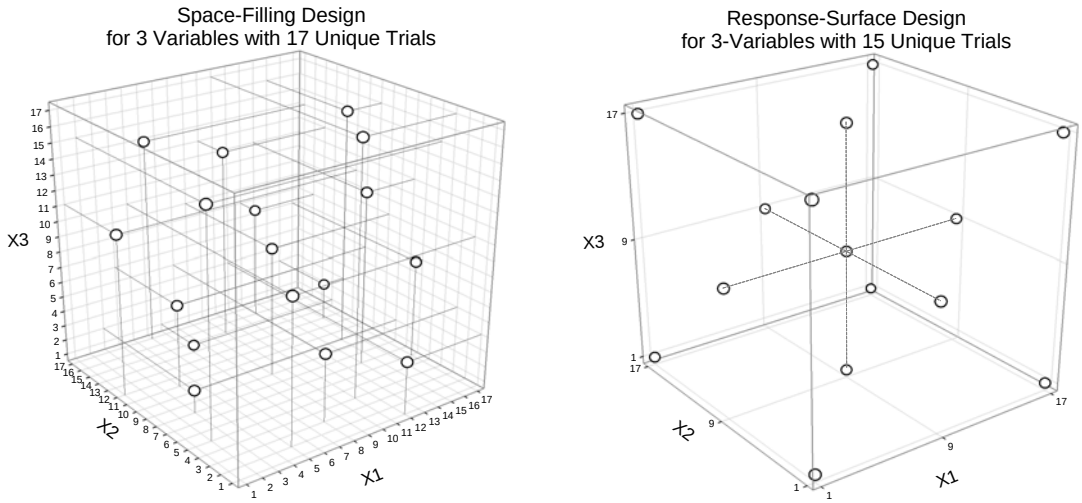

Figure 2: Comparison of space-filling design (left) with 17 unique trials spread throughout the design volume to a traditional response-surface design (right) with 15 unique trials with its trials located predominantly at the extremes of the volume. The response-surface design is a central-composite-in-a-cube design and its 15 unique trials were previously broken out into 3 blocks of trials in Figure 1.

Sacks et.al. (1989) were the first to propose using kriging analysis to analyze the response data generated by computer experiments using space-filling designs. Kriging is a technique originally conceived for geospatial analysis of gold deposits by Daniel Krige, a South African mining engineer. Matheron (1963) developed the theory behind interpolation by kriging. Because kriging is a spatially weighted regression technique it will fit all the data exactly. Were the data stochastic it would fit this data just as perfectly which is why this technique is currently used only with non-stochastic data. Emerging modifications to kriging analysis methods called "Blind Kriging" by Joseph et. al. (2007, unpublished paper to appear in the *ASME Journal of Mechanical Design*) suggests that these new methods may be applicable to stochastic data. The authors plan to explore these new methods in future work.

DOE SEQUENCE USED WITH MILITARY SIMULATION DATA

This example re-uses data collected by one of the authors in a previous simulation study that was part of the military effort. This effort, which studied a range of BW attacks that might be encountered at a USAF Air Base, was sponsored by what is now the US Air Force Deputy Director of Strategic Plans and Policy (HQ

USAF/A5XP). The data set re-used for this report was originally presented in the report "Developing Masking Guidance with Respect to BW Trigger Events" by Cinotti (2007).

The subset of data used here is from just one of five case matrices studying different types of BW attacks on a military base. The type of attack in this data set is that using Tactical Ballistic Missiles (TBMs). The response on which the analysis is focused is the Probability of Casualty (PCAS) averaged over 14 point detectors at which BWA concentration data were simulated. A map of detector locations and notional concentration contours of the BWA cloud is shown in Figure 3.

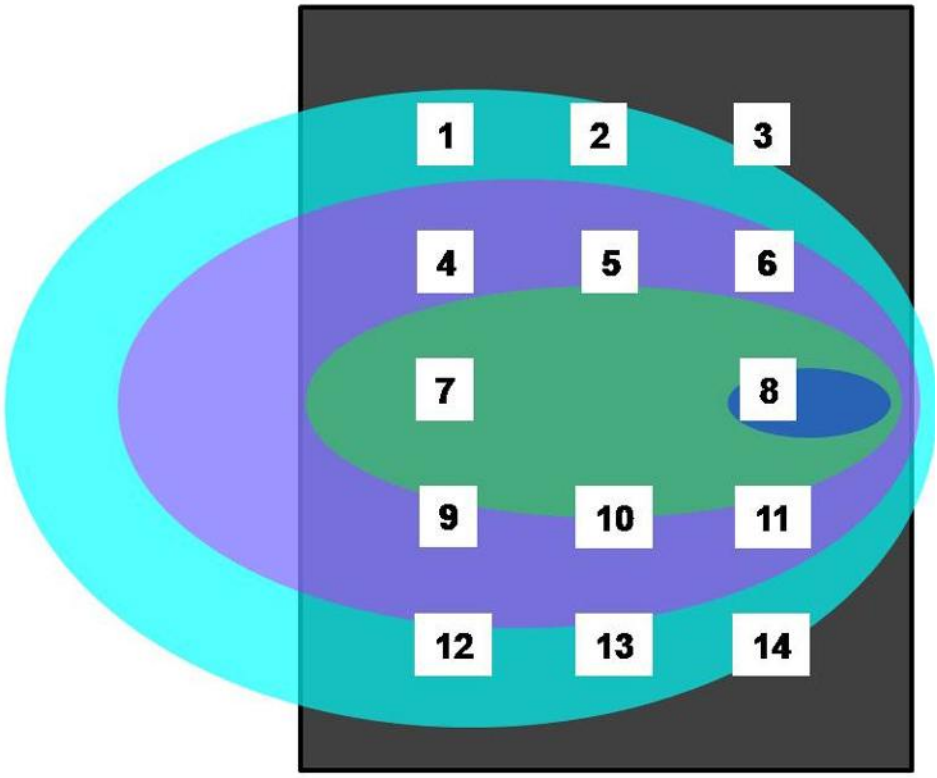

Figure 3: Notional Tactical Ballistic Missile attack with Biological Warfare Agent (BWA) on a military base is depicted. At this time the aerosol cloud has identified agent at 2 or more detectors resulting in a detector alarm. Some personnel may have already been exposed to an effective dose. Others would benefit from donning their mask.

The goal in this work is to see how well the PCAS response calculated by the long-running simulation model could be predicted with a metamodel fit to subsets of the full data matrix. Half of the 648 simulations were never used to fit the metamodels. The PCAS values for these 324 simulations were only used as checkpoints against which metamodel predictions were compared.

The full case matrix consisted of all 648 combinations of settings of 6 variables. Table 1 shows the variables and associated settings. Although for most of the analyses all the variables are treated as categorical, three of the variables are continuous in nature.

Case Matrix for Simulation of BWA Attack on Military Base

Variable Names	Number of Levels	Level Names	Variable Type
Agent Codes	6	A, N, T, H, R & Y	categorical
Season	3	Winter, Summer & Spring/Fall	categorical
# of TBMs & Spread Radius	2	1 TBM & 1 m & 2 TBMs & 1000 m	categorical
Time of Attack	3	0500, 1200 & 2200 Local Time	continuous
Mass (nested in Agent)	3	1.00, 1.57 & 2.00 (coded levels)	continuous
Height of Release	2	0 & 10 m	continuous
Total Cases	**648**	(648 = 6 X 3 X 2 X 3 X 3 X 2)	

Table 1: The case matrix for 6 variables showing the number of levels, level names and variable types from one of the simulation studies run as part of the military effort. The Mass variable actually used different amounts of material depending upon the agent. The numbers shown are the relative amounts of agent used for every agent regardless of actual mass levels. As a result the variable Mass is treated as a nested variable because of its dependence on the choice of agent.

Figure 4 plots all 648 possible combinations of variable settings on 24 cubes, with each cube showing all 3 X 3 X 3 = 27 combinations of the variables, Mass, Season, and Time of Attack ("Time of Attack" in figures and plots is shortened to "Hour.") There are six columns of cubes with one column for each agent by four rows of cubes, with each row associated with one of the combinations of the two levels of Height of Release (abbreviated as HOR) at each of the two levels of # of TBMs (Tactical Ballistic Missiles) & Spread Radius.

All 648 Possible Combinations of Settings for the 6 Variables in the Case Matrix in Table 1.
(The Solid Circles Indicate the Variable Settings for a 36-trial Orthogonal Array.)

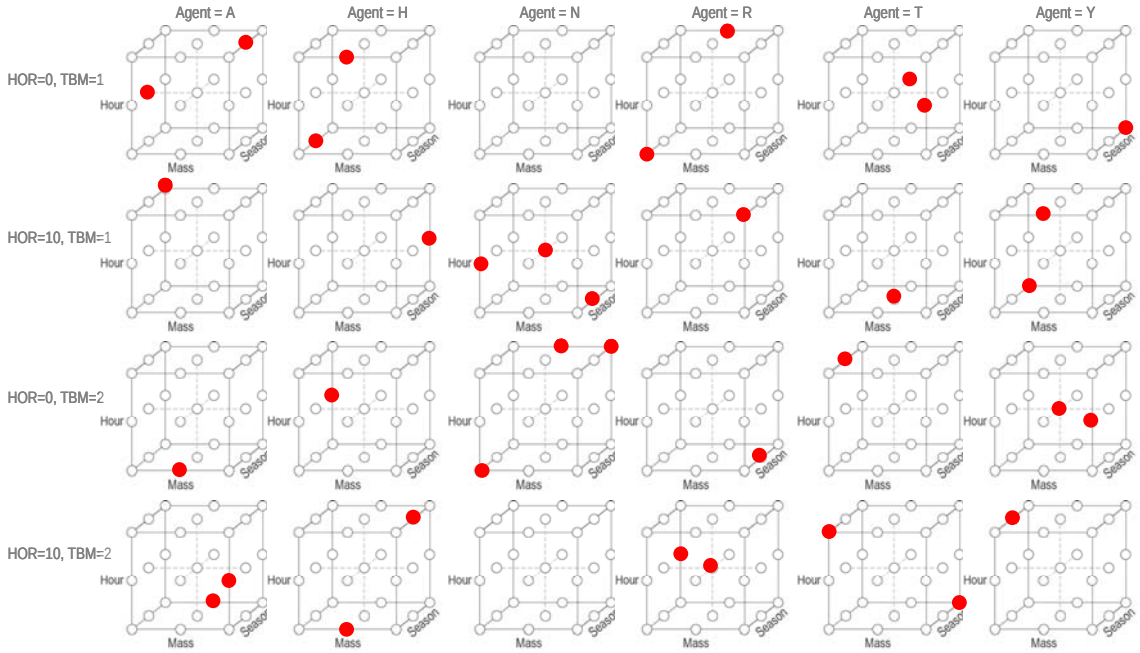

Figure 4: All 648 possible combinations of variable settings are spread over 24 cubes with each cube showing all 3 X 3 X 3 = 27 combinations of the variables, Mass, Season, and Hour (Time of Attack). There are six columns of cubes with one column for each agent by four rows of cubes with each row associated with one of the combinations of the two levels of Height of Release (HOR) at each of the two levels of # of TBMs & Spread Radius. The 36 locations marked by solid dots make up an orthogonal array design suitable for estimating the main effects of all 6 variables. These 36 locations make up only 1/18th (5.6%) of all 648 possible combinations.

Even though 2 of the 24 cubes in Figure 4 have no design trial locations indicated in them by the presence of solid dots - specifically the top and bottom cubes in the Agent N column - there is still be seen a balance in the numbers of levels across the variables. One can inspect Figure 4 to see that for any agent the six dots are spread among the four cubes such that:

2 each always fall on the top, middle and bottom slices of the Mass-Season plane,
2 each always fall on the front, middle and rear slices of the Hour-Mass plane, &
2 each always fall on the left, middle and right slices of the Hour-Season plane.

STAGE-1 DESIGN: 36-TRIAL ORTHOGONAL ARRAY

The 36-trial orthogonal array (OA36) shown in Figure 4 is similar to those found in Wu & Hamada (2000) and at a web site of orthogonal arrays maintained by N.J.A. Sloane, http://www.research.att.com/~njas/oadir/index.html. A pair of columns in a matrix is called orthogonal if all possible combinations of levels in the two columns appear equally often. Orthogonal arrays come in different "strengths" which are comparable to "resolutions" for fractional factorial designs. This is a strength 2 (or resolution III) design because the main effects are free of confounding (or aliasing) with each other, while some 2-way interaction effects are confounded with the main effects.

The OA36 shown in Figure 4 was created with an algebraic design generator using computer code obtained from Professor Hongquan Xu, of the UCLA Department of Statistics. Source code for this orthogonal array design tool is available here: http://www.stat.ucla.edu/~hqxu/nsf/. Four slightly different OA36 designs that each contains an equivalent number of variables and levels as to what has been shown in Figure 4 were published by Zhang, et. al. (2001).

The advantage of using computer code is that it can find orthogonal arrays for combinations of variable levels that meet specific real-world needs of the experimenter. Too often when all that is available is a catalog of designs, experimenters reduce their problem to fit the available designs rather than create a design for a combination of settings yet to be cataloged. Still using these codes do not guarantee that an orthogonal array can be found for all numbers of variables and levels.

As an example of the efficiency that is possible using this design tool in a case involving a larger numbers of variables, the authors for a separate project created a 144-trial orthogonal array (OA144). The numbers of levels were distributed such that three variables had 6 levels, two variables had 4 levels, two variables had 3 levels and two variables had 2 levels. The 144 trials in the design amount to being only 0.12% of the 124,416 = 6 X 6 X 6 X 4 X 4 X 3 X 3 X 2 X 2 possible combinations of settings of the levels of these 9 variables. This OA144 design is just 4 times as large as the OA36 in the case 1 example, but the total number of possible combinations, 124,416, is 192 times as many as the 648 shown in Figure 4.

Once the OA36 trials were determined, the full set was sorted in order to split the 648 trials into two halves of 324 trials with the OA36 trials all in one half. The other half not containing any of the OA36 trials was set aside to be used as checkpoints.

STATISTICAL DETAILS

Before discussing how the analysis proceeded in three stages, it is worth noting a couple of statistical details in the analysis. The first has already been mentioned and is the matter of nesting the Mass variable in the Agent variable. The second involves transforming the response data values, PCAS, to a new scale.

Because a different set of mass values were used for each agent, the variable Mass is “nested” within the variable Agent. When the three coded levels 1, 1.5714 and 2 are used, it is known that these are the relative masses and not the actual mass values run in the simulation. This method of coding the nested variable is also called “sliding levels” in Wu and Hamada (2000). To properly treat this issue Mass as a standalone variable is removed from the model and is instead entered in the model only through 12 interaction terms (6 levels of Agent X 2 levels of Mass) with the variable Agent. Thus the “nested 1-way model” has 24 terms, which are 10 more than a straight 1-way main-effects model would have for the six variables with their given numbers of levels.

The response Probability of Casualty (PCAS), which is bounded within the range (0, 1), was transformed to a new scale using $2*\text{Arcsin}((\text{PCAS})^{1/2})$ which maps the range (0, 1) to the range $(-\infty, +\infty)$. This transformation made the error fit the usual regression assumption of the data being normally distributed. Using this transformation had the added benefit of preventing the regression model from predicting PCAS values and prediction limits that were above 1.0 and physically impossible.

PREDICTIONS & CHECKPOINTS

Although the regression analysis was done using the transformed values of PCAS, prediction plots (and 95% prediction limits) are presented after transforming the values back into raw units of PCAS. As an example, Figure 5 shows plots of predicted PCAS versus Mass – with Mass treated here as a continuous variable – for five models fit to three different sized data sets. The choices of models fit to which particular sets of trials are discussed in following sections.

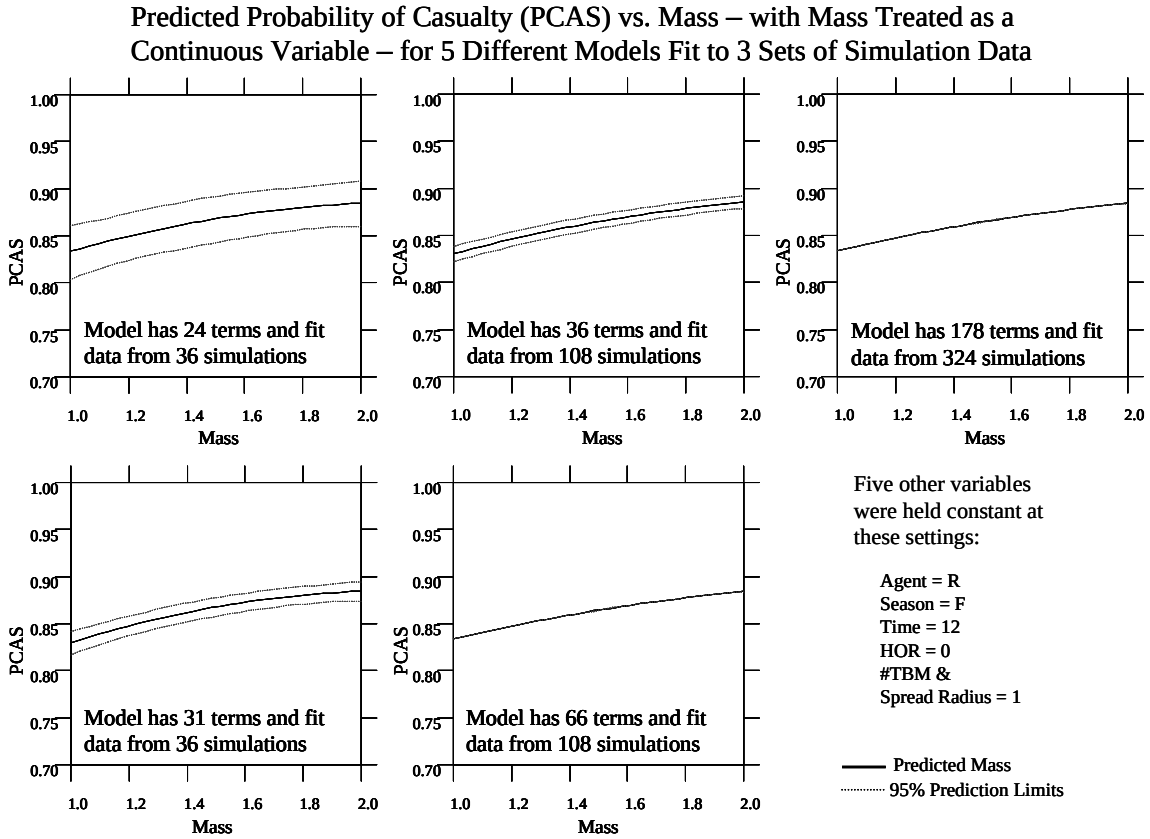

Figure 5: Plots of predicted Probability of Casualty (PCAS) versus Mass are shown for five different models fit to 36, 108 or 324 observed PCAS values out of the 648 PCAS values collected from the running of all possible variable combinations shown in Figure 4. The two plots on the left show results of fitting 36 trials (5.6% of 648). The two plots in the center show results of fitting 108 trials (16.7% of 648). The single plot at the right shows the result of fitting 324 trials (50% of 648). The prediction curve is very similar in all plots. The 95% prediction limits – the window into which checkpoint trials should fall – are seen to shrink as one views the plots from left-to-right and from top-to-bottom. The limits are sufficiently tight in the top-right and bottom-center plots that the limits are coincident with the prediction curve.

The point of this plot is to raise the question, "How good does the fit have to be before one can stop taking data and start answering the questions for which the simulation study was initiated?" Another question that the graph raises is, "How well do these predictions hold up to validation with checkpoint trials?"

Before getting into the regression error numerical estimates the answer to this question can be seen graphically in Figures 6 and 7. Histograms of the "Percent Off Target" that the PCAS predictions fell relative to 324 checkpoint observations are plotted for the corresponding models used to draw the PCAS versus Mass plots in Figure 5.

Histograms of the "Percent Off Target" that Probability of Casualty (PCAS) Predictions Fell Relative to 324 Checkpoint Observations

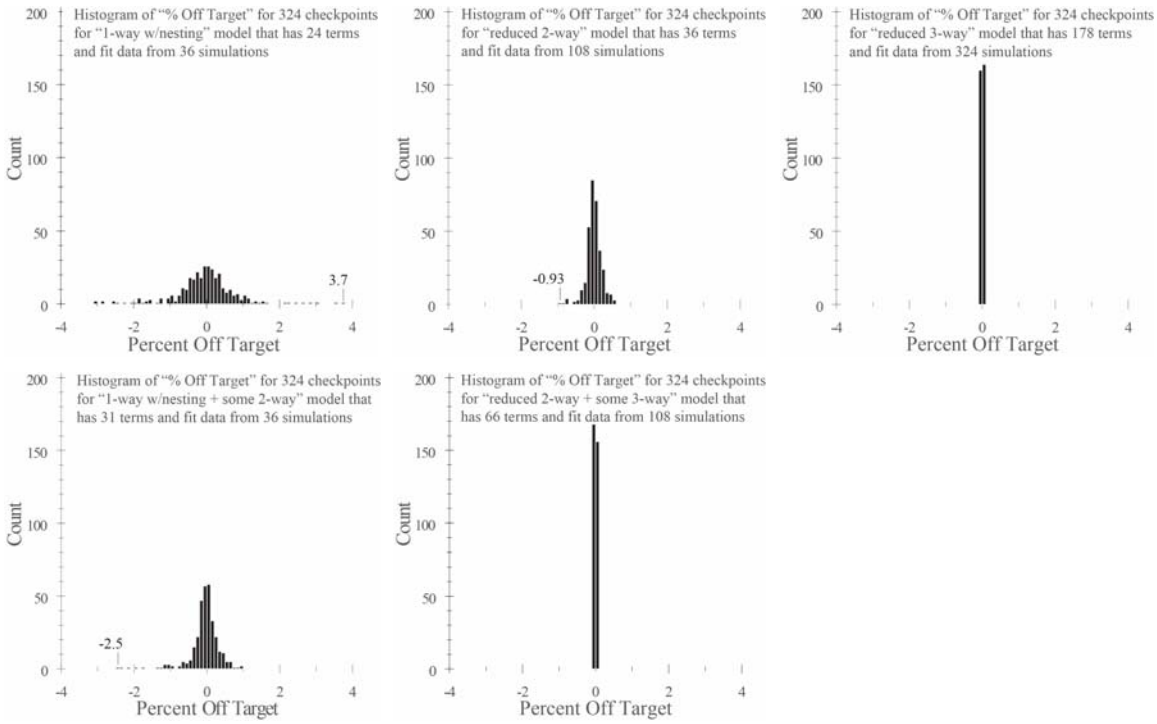

Figure 6: Predictions of Probability of Casualty (PCAS) were made for five models (description of model is given in each plot) at the settings of the six control variables for the 324 simulations not used in fitting any of the models. The percentage above or below the simulation observation that the predicted value fell was calculated and called the "Percent Off Target." Histograms showing the counts in the eighty 0.1% wide bins between -4% and 4% are shown for each model. Locations and magnitudes of the worst case predictions are shown in the two left most and the top center plots. The two plots not showing worst case prediction values are redrawn in Figure 7 with increased resolution of the horizontal axis.

In Figure 6 it can be seen that for 36 simulation trials the better predicting model in the lower-left plot has a worst case prediction that is off target by 2.5%. For the data used to make this histogram the median of the 324 absolute values is 0.16%. This plot looks almost as good as the top center plot whose model was fit to 108 simulation PCAS data values – 3 times as much data. Here the worst case prediction is more than 2.5 times

smaller, off target low by 0.93%. The median of the 324 absolute values in this plot is reduced 31% from 0.16% to 0.11%.

Higher Resolution Histograms of the "Percent Off Target" that Probability of Casualty (PCAS) Predictions Fell Relative to 324 Checkpoint Observations

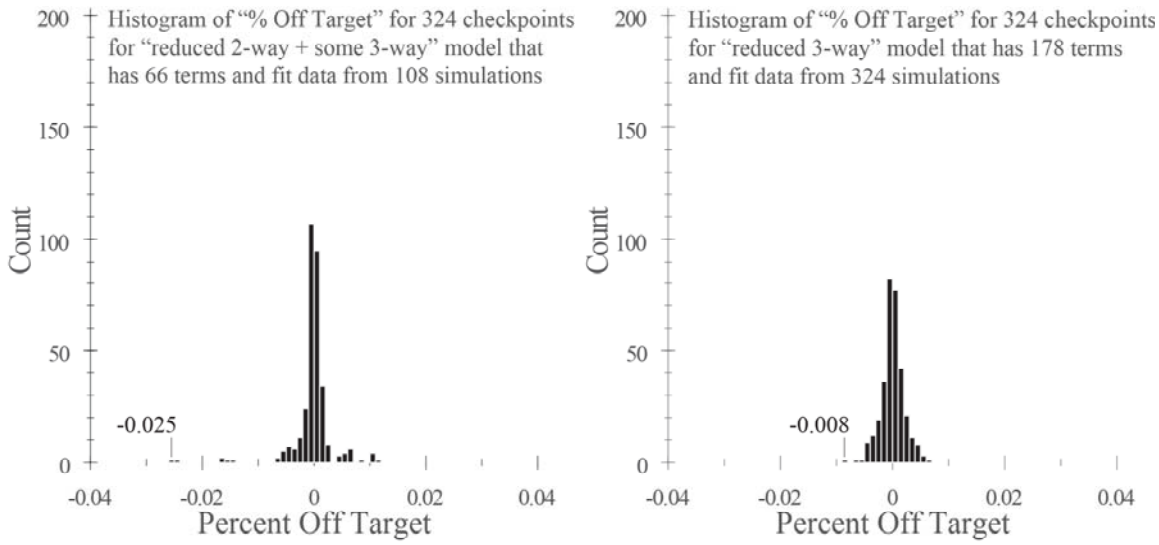

Figure 7: Two plots from Figure 6 are redrawn with a horizontal scale that is one-hundredth as wide as that of Figure 6 for the response "Percent Off Target." Histograms showing the counts in the eighty 0.001% wide bins between -0.04% and 0.04% are shown for both models. The worst case predictions are identified for each model.

In Figure 7 it can be seen that for 108 simulation trials (left plot) the model has a worst case prediction that is off target by -0.025%. For the data used to make this histogram the median of the 324 absolute values is 0.0010%. This plot looks almost as good as the right plot whose model was fit to 324 simulation PCAS data values – again, 3 times as much data. Here the worst case prediction is more than 3 times smaller, off target low by 0.008%. The median of the 324 absolute values in this plot is reduced 30% from 0.0010% to 0.0007%.

Looking at Figure 7 it is fair a question to ask at this point, "Has the analysis gone beyond the point of practical significance?" "Does it make sense to continue to run hundreds or even thousands of hours of additional simulation time just to improve the accuracy of prediction of a response by a few hundredths of a percent?" "When do we have enough practical information to solve the problem for which the simulation experiments were designed?"

With these rhetorical questions in mind and Figures 5, 6 and 7 graphically providing some of the answers, the balance of this example will discuss the reasoning and methods behind choosing both the design trials of the latter two stages, and the particular models used to generate the plots shown so far. It will be seen that sometimes models are reduced in complexity and things get better and sometimes they are increased in

complexity and things get better. Much of the discussion will involve comparing three different error estimates, two are calculated from the data that are fit with the model, and one is calculated from the checkpoints. These error estimates help guide us in choosing how to modify the model complexity to improve its prediction accuracy. These error estimates will quantify in more detail what Figures 5, 6 and 7 have already shown.

STAGE-2 DESIGN: 72 TRIALS ALGORITHMICALLY ADDED TO INITIAL 36

It has already been stated that the OA36 design making up the first stage of trials has its main effects partially confounded with the 2-way interactions. In order to enhance the model to fully support analysis of all the 2-way interactions a set of 72 new trials were chosen from 288 candidate trials. These candidates were comprised of the 324 non-checkpoint trials minus the 36 OA trials already run. An algorithmic design software package, ECHIP®, was used to generate the D-optimal design. This algorithm picks the best subset of trials to minimize the predicted variance for the proposed model. For a review of optimal design theory and "alphabetic" optimality criteria, the reader is referred to Atkinson and Donev (1992). Although ECHIP® is no longer commercially available, other software such as JMP® (www.jmp.com) and Design-Expert® (www.statease.com) can accomplish the same task.

D-optimal design algorithms do not do a particularly good job of selecting trials to support main effects models because these designs frequently have the main effects confounded, i.e. the design columns are not orthogonal. For this reason the pairing of the orthogonal array code to pick the first 36 with the D-optimal design code to pick the next 72 is particularly effective.

The minimum number of trials that could have been added algorithmically to the OA36 design to support the full 2-way interaction model would have brought the total number of trials equal to the number of terms in the model, 79. This means that as few as 43 = 79 – 36 could have been algorithmically added. The resulting number of trials would have allowed no degrees of freedom for estimating model error. The design would also have had a low G-efficiency (6.9%), a measure of how well the design is filling out the space to support the proposed model. Adding 72 trials to bring the total to 108 brought the G-efficiency up to 39.5%. For the numbers of variables and levels involved, G-efficiencies in excess of 25% are considered acceptable and G-efficiencies in excess of 50% are considered very good (Wheeler, 2002). The decision to add 72 was influenced by a desire to have a good G-efficiency and it also divided the data sets into groups that at each stage in the process tripled the number of trials analyzed in the previous stage.

STAGE-3 DESIGN: ADD THE BALANCE OF THE NON-CHECKPOINT TRIALS

The stage 3 design amounted to using all 324 non-checkpoint trials. This means that the 216 trials not chosen in stage 2 were added – but not algorithmically chosen – to the 36 stage 1 trials and the 72 trials added at stage 2 (108 total), and ultimately were used to fit the 3-way model. Checking how well these 324 trials support the "1-way

w/nesting," "full 2-way" and "full 3-way" models, the respective G-efficiencies are 52.3%, 51.5% and 7.5%. This low result for the 3-way model would not be encountered if the candidate set from which 324 trials could be chosen was the full 648 possible combinations. In that case the G-efficiency would be a respectable 56.7%. Since it was desired to use the exact same checkpoints to test all models, no attempt was made to swap out checkpoints to improve the design. In hindsight this could easily have been done and will be if future design sequences involve at least three stages.

IMPROVING MODEL FIT BY DELETING UNNECESSARY TERMS

Although full models with all their possible N-way interactions were first fit, in the cases of the stage 2 and stage 3 analyses, reduced models were sought to prevent "over-fitting" the data and to improve the accuracy of the predictions. The general approach to finding a reduced model was to first pick a design to support a proposed full model and look at the regression output to see if there were terms not contributing to explaining the variability in the process. Using ECHIP™ software that employs criteria proposed by Sawa and Hiromatsu (1973), model terms were identified that when eliminated would reduce the model complexity and correspondingly shrink the Residual Standard Deviation (model error calculated from the data used in the fit), the Checkpoint Root Mean Square (RMS) (checkpoint error calculated from the data not used in the fit) and especially the Cross-Validation RMS values (calculated from the data used in the fit).

Following Atkinson (1985:p22), the calculation of the Cross-Validation RMS is equivalent to successively deleting each observation individually, fitting the balance of the data, calculating the deletion (or "one-left-out") residual between the observed value and the prediction from the fit of the balance of the data, and then calculating the root-mean square (RMS) of all the deletion residuals. The Cross-Validation RMS is sensitive to over-fitting data and its value will be inflated if the removal of individual data points dramatically changes the fit. Thus, a relatively smaller Residual SD as compared to the Cross-Validation RMS, especially when there are many insignificant terms, suggests that elimination of unnecessary model complexity will improve the robustness of the model and makes it a better predictor. This is what was observed in the analysis of the 2-way and 3-way models and is shown in Table 2.

Error Estimates for 8 Models Fit to Data Sets of 36, 108, & 324 Observations

Number of Trials	**Model Used to Fit Data**	**Number of Model Terms**	**Residual SD** (model error from data used in fit)	**Cross-Validation RMS** ("one-left out" error from data used in fit)	**Checkpoint RMS** (model error from 324 data values NOT used in fit)	**Adjusted R-squared**
36	1-way	14	0.043623	0.055802	0.037217	0.977
36	1-way w/nesting	24	0.026557	0.047269	0.035424	0.992
36	1-way w/nesting + some 2-way	31	0.008212	0.025188	0.016153	0.999
108	2-way	79	0.011197	0.022207	0.010772	0.998
108	reduced 2-way	36	0.008469	0.010933	0.008612	0.999
108	reduced 2-way + some 3-way	66	0.000045	0.000132	0.000179	1.000
324	3-way	242	0.000039	0.000078	0.000083	1.000
324	reduced 3-way	178	0.000037	0.000058	0.000064	1.000

Table 2: The Residual Standard Deviation (SD), Cross-Validation Root Mean Square (RMS), and Adjusted R-squared values are shown for 8 different models fit to three increasingly larger sets of Probability of Casualty (PCAS) simulation data. The first three columns show the number of unique trials for which data were fit, a description of the model, and the number of terms in the model. The Checkpoint RMS is also shown and is calculated from the differences between the model prediction and observed simulation data at 324 unique checkpoint trials not used in fitting the models. The three columns of error estimates generally get smaller and the Adjusted R-squared estimate gets larger as one reads down the rows of the table as more and more of the variability in the process is explained by succeeding models.

ANALYSIS OF 36 STAGE-1 TRIALS

The first two rows of Table 2 show results for the stage 1 design, the 36-trial orthogonal array, after fitting a 14-term 1-way model and a 1-way model plus 10 additional terms to properly support the nesting of the Mass variable in the Agent variable. Because the Mass was known to be different depending on the Agent, adding these terms was strongly expected to reduce the error in the model. Row 2 is associated with the top-left plot in Figures 5 and 6.

IMPROVING MODEL FIT BY ADDING SELECTED TERMS

The third row for the stage 1 trials in Table 2, show results of adding 7 more 2-way terms to support all other interactions among the 3 variables having the largest main effects; Agent, Mass, and #TBMs & Spread Radius. This brings the total number of terms to 31. The favorable projection properties of the orthogonal array support analysis of 2-way interactions for subsets of the variables as long as the total number of terms does not exceed the number of trials. The more terms added beyond the original main effects, the greater the level of confounding among the effects. Even so, with the Checkpoint RMS for the 324 checkpoints reduced 54% from 0.035424 to 0.016153, there appears to be justification for adding these terms. Row 3 is associated with the bottom-left plots in Figures 5 and 6.

This analysis of the OA36 trials with the 31-term model exploits two principles as discussed by Wu & Hamada (2000) and Xu et. al. (2004). The first principle is "factor sparsity," as called by Box & Meyer (1986), and states that only a few factors are active in any factorial experimentation. The second principle is that significant interactions occur only among the active factors and was called strong "effect heredity" by Chipman (1996). The effect heredity principle as stated in Wu and Hamada (2000:p112) is, "In order for an interaction to be significant, at least one of its parents should be significant." In this analysis these principles certainly appear to hold true. It is still strongly recommended that the 72 stage-2 trials be added to the original 36 stage-1 trials in order to support analysis of all possible 2-factor interactions.

In practice when faced with a new process to characterize, one will not already have the luxury of a large set of checkpoints with which to test the predictions of the initial model. It makes sense that whatever checkpoints are to be taken that when they are combined with the trials from the original design the checkpoints serve the purpose of supporting analysis with a more complex model. In particular one will want to make sure that the trials added will reduce the confounding among the interaction effects, which is what the D-optimal design does when it augments the original design to support the full 2-way interaction model. The addition of sufficient trials to support the full 2-way interaction analysis will also afford the opportunity to again apply the principles of factor sparsity and strong effect heredity to examine the most likely possible subset of 3-way interactions before actually collecting trials to support the full 3-way model.

ANALYSIS OF 108 STAGE-2 TRIALS

The fourth and fifth rows in Table 2 show results after fitting a full 2-way and a reduced 2-way model to the 108 trials of the stage 2 design. It is no surprise that the biggest difference between these two rows is for the Cross-Validation RMS where the value for the reduced model, 0.010933, is reduced 51% relative to the value for the full model, 0.022207. When one considers that 43 terms were removed from a 79 term model, it is a welcome result to see the Residual SD reduced by 24% from 0.011197 to 0.008469, and the Checkpoint RMS reduced 20% from 0.010772 to 0.008612, when fitting the smaller and less "flexible" model. The fifth row is associated with the top-center plot in Figures 5 and 6.

Comparing the 36-term "reduced 2-way" model analysis of 108 trials to the 24-term "1-way w/nesting" model analysis of 36 trials, it can be seen that the value of the Residual SD for the "reduced 2-way" analysis, 0.008469, is reduced 68% from 0.026557. Similarly, the Checkpoint RMS for the "reduced 2-way" model, 0.008612, is reduced 76% from 0.035424 of the "1-way w/nesting" model. Plots associated with this comparison in Figures 5 and 6 are respectively at the top-center and top-left.

Comparing the 36-term "reduced 2-way" model analysis of 108 trials to the 31-term "1-way w/nesting + some 2-way" model analysis of 36 trials, it can be seen that the values of the Residual standard deviations for the two analyses differ by just 3%,

0.008469 and 0.008212. However, the Checkpoint RMS for the "reduced 2-way" model, 0.008612, is reduced 47% from 0.016153 of the "1-way w/nesting + some 2-way" model. Plots associated with this comparison in Figures 5 and 6 are respectively at the top-center and bottom-left. These two plots look fairly similar in both Figures 5 and 6.

In reducing the 2-way model, terms were only eliminated in groups such as "all the interaction terms between two variables" not just the subset of terms that had high p-values indicating they were not contributing to the explanation of the variability in the process. That is to say if there were five interaction terms between Agent and HOR (Height of Release) and only the particular term "Agent R*HOR" had a sufficiently high p-value indicating it had virtually no effect, the term would still be kept if the other Agent "X"*HOR terms had low p-values and were thus significant or nearly significant.

The sixth row for the stage 2 trials in Table 2 shows the results of again applying the principles of factor sparsity and effect heredity to try to enhance the model to support higher order interactions - even though the design was not constructed to specifically support these terms. These are the results of adding 27 terms to support all the 3-way interactions among the 4 variables having the largest main effects; Agent, Mass, #TBMs & Spread Radius and now Height of Release (HOR) as well as 3 two-way terms between HOR and Mass and HOR and #TBMs & Spread Radius that were not in the reduced 2-way model. This brings the total number of terms to 66.

D-optimal designs generally do not have high G-efficiencies for models for which they are not designed. It was previously stated that the 108 trials in stage 2 had a G-efficiency of 39.5 for the full 2-way model. For the reduced 2-way model it rises to 64.4%. However, when the G-efficiency is calculated for these 108 trials with the "reduced 2-way + some 3-way" model the G-efficiency drops to 8.2%. If trials were to be algorithmically augmented to the 108, the addition of just 5 more trials raises the G-efficiency to 60.0%.

Returning to the values in row six of Table 2 we see that across the board this analysis of the 108 trials in stage 2 with the "reduced 2-way + some 3-way" model has by far the smallest error estimates. All three error estimates for the "reduced 2-way + some 3-way" model dropped at least 98% relative to the "reduced 2-way" estimates in row 5. The Residual SD dropped from 0.008469 to 0.00045, the Cross-Validation RMS dropped from 0.010933 to 0.000132, and the Checkpoint RMS for the 324 checkpoints dropped from 0.008612 to 0.000179. There again appears to be a strong justification for adding these terms. Row 6 is associated with the bottom-right plots in Figures 5 and 6, and the left plot in Figure 7. These plots are easily identified as the ones that are associated with the model that gives predictions near to the best while requiring one-third as many trials as the analysis producing the best values in rows 7 and 8.

ANALYSIS OF 324 STAGE-3 TRIALS

The seventh and eighth rows in Table 2 show results after fitting a full 3-way and a reduced 3-way model to the 324 trials of the stage 3 design. The various error estimates

in both rows are running almost 3 orders of magnitude smaller than those seen in the row 2 for the "1-way w/nesting" model. Compared to the error estimates

By comparison to the difference seen between the full and reduced 2-way error estimates in rows 4 and 5, there is not as large a difference between the results in rows 7 and 8. The largest difference between these two rows is for the Cross-Validation RMS where the value for the reduced model, 0.000058, is reduced 26% relative to the value for the full model, 0.000078. The result of removing 64 terms from a 242 term model was to also reduce the Residual SD by 5%, from 0.000039 to 0.000037, and to reduce the Checkpoint RMS 23%, from 0.000083 to 0.000064. The eighth row is associated with the right-most plot in Figures 5, 6 and 7.

When the values in row eight are compared to those in row six the percentage improvements are generally comparable to those seen between row 3 and row 5. The Residual SD is reduced 18% from 0.000045 to 0.000037. The Cross-Validation RMS is reduced 56% from 0.000132 to 0.000058. The Checkpoint RMS is reduced 64% from 0.000179 to 0.000064. The same questions that came to mind when looking at Figure 7 should come to mind again. Phrased slightly differently, "Are the reductions in the various error estimates and the corresponding improvement in prediction worth the price - given that three times as many simulations were run?"

SUMMARY AND CONCLUSION

A sequence of three sets of simulation trials were run, consisting of one-eighteenth (5.6%), one-sixth (16.7%) and one-half (50%), or 36, 108 and 324 of all 648 possible combinations of the 6 variables in the case matrix in Table 1. The 324 trials not used in the analyses of the three stages were used as checkpoints. The first two stages used computer generated experimental designs not available in textbooks. The second stage built on the first to initially support a more complex 2-way interaction model. Statistical criteria were used to eliminate terms in whole variable groupings and cross-validation was performed to measure the improvement. Factor sparsity and effect heredity principles were employed to add complexity to the model. Reduction in the checkpoint error was used to validate the improvement of adding these new terms.

Histograms of the "Percent Off Target" that the predicted values of PCAS were from the 324 checkpoint values indicated how well various models performed. The biggest improvement within a given stage came when guided by the principles of factor sparsity and effect heredity, complexity was added in the form of an appropriately chosen subset of the N-way interactions from the next higher model. Using this method the stage 1 trials showed that running 5.6% of all combinations yielded a model that had a worst case prediction that was 2.5% off-target. Using this method the stage 2 trials showed that running 16.7% of all combinations yielded a model that had a worst case prediction of 0.025% off target.

All data sets may not model as efficiently as was seen with the military simulation data. However, the strategy demonstrated here is robust. Taking data

in sequential blocks of trials that build support for increasingly complex models means that if a sufficiently accurately predicting model is not found in the first block, the next block will surely bring one closer to obtaining one. Once a surrogate metamodel has been validated with checkpoints, simulationists will then have a powerful tool that should be able to give a suitably accurate prediction - instantaneously - of what will be produced hours later by a long-running simulation.

REFERENCES

Atkinson, A. C. (1985), *Plots, Transformations and Regression*, Oxford U. Press, New York

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

Chipman, H. (1996), "Bayesian Variable Selection With Related Predictors," The Canadian Journal of Statistics, 24, 17–36.

Cinotti, D. P., (2007) "Developing Masking Guidance with Respect to BW Trigger Events" 75th Military Operations Research Society (MORS) Symposium.

Kleijnen, J. P. C., Sanchez, S. M., Lucas, T. W., and Cioppa, T. M. (2005). "State-of-the-Art Review: A User's Guide to the Brave New World of Designing Simulation Experiments.'" INFORMS Journal on Computing 17 (3): 263–289.

Matheron, G. (1963), "Principles of geostatistics." Economic Geology, 58, no. 8: 1246-1266.

Mee, R. W. (2004), "Efficient Two-Level Designs for Estimating Main Effects and Two-Factor Interactions," Journal of Quality Technology, 36, 400-412.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," Statistical Science, 4, 409-435.

Sawa, T. and Hiromatsu, T. (1973). "Minimax regret significance points for a preliminary test in regression analysis." Econometrika. 41, 1093-1011.

Wheeler R.E. (1974) "Portable power." Technometrics. 16 (2). 193-201.

Wheeler R.E. (2002), *ECHIP Reference Manual*, ECHIP, Inc., Hockessin, DE

Wu, C. F. J., and Hamada, M. 2000, *Experiments, Planning, Analysis and Parameter Design Optimization*, Wiley, New York

Xu, H., Cheng, S.W. and Wu, C. F. J. (2004). "Optimal Projective Three-Level Designs for Factor Screening and Interaction Detection." Technometrics, 46, 280-292.

Y. Zhang, S. Pang and Y. Wang, (2001), "Orthogonal Arrays Obtained by Generalized Hadamard Product," Discrete Math., 238 151-170.

BIBLIOGRAPHY

Atkinson, A. C. and Donev, A. N. (1992), *Optimum Experimental Designs*. Oxford, Clarendon Press

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (2005), *Statistics for Experimenters*, 2nd ed., Wiley, New York

Fang, K. T., Li, R. Z., and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, Chapman & Hall/CRC Press, New York

Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd

Kleijnen, J. P. C. (2008), *DASE: design and analysis of simulation experiments*. Springer, New York.

Montgomery, D. C. (2005), *Design and Analysis of Experiments*, 6th ed., Wiley, New York

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer, New York

ACRONYMS

BW—Biological Warfare

BWA—Biological Warfare Agent

DOD—Department of Defense

DOE —Design of Experiments

HOR—Height of Release

LHC—Latin HyperCube

M&S—Modeling and Simulation

MORS—Military Operations Research Society

NOLH—Nearly Orthogonal Latin Hypercube

NPS—Naval Postgraduate School

OA—Orthogonal Array

OMLHD—Orthogonal-Maximin Latin Hypercube Design

PCAS—Probability of Casualty

RMS—Root Mean Square

SD—Standard Deviation

TBM—Tactical Ballistic Missile

DESCRIPTORS

Design of Experiments (DOE)

Efficient Experimentation

Computer Experiments

Modeling & Simulation

Surrogate Model

Metamodel

Biological Warfare

Regression Analysis

Cross-validation

Checkpoints

SAS Institute Inc. World Headquarters +1 919 677 8000 To contact your local JMP office, please visit **www.jmp.com**