Proceedings of the 1989 Winter Simulation Conference
E.A. MacNair, K.J. Musselman, P. Heidelberger (eds.)

PRE- AND POST-PROCESSING FOR DATA EFFICIENCY IN LARGE-SCALE COMPUTER SIMULATIONS

Turkan Kumbaraci Gardenier
TKG Consultants, Ltd.
Vienna, VA 22180, U.S.A.

ABSTRACT

Large-scale computer simulations demand the availability of methods to plan for simulation runs, and to economize on data requirements. Efforts to link input to various output measures have used the expertise of statistical experiment design in the following ways: (a) generating sub-fractions of a full factorial experiment design, still maintaining orthogonality for efficient estimation; (b) capitalizing on the primary network of synergistic and modulative interactions so that those parameters can be planned into the pre-processor requirements; (c) deriving methods to conduct a profiling analysis on a set of output measures of merit. The author has developed the PRE-PRIM and POST-PRIM systems, the following major features of which are presented here: (a) pre-processor types oriented toward screening versus estimating synergies/antagonisms; (b) mathematical and statistical requirements of orthogonality and balance; (c) regression-oriented software and POST-PRIM profiling approach to dealing with output measures; use of rotatable graphics; (d) tracing time-indexed output through CTSS, a major component of POST-PRIM to monitor output trends.

1. INTRODUCTION

Computer simulation has been applied to management decisions in a variety of settings such as queuing, military battle management, environmental and health risk assessment. One simulates in order to answer questions which are difficult to address analytically when a large number of input parameters are explored. Yet, the larger the simulation model, the interrelationship among variables and the relationship of inputs to output measures becomes difficult to analyze.

Metamodeling, as a technique, has been advocated to link inputs to outputs (Kleijnen 1985; Law and Kelton 1982). The objective in metamodeling is to derive a suitable linear or non-linear regression function relating input variables to an output criterion. The regression coefficients of the input parameters can be tested statistically and reduced to an efficient minimum. Expected values can be substituted for subroutines which are non-significant and result in economy in computer run time.

The importance of planning for the simulation runs using orthogonality principles of statistical experiment design in order to increase the reliability of results has been stressed by Gardenier (1982, 1988). In PRE-PRIM, the resultant pre-processor designs are interfaced with metamodeling regression software. In POST-PRIM, metamodel regression equations are submitted to a profiling analysis and interactive graphical evaluation. Time-series oriented tracing of criterion measures is accomplished by CTSS, a sub-component of POST-PRIM. The procedures embedded in PRE- and POST-PRIM thus increase the likelihood that the best combination of inputs are used to increase the effectiveness of the output variables.

2. PRE-PROCESSING CONSTRAINTS IN SAMPLE SIZE

In designing real-world experiments or computer simulation runs, a full factorial or Galois field is often considered initially. Within this context, the number of runs or trials can be determined by:

$$n = l^k$$

where n refers to sample size or the number of runs, l refers to the number of levels or partitions in each input variable, and k to the number of factors or variables. For example, a 2-level design with six variables would call for 32 runs, a 3-level design with 4 variables would need 81 runs.

A full factorial design can be interfaced with multivariate regression software. The output would yield unbiased, orthogonal estimates of all main effects, all two-way interactions, 3-way interactions, up to n-way interactions of model input parameters as well as a constant and error term.

When a large number of input variables exist, as often is the case in large-scale computer simulations, full factorial designs are infeasible and uneconomical. Interactiolns of higher order than second order are usually non-significant and bevome difficult to interpret. Then it is possible to use fractional factorial designs. They offer orthogonal estimates of all main effects and a subset of 2-way interactions for a fraction of the experiment costs of the full factorial. Fractional replicates can be constructed from a 2^p factorial design by defining p original defining contrasts for partition. This generates 2^{p-1} other defining contrasts, the principles of which exist in modulo notation. For example, a $\frac{1}{2}$ replicate of a full factorial design can be constructed by one defining contrast, a $\frac{1}{4}$ replicate by two defining contrasts. As an illustration, let us consider a 6-variable design which has input factors A, B, C, D, E and F. If we wish to reduce the total number of trials to one-eighth of what would be dictated in a full factorial design, the experimenter would need to specify 3 defining contrasts. Let us assume that we specified them to be the 3-way interactions ABC, ADE and BCF. Taking all combinationns of the variables and using modulo notation, the following additional contrasts would be generated:

$$(ABC)\ (ADE) = A^2\ B\ C\ D\ E = B\ C\ D\ E$$

$$(ABC)\ (BCF) = A\ B^2\ C^2\ F = A\ F$$

$$(ADE)\ (BCF) = A\ B\ C\ D\ E\ F$$

$$(ABC)\ (ADE)\ (BCF) = A^2\ B^2\ C^2\ D\ E\ F = D\ E\ F$$

The confounding pattern shown above may not suit the interests of the specific user. Therefore interaction is needed between the simulation user and design minimization component of pre-processors in order to attain a suitable subset of interactions which will be estimated orthogonally.

Fedorov (1972) summarized the procedures to construct optimal designs which maintain orthogonality and balance in the input study design matrix X. D-optimal ddsigns minimize the determinant of the inverse of $(X'X)$; A-optimal designs minimize its trace; E-optimal designs minimize its maximal eigenvalue. Fractional factorials have also been classified by a "resolution" concept. In Resolution III designs, some 2-way interactions may be confounded amongst themselves and with main effects. Resolution IV designs assure that 2-way interactions are not confounded with main effects, but they may be confounded amongthemselves. Resolution V designs assure that 2-way interactions are not confounded with main effects or among themselves and are most preferable to use from a simulation standpoint.

3. PROTOTYPE DESIGN PLANS

Pre-processor plans may be classified into the following categories:

(a) screening designs which are derived from Hadamard matrices; they estimate only main effects. Orthogonal estimates for k number of input factors may be obtained with these designs with as few as k+1 runs or observations. In these designs only two levels are used; non-linearities are unestimable.

(b) fractional factorials at two or three-levels; Taguchi (1986) has systematized a number of these plans within the Quality Function Deployment context. Interaction patterns are displayed in the form of linear graphs or triangular matrices in Taguchi plans.

(c) lattice designs used in mixture experiments; the sum of relative preference among alternatives is set to 100%. A q-dimensional factor space is represented by a q-1 dimension polyhedron; a polynomial of degree n is solved for to obtain the optimal mix.

(d) Latin squares and Graeco-Latin squares which estimate no interactions but may handle 3 or more level factors. The levels of all factors need to be equal. Graeco-Latin squares handle four factors; Latin squares use three factors.

The following section presents a contextual application for a number of these design plans.

4. APPLICATION OF VARIOUS PRE-PROCESSORS

The context for the choice and use of fractional factorials, interaction designs and mixture experiments will be given here as applied within a Monte-Carlo layered defense model setting.

4.1. Fractional Factorials

A number of main effects and a subset of interactions were explored, as described in Prouse and Gardenier (1987). A sequence of fractional factorials reduced the full factorial design requirements by $\frac{1}{2}$ at each stage. An example of the use of 10 parameters, five main effects and five interactions in the input design matrix of the pre-processor is shown in Table 1.

Table 1. Fractional Factorial Design with 5 Main Effects and 5 Interactions

	PLATFORMS	THREAT	P_K	P_{TRACK}	R_{WEAPON}					
RUN	X_1	X_2	X_3	X_4	X_5	X_1X_2	X_1X_3	X_1X_4	X_2X_4	X_3X_4
1	-	-	-	-	-	+	+	+	+	+
2	+	-	-	-	+	-	-	-	+	+
3	-	+	-	-	+	-	+	+	-	+
4	+	+	-	-	-	+	-	-	-	+
5	-	-	+	-	+	+	-	+	+	-
6	+	-	+	-	-	-	+	-	+	-
7	-	+	+	-	-	-	-	+	-	-
8	+	+	+	-	+	+	+	-	-	-
9	-	-	-	+	-	+	+	-	-	-
10	+	-	-	+	+	-	-	+	-	-
11	-	+	-	+	+	-	+	-	+	-
12	+	+	-	+	-	+	-	+	+	-
13	-	-	+	+	+	+	-	-	-	+
14	+	-	+	+	-	-	+	+	-	+
15	-	+	+	+	-	-	-	-	+	+
16	+	+	+	+	+	+	+	+	+	+

A total of 16 simulations represent $\frac{1}{2}$ fraction of the 2^5 or 32 observations which would have been required in a full factorial. Results obtained for a number of output parameters; e.g., leakage, time to RV impact, CP time, were analyzed by a multivariate regression routine in order to estimate coefficients. An equation was formulated for each output parameter in the format showing the regression constant and regression coefficients associated with each main effect and interaction. For example:

$$Y = 73.7 - 8.1\ X1 + 4.4\ X2 - 3.0\ X3 - 1.7\ X4 - 4.5X5 + 1.4\ X1x2 + .2\ X1X3 + .1\ X1X4 - .8\ X2X4 - .4X3X4$$

The coefficients of each output variable were evaluated statistically through the t-value of each coefficient. Only those parameters with significant coefficients were included in the next stage of model generation; in this stage the multivariate regression algorithm was applied again, using only the subset of critically relevant input variables. For example, in the equation above, X4 and all interaction terms associated with it were non-significant. The second stage model thus became:

$$Y = 73.7 - 8.1\ X1 + 4.4\ X2 - 3.1\ X3 - 4.5\ X5 + .3\ X1X3 + .4\ X1X4 - .6\ X3X4$$

Using the simpler model resulted in an increase in the coefficient of determination from .86 to .88.

4.2. Mixture Experiment

In a similar context, the query was formulated as to whether a model subsystem dealing with battle management weights had a significant effect upon the output. This subsystem was a man-in-the-loop configuration; the user interfaced with the battle management process rating, on a scale from 0 to 10, the importance to be given to five characteristics such as RV impact time and kill probability. In this context, the user decides on relative merits based upon competitive preferences, not on the assumption of independence inherent in full or fractional factorials.

With five input factors, the pre-processor used 21 runs, the mixture components of which are shown in Table 2. The first 5 trials gave held each of the five factors at highest weight, giving zero weight to the remaining four. The next 10 observations took combinations of two out of five in each run, giving them a weight of zero and a weight of .33 to the remaining three variables. Five trials were also allotted to a scheme, giving a weight of zero to each variable in turn and equal weights to the remaining four. The twenty-first or last run apportioned equal weight to all the five input variables. The design matrix and associated results was, again, submitted to multivariate regression analysis.

4.3. Interaction Design Constituents

A researcher is usually interested in synergistic effects of a specific pattern of sub-component input variables. In military battle-management oriented simulations, we may be interested in the degree to which electoric counter measures and decoy use enhance the probability of success more than what would be expected if each were used singly. Such synergies can be evaluated reliably if they are incorporated into the pre-processor as rows of the orthogonal input matrix.

Table 2: Preprocessor Design Combinations in a Mixture Experiment

#	RVNUMBER	TARGETTYPE	IMPACTTIME	KILLPROB	PLATFORMRES
1	0.000	0.000	0.000	0.000	1.000
2	0.000	0.000	0.000	1.000	0.000
3	0.000	0.000	1.000	0.000	0.000
4	0.000	1.000	0.000	0.000	0.000
5	1.000	0.000	0.000	0.000	0.000
6	0.250	0.250	0.250	0.250	0.000
7	0.250	0.250	0.250	0.000	0.250
8	0.250	0.250	0.000	0.250	0.250
9	0.250	0.000	0.250	0.250	0.250
10	0.000	0.250	0.250	0.250	0.250
11	0.333	0.333	0.333	0.000	0.000
12	0.333	0.333	0.000	0.333	0.000
13	0.333	0.333	0.000	0.000	0.333
14	0.333	0.000	0.333	0.333	0.000
15	0.333	0.000	0.333	0.000	0.333
16	0.333	0.000	0.000	0.333	0.333
17	0.000	0.333	0.333	0.333	0.000
18	0.000	0.333	0.333	0.000	0.333
19	0.000	0.333	0.000	0.333	0.333
20	0.000	0.000	0.333	0.333	0.333
21	0.200	0.200	0.200	0.200	0.200

Tables 3 and 4 encapsulate an example of this type of formulation. In Table 3, a set of interactions of interest which need to be formulated are displayed. These interactions are among 8 variables, four of which relate to database changes and four to man-in-the-loop oriented human factors in a simulation environment. Thus, X1-X4 are called input variables and X5-X8 process variables. The cluster of interactions specified are denoted with arrows.

Table 3: Formulated Cluster of Interactions in a Man-in-the-Loop Model

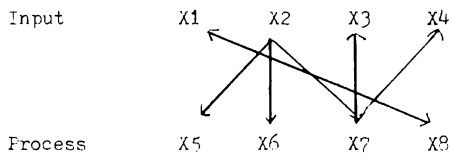

Table 4 shows a way of translating the set of interactions formulated into a metamodeling regression framework. A total of 14 regression parameters are generated, 8 main effects exploring the variables singly and 6 in combination. These are expressed in multivariate model form which can be regressed against an output criterion variable such as percent effectiveness of the battle management strategy used in the specific scenario. This specific preprocessor design required 47 simulation runs of the model in a user-oriented interactive system.

Table 4: Regression Metamodel Inputs of the Variable Interaction Cluster

Primary Interactions of Interest

	Input	Process
1.	X1	X8
2.	X2	X5
3.		X6
4.		X7
5.	X3	X7
6.	X4	X7

Multivariate Model Form

$$Y = b_0 \pm b_1 X1 \pm b_2 X2 \pm b_3 X3 \pm b_4 X4 \pm b_5 X5 \pm b_6 X6 \pm b_7 X7 \pm b_8 X8$$

(Additional Interactions)

$$\pm b_9 X1X8 \pm b_{10} X2X5 \pm b_{11} X2X6 \pm b_{12} X2X7 \pm b_{13} X3X7 \pm b_{14} X4X7 + \text{error}$$

5. COEFFICIENT ESTIMATION AND INTERPRETATIVE GRAPHICS

The coefficients estimated through the application of multivariate regression analysis are then used to formulate an equation which has been denoted as a meta-equation for surrogate models within the military battle-management milieu. The reliability of the overall equation is the coefficient of determination or the square of the multiple correlation coefficient used by statisticians.

Interpretative evaluations of the model can be enhanced by the use of graphical tools. POST-PRIM includes 3-dimensional rotatable graphics where the value of the output can be plotted for any combination of input variables. The response surface displayed gives the uder a feel for how a variation in input variables affects output. Tradeoffs can also be explored by holding the value of the criterion measure, Y, constant and displaying the value of three different inputs. Figure 1 shows a display of this type. The three axes show the plane created by the variation of three variables as well as the expected range of model variability.

versatile indices for analysis and display and use discrete partitioning of time-series data. User interface with an evaluative scheme to time-indexed changes adds uniqueness to the method. Shift in trends are graphically displayed and can be tested by nor-parametric statistical procedures. An illustrative sample graph is shown in Figure 2. Neither a plot of original values, nor Autoregressive Integrated Moving Average techniques (ARIMA) as described in Box and Jenkins (1976) would entail the clarity in display and show the change in trend near the terminal point of the time-series in the first part of Figure 2.

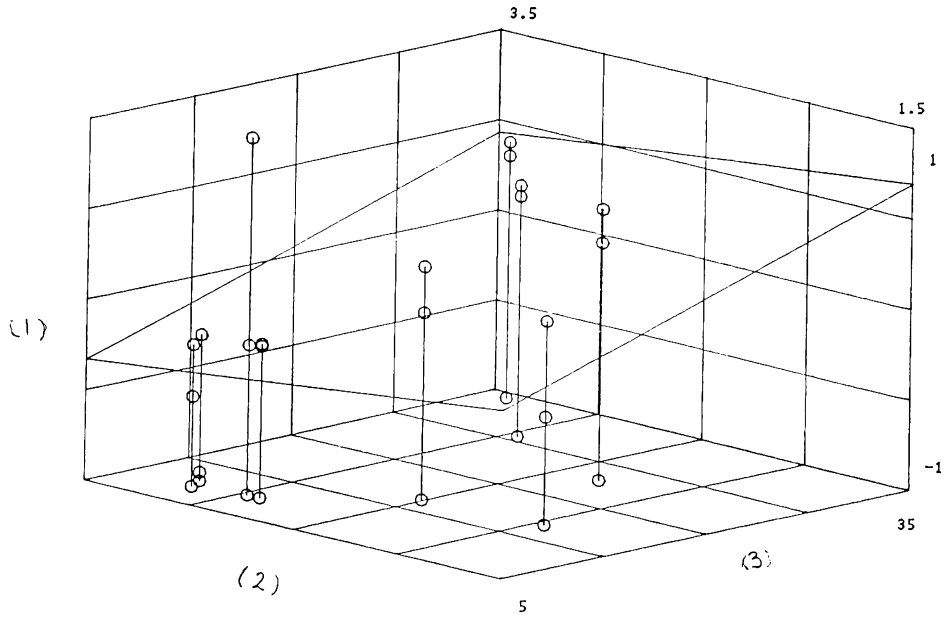

Figure 3: Rotatable Response Surface Sample Display used in POST-PRIM

POST-PRIM also includes a multiple criterion output screening module for criterion measure profiling analysis. The interplay of various endpoints and the importance of multiple criteria has been explicated by Zeleny (1982). POST-PRIM profiling takes into account the multicollinearity of output variables and leads to data efficiency by eliminating partially the data collection burden of monitored studies (Gardenier, et. al. 1989).

6. MONITORING USING CTSS IN POST-PRIM

Adaptive indices for monitoring, the Transitional State Score (TS) and the Cumulative Transitional State Score (CTSS) developed by Gardenier (1982 b) are also incorporated into POST-PRIM. They have proved to be

6. CONCLUSIONS

This presentation has demonstrated the importance of (a) pre-planning for simulation runs in order to reduce the experimentation costs and increase reliability of results, (b) post-processing of metamodels with interactive graphics, multi-criteria profiling and monitoring indices. The various components of PRE- and POST-PRIM developed by the present author assist large-scale computer simulations in this effort. Competitive preference decision model can be effectively analyzed and displayed using the surrogate modeling procedures discussed above. The utility-oriented regression equation, along with the pre- and post-analysis components of the system described here, creates an effective tool for decision making.

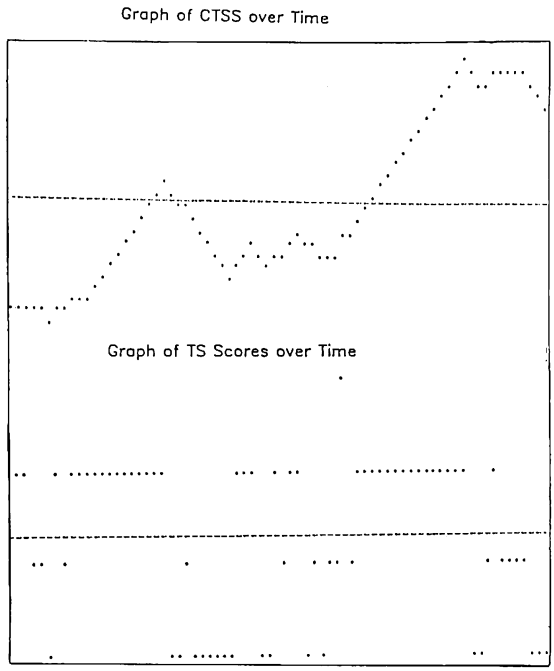

Figure 2: Adaptive Monitoring Index of POST-PRIM

ACKNOWLEDGEMENTS

The author gratefully acknowledges the encouragement and support of IIT Research Institute, particularly Dr. Frederick Bock, during the concept development of pre-processors to multivariate regression. The specific computer simulation applications were conducted under subcontract number B-716 with ANSER Corporation for work funded by the Strategic Defense Initiative Organization and National Aeronautics and Space Administration. Special thanks go to Dunell Schull, Douglas Brouse and Robert Petkewicz of ANSER while the author was applying the procedures to the Airlift Options Evaluation Model (AOEM), Blue Defender and the Direct Ascent Nuclear ASAT (DANASAT) simulations. CTSS and TS as indices were initiated by the author while at Pfizer Corporation; applications were conducted with support from the Centers for Disease Control, U.S. Public Health Service and Energy Information, U.S. Department of Energy. The author wishes to especially thank Drs. David Salsburg, Gladys Reynolds and Yvonne Bishop for their encouragement in the use and applications of CTSS in time-series monitoring. Dr. Richard Soland of George Washington University pointed me to metamodeling procedures for simulations.

REFERENCES

Brouse, D. and Gardenier, T. K. (1987). Regression metamodels for strategic defense simulation analysis. Presented at the 1987 Summer Computer Simulation Conference, Montreal, Canada.

Box, G. and Jenkins, G. M. (1976). Time-Series Analysis: Forecasting and Control. Holden-Day, California.

Fedorov, V. V. (1972). Theory of Optimal Experiments. Academic Press, New York.

Gardenier, T. K. ed. (1989). Data Efficiency Using Pre-Processing: Proceedings of First Symposium. Pro-File Computer Institute, Virginia.

Gardenier, T. K. (1988). Optimizing Computer Simulations: Lecture Notes and Practice Forms. Pro-File Computer Institute, Virginia.

Gardenier, T. K. (1982). Some uses of statistics in simu ation. In: Computer Models and Simulation: Principles of Good Practice (J. McLeod ed.). Society for Computer Simulation, California, 129-139.

Gardenier, T. K. (1982 b). Guidelines for risk monitoring. Presented as invited paper at joint meetings, American Statistical Association, Biometric Society and Institute of Mathematical Statistics, Cincinnati, Ohio.

Kleijnen, J. P. C. (1985). Statistical Techniques in Simulation: Vols I and II. Marcel Dekker, New York.

Law, M. M. and Kelton, W. D. (1982). Simulation Modeling and Analysis. McGraw Hill, New York.

Taguchi, G. (1986). Introduction to Quality Engineering. Asian Productivity Organization, Tokyo, Japan.

AUTHOR'S BIOGRAPHY

TURKAN KUMBARACI GARDENIER is founder and president of TKC Consultants, Ltd. specializing in simulation optimization and time-series analyses for government contracts. She received her undergraduate degree from Vassar College and M.A. and Ph.D. from Columbia University. She started and chaired the first Industrial Engineering Department at M.E.T.U., Turkey and taught at George Washington University, American University.

301 Maple Ave. West, Suite 100, Vienna, VA 22180, USA
(703) 938-9239