Proceedings of the 1980 Winter Simulation Conference
T.I. Ören, C.M. Shub, P.F. Roth (eds.)

DERIVING EMPIRICAL EQUATIONS FROM SIMULATION RESULTS

by

Leo J. Boelhouwer

System Communications Division
International Business Machines Corporation
Kingston, New York

ABSTRACT

Analysis of variance is performed in a formal sensitivity analysis of a simulation model. This technique identifies those variables that significantly affect the response variable(s). Then curves are fitted to the results to explain the response in terms of the significant variables. Finally, the equations are available to optimize the system under study.

1. INTRODUCTION

Although discrete simulation is a well-established technique, a difference of opinion exists between its adherents and others who prefer analytic methods. Simulation models offer great detail of representation but require excessive computer time for analyzing how different variables affect performance. On the other hand, analytic techniques clarify relationships among variables, but they may be too gross to adequately represent a given system.

A method is presented that combines advantages of both techniques. It produces representative equations based on a pre-defined sequence of discrete simulations.

To identify the most significant variables, factorial analysis is applied to the simulation results. Curves are then fitted to explain the results in terms of these variables.

2. FACTORIAL ANALYSIS

Factorial analysis has been around for decades. Used to analyze industrial and agricultural processes for optimization, the method identifies the most influential variables. It is important to note that the technique does not prove that there is a relationship between the controlled input variables and the observed response. It does establish that there is a high probability that such a relationship exists. Since the process under study is treated as a black box, the technique can easily be applied to the study of simulation models.

3. PROCEDURE

The first step is to select the independent variables that affect the response variable. To make sure that they are independent, select only those variables that can be directly controlled in simulation runs. Examples of such variables in a teleprocessing system are: line speed, MIP rate, message size, etc.

Next, a range is chosen for each variable. It must be carefully estimated because the results may only be valid within this range. Say a formula is derived that assumes a message rate to range from .5 to 1.5 messages/second. If the message rate is later shown to be 2 messages/sec., the formula is likely to give incorrect results. On the other hand, it is desirable to limit the range of a given variable, because the response is more likely to vary linearly. In any case, if the range is too small or too large, it will probably lead to doing another set of measurements.

With the ranges established, the measurements can be outlined as follows. Each variable is allowed to take on only its minimum or maximum value. For n variables, there are 2^n measurements. If the number of measurements becomes unreasonable, it may be possible to combine several variables into a dummy variable. That way, the number of measurements can be made manageable. This point will be addressed in more detail with various other aspects of using the technique.

Once the measurements have been performed, the results are tabulated for each response variable. Simple arithmetic identifies the significant controlled variables. Finally, a curve is fitted to the results, giving a formula for every response variable.

4. EXAMPLE

Assume the following hypothetical situation. Ten teleprocessing lines, with 10 terminals each, are connected to a control unit. The control units funnel the traffic into a processor with disk as the secondary storage medium. The objective is to study response time as a function of a few variables, with assumed ranges as shown. The cost column lists an assumed difference in monthly cost between the two extremes.

Variable	Range From	Range To	Δ Cost/mo.
C = CPU MIP (M inst/sec)	.36	.90	60K
L = line speeds (cps)	600	800	10K
M = message rate (mess/sec)	.75	1.5	
D = disk access time (ms)	30	50	2K

FIGURE 1

There are 4 variables, each of which is allowed to take on the 2 values at the extremes of its assumed range. Let the high value of the range be indicated with a lower case letter, and the low value by a blank. (For instance, C=.90 is represented by c.) Further, let the measurement with all controlled variables at their lowest value be represented as the integer 1. Then the other required measurements can be specified by means of a simple rule.

First, the integer 1 is multiplied by the symbol for the high value of variable M, or m. So far, we have two measurement runs to be performed, represented by 1 and m. Next, this series of symbols is multiplied by c. This indicates two more measurements, represented as m and mc. Repeating this procedure with d and ℓ leads to the first column of Figure 2. This column is the key to the measurement process and subsequent analysis.

The next step in the process consists of running the model 16 times. All 16 possibilities are covered by setting the variables to their respective high and low values as outlined above.

If each model run starts with no transactions in the model, the actual sequence of model runs is at the convenience of the analyst. If transactions are left in the model between runs, the sequence of model runs should be randomized. This precaution prevents the sequence itself from becoming a variable affecting the results.

When the modeling results become available, the values of the response variable are arranged as in column 2 of Figure 2. (If there are several response variables, similar charts are made for each one.)

Next, the required arithmetic is performed. This means adding the results of adjacent pairs in Column 2 to make up the first half of Column 3. For the second half of Column 3, the first entry is subtracted from the second entry in each column pair. For example, 1.52 + 1.58 = 3.10, 1.87 + 2.56 = 4.43, . . . ; 1.58 - 1.52 = 0.6, 2.56 - 1.87 = .69, etc. In this manner, Columns 3 through 6 are generated (one column per controlled variable).

The first entry in Column 6 is the sum of the results of all the runs. The remaining entries in the column are used to calculate the level of significance for each variable or combination of variables. Column 7 shows which combination is linked to the entry in Column 6. For instance, 6.81 indicates the relative strength of variable M (message rate). Variable C (CPU speed) has a relative strength of 8.79 and there is a second-order effect or interaction between M and C with a relative strength of 5.15. There is even a third-order effect, MCD, due to a joint contribution of message rate, CPU speed, and disk access time. (Note the one-to-one correspondence between Columns 1 and 7.)

To determine whether a variable is significant, divide its effect by the effect of the largest order interaction, LMCD (.07), and square the result. To test for significance, compare this squared ratio to the corresponding entry of an F distribution table for $F_{v1,v2,p}$, where v_1 indicates the number of degrees of freedom of the single variable ($v_1 = 1$), v_2 is the number of high-order effects lumped together to form the denominator of the ratio in question, and P is the desired level of confidence.

Since only the LMCD interaction is used here, v_2 is 1. For $v_1 = v_2 = 1$ and P = .95, the table entry is 161.4. So if the squared ratio is greater than 161.4, a 95 percent confidence factor can be established for the fact that the effect is signicant. And since the squared ratios are all greater than 1,000, all interactions are significant, except for those involving line speed. It appears to act independently from the other variables.

5. INTERPRETING THE RESULTS

The CPU speed, message rate and disk access time also directly affect response time. Since they act in concert, individually, and also via paired interactions, the next step is to quantify the effect from each of the three causes.

Let Δ represent a variable's range from 0 to 1. Thus, $\Delta_L = 0$ represents the line speed's low value and $\Delta_L = 1$ reflects its highest value. Actually, the designations 'low' or 'high' are arbitrary. In fact, in one case in this example, the usual convention is reversed. This means that in Figure 1, the 'high' value of CPU speed actually was assigned to the variable's lower numeric value. Although not strictly necessary, this was done to show only positive effects and to avoid confusion due to minus signs. What Figure 2 shows then is that decreasing the CPU speed increases the response time. Increasing the message rate and disk access time tends to increase response time as well, as one would expect.

Now an equation can be written in terms of Δ M (message rate), Δ C (CPU speed), Δ D (disk access time) and ΔL (line speed). When two variables interact such as M and C, the interaction is represented by Δ M. Δ C. Thus, when either ΔM or ΔC is equal to zero, the term's contribution is also zero. When both variables are equal to one-half, the term contributes one-fourth the full strength of the effect. The final formula is expected to be a function of ΔM, Δ C, Δ D, Δ L, Δ C. Δ D, Δ M. ΔC, ΔM. ΔD, and Δ M. Δ C. ΔD.

When a variable such as the line speed acts only by itself, the coefficient of Δ L can be found by dividing its entry in Column 6 of Figure 2 by half the size of the experiment. Because the size of this experiment is equal to 16 runs, the coefficient of Δ L is -2.35/8 = -.294. (The minus sign reflects the fact that increasing line speed decreases response time.) In this case, it is easy to verify that the last eight entries of Column 2 (where the line speed is high) differ from the first eight entries by about .29. Thus, changing the line speed from 600 to 1200 cps decreases the response time by .3 seconds regardless of any other changes.

For the other variables, their various effects must be isolated. There are sophisticated techniques for doing so (Johnson & Leone, 1964), but approximate results can be found as follows.

The only variable acting when Result 2 is compared to Result 1 is the message rate, accounting for .06 seconds (1.58 - 1.52) of response time. Since line speed has only an additive effect, another estimate for the effect of message rate is available by subtracting Entry 9 from Entry 10 (1.27 - 1.23) i.e., .04. Thus, the total range of message rate changes the response time by about .05 seconds. Similarly, the change in disk speed accounts for .52 seconds and the range of CPU speeds is responsible for .35 seconds.

To find the contributing effects of CPU and disk speed interaction, compare Entry 1 to Entry 7. Entry 7 has ΔD = ΔC = 1, accounting for .87 seconds. Added to the value of Entry 1, this gives 2.39, leaving .22 seconds to be accounted for by the Δ C. Δ D term. But Entry 15 provides a check. By subtracting the effect of line speed, .21 seconds is given for the Δ C. ΔD term. And continuing in this vein results in a response time equation:

$$RT = 1.52 - .29\,\Delta L + .52\,\Delta D + .35\,\Delta C + .05\,\Delta M + .22\,\Delta D.\,\Delta C + .33\,\Delta D.\,\Delta M + .64\,\Delta C.\,\Delta M + 1.27\,\Delta D.\,\Delta C.\,\Delta M$$

This equation accurately fits the observed values and offers some immediate insight into tradeoffs. For instance, if the incremental costs of changing the variables from their low to their high value are as given in Figure 1, the most economical approach is to increase disk speed in order to decrease response time. In this hypothetical example it costs \$3.85K (\$2K/.52) to decrease response time by 1

second via faster disk (although the range of disk speeds only allows a possible decrease of .52 seconds.) Note that all cost data is this and succeeding paragraphs are for illustrative purposes only.

Increasing line speed buys a second of response time for $34.5K ($10K/.29). And increasing CPU power decreases response time at a cost of $60K/.35 or $171K per second. (For simplicity, the higher-order interactions were ignored. The point is that the equation allows quick investigations of the trade-offs and a zeroing in on the most desirable configuration.)

Once the optimum configuration has been calculated, it is good procedure to check it with another run because the postulated model is quite simple. In case of a serious mismatch, it might be necessary to bound a new and smaller response surface around the calculated point for further investigation.

Several equations for other response variables can be developed without additional machine time. Thus, equations can be developed for utilization, or variance of response time, or any other measured response variable.

6. DISCUSSION

Several questions may arise :

How does one handle the situation where the number of measurement runs threatens to become prohibitive?

What happens when no variable appears to be significant?

How can the accuracy of a formula be improved?

When the number of variables of interest reaches 6 or more, the number of modeling runs quickly becomes unfeasible. There are at least two methods available to reduce the number of runs to a reasonable level. One way is known as confounding. This method sacrifices knowledge about high-order interactions in order to obtain information about most of the low-order ones. This is usually a reasonable compromise because interactions of 3 or more variables are rare. It is outside the scope of this paper to discuss confounding in detail; the subject is covered to the necessary depth by Duckworth (1968).

The other method of limiting the number of runs is to combine control variables into dummy variables and test these dummy variables for significant interactions. If none exist, the groups of variables can be examined one at a time, reducing the number of runs sharply.

For example, in an actual situation, the number of buffers in an operating system were thought to be dependent on (1) input message length, (2) output message length, (3) number of terminals, (4) processor instructions/message, (5) file accesses per message, (6) line speed, and (7) number of messages processed in parallel. To fully examine all possible interactions required 2^7 or 128 runs. Instead, the variables were split into 3 groups. One consisted of variables that affected the load placed on the system, namely input message length, output message length, and number of terminals. This group is referred to as the demand variable. The second group comprise the server parameters. It consists of processor instructions/message, file accesses/message, and line speed. The third variable is the number of messages that can be processed simultaneously.

To properly vary the dummy variables, it is necessary that all the contributing variables affect the response in the same way. If some variables at their high values tend to raise the value of the response variable while others lower it, the results will be inaccurate. Sometimes it may be necessary to hold all but one variable constant at their average value and vary the one variable to know how it tends to affect the response variable.

With the dummy variables properly constructed, it turned out that the server parameters were unimportant. Thus the number of controlled variables was reduced to 5 and the number of runs to 32. (The insignificant variables were left at their average value to prevent any bias.)

6.1 Indeterminate results.

It may turn out that the highest order interaction is too large for any effect to pass the significance test. The simplest solution to this problem is to rerun all of the measurements with another seed in the random number generator. If one considers the random number generator as another (insignificant) variable, the highest order interaction should be small enough to be useful.

If this approach fails to work, one probably needs to change the list of controlled variables. A significant one may have been left out of consideration. This may be tested by again assigning all variables their average value and varying a single candidate variable to test the effect on response variables.

6.2 Improved Accuracy

The formulas developed so far are based on the highest and lowest values of the controlled variables. In general,the behavior of a response variable has been predicted from its behavior on the response surface's boundaries. As a check, one might compare the formula and the model when all the control variables are in the middle of their respective ranges. If there is an objectionable discrepancy, it may be necessary to use 3 instead of 2 levels for one or more of the controlled variables. In addition to the two extreme values, one also uses the midpoint value of a variable. For a full treatment of this subject, refer to Johnson and Leone (1964).

However, sometimes it is merely necessary to change the assumed nature of the formula. For instance, in some cases it was found that $1/(1-a\Delta)$ gives a much better fit than $a\Delta$. The form is that of a single server queue.

7. CONCLUSION

Trade-offs considered in a design situation can be quickly evaluated using factorial analysis. Confusing actions and interactions of the different variables are clarified and approximate design equations can be easily formulated -- equations that allow a user to be aware of the effects of changing his design assumptions.

However, any of these equations are only valid within the boundaries defined by the ranges of the controlled variables. Extensions outside those ranges are likely to require more experimental design and more runs. Also, if a more detailed investigation of the response surface is required, a similar set of runs may be necessary with three values of some variables.

Once empirical functions (like the one described here) become generally established, new problems may fall into the bounded region of a previous problem. Thus, a formula derived earlier may also be applicable to the new situation. And, for the first time, results gained from one simulation model will reduce or eliminate simulation runs for a new problem.

ENTRY	COL 1	COL 2	COL 3	COL 4	COL 5	COL 6	COL 7
1	1	1.52	3.10	7.53	19.50	36.65	
2	m	1.58	4.43	11.97	17.15	6.81*	M = Message Rate
3	c	1.87	4.47	6.33	3.40	8.79*	C = CPU
4	mc	2.56	7.50	10.82	3.41	5.15*	MC
5	d	2.05	2.50	.75	4.36	8.93*	D = Disk Speed
6	md	2.42	3.83	2.65	4.43	3.89*	MD
7	cd	2.61	3.86	.71	2.54	3.47*	CD
8	mcd	4.89	6.96	2.70	2.61	2.63*	MCD
9	ℓ	1.23	.06	1.33	4.44	-2.35*	L = Line Speed
10	ℓm	1.27	.69	3.03	4.49	.01	LM
11	ℓc	1.58	.37	1.33	1.90	.07	LC
12	ℓmc	2.25	2.28	3.10	1.99	.07	LMC
13	ℓd	1.75	.04	.63	1.70	.07	LD
14	ℓmd	2.11	.67	1.91	1.77	.09	LMD
15	ℓcd	2.31	.36	.63	1.28	.07	LCD
16	ℓmcd	4.65	2.34	1.98	1.35	.07	LMCD

*NINETY FIVE PERCENT CONFIDENCE LEVEL THAT VARIABLE IS SIGNIFICANT.

FIGURE 2

Duckworth (1968), Statistical Techniques in Technological Research, pp. 102-107.

Johnson & Leone (1964), Statistics and Experimental Design.