Proceedings of the 1991 Winter Simulation Conference
Barry L. Nelson, W. David Kelton, Gordon M. Clark (eds.)

TOWARD A HIGHER LEVEL, OUTPUT ANALYSIS INTERFACE

Edward A. MacNair
Peter D. Welch

IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

To match the greatly increasing power of scientific engineering workstations with the slowly changing capability of the simulation practitioner, this paper examines possibilities for a high level output analysis interface. The goal is an interface which considers a parametric family of models and accepts commands stated in terms of understanding model characteristics over this family and/or making practical decisions with respect to it.

1 INTRODUCTION

With the increasing availability and power of engineering workstations, the data processing challenges to simulation practitioners will soon shift from problems involving the careful allocation of scarce computing resources to problems involving the effective utilization of plentiful computing resources. This will result in a challenge to the creators of simulation packages to provide a high level output analysis interface so that substantially less effort on the part of the experimenters will result in substantially more effort on the part of the machines. This interface will have to match the relatively unchanged capability of the user with the vastly increased capability of the workstation so as to keep both of them gainfully occupied. For example, instead of issuing a command for the machine to run a given model for a certain simulated time and provide certain output statistics, the user would issue a command for the machine to explore a set of performance measures over a specified parametric family of models and obtain a summary of the response surfaces to some specified resolution and accuracy. The interface will have to have the capability to take broad statistical goals as commands and produce meaningful summaries. These commands will be tied as closely as possible to the user's need for understanding the implications of the model and utilizing this understanding to make decisions. The responses may not be exactly what the user requires but they should provide valuable insight and provide the basis for a dialogue which will result in understanding and well founded decisions.

As a secondary requirement the interface should be able to produce a running summary of its progress in one or more of the windows of the workstation display. This will provide the user with the capability to abort or modify the analysis depending on what is currently evident. For example, he may wish to expand or contract the region of the parameter space being explored. This secondary requirement means that the system should proceed to explore the "important" things first, presenting information which is immediately useful and which becomes more and more refined until the final goal is reached. Hence it should not proceed in some exhaustive fashion which will not produce any conclusions until the end. It should also not present yes/no results of hypothesis tests without providing the type of data summary which can lead to additional experimentation. What we have in mind by this will become apparent as we consider the examples.

In such an environment the simulation package will have to, of necessity, proceed in a sequential fashion much the same as a person would. It will have to make some runs, see what information they provide, decide what additional runs to make, reassess the information, and so on until the final goal has been reached. In this paper we will explore possibilities for the form of such an interface. We will consider a number of situations proceeding from the simple to the complex.

2 CHARACTERIZING A SINGLE MODEL: SOME EXPERIMENTS WITH SEQUENTIALLY CONTROLLED CONFIDENCE INTERVALS

We first consider the validity of sequentially controlled confidence intervals. Such confidence intervals will be important throughout this discussion and hence it is appropriate to discuss their validity. They are also important in the simplest practical application of simulation: the case where an experimenter has a single model whose performance he wishes to characterize. We now discuss this case. We assume the characterization takes the form of a set of response variables whose expectations must be estimated. The most common type of control for simulation experiments involves the specification of limits on characteristics of the simulation mechanisms themselves such as simulated time, numbers of events, etc. They are not stated in terms of any statistical or decision oriented goal. If such controls are used in this case and confidence intervals are generated they may or may not meet the accuracy requirements. If they do not the experimenter would have to set new run control parameters and continue this process until he was satisfied. In this simple case, with a higher level interface, the user would give the simulation package his accuracy requirements either in the form of confidence interval half widths or relative half widths and the system would automatically run until these accuracies were achieved. The window would display the running point estimates and confidence intervals as a simple range plot. This process (without the display) is, in fact, implemented in some systems today. By a relative half width we mean the confidence interval half width divided by the point estimate. Such relative or proportional confidence interval requirements appear to us to be more natural and of greater practical importance than absolute confidence interval requirements so we will confine our discussion to them. Such sequentially generated relative confidence intervals have been shown to be asymptotically valid (see Nadas 1969). However if they are implemented sequentially in a fashion exactly parallel to their non-sequential application they display lower than expected coverage because they pass the criteria on the average too readily at low estimates of the standard deviation of the point estimate. We ran some experiments to investigate this effect and some simple techniques to correct for it.

Nadas showed the confidence intervals to be asymptotically valid as the criteria for the relative half width approached zero. The practical question is how valid are they for the accuracy levels common in practice and what precautions can be taken to make them valid. These questions have been studied for absolute accuracy criteria (see Anscombe 1953, Starr 1966a, b, Woodroofe 1986 and Edelman 1990) but not for the relative accuracy criterion. We now describe some simple experiments which investigate the effectiveness of two techniques for achieving good small sample coverage: creating a minimum sample size and inflating the t-multiplier.

We consider the method of independent replications and assume that more accurate estimates are obtained by increasing the number of replications rather than the run lengths of a fixed number of replications. We further assume that the outputs are normally distributed. This is not a controversial assumption since the outputs are generally averages. Hence we assume we have a sequence

$$X_1, X_2, X_3, \ldots$$

of i.i.d. normal random variables with mean μ and standard deviation σ. In the straight forward procedure for sequentially generating a confidence interval for μ with a given relative half width, at each stage ($N \geq 2$) we form the sample mean

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$$

and the sample standard deviation

$$s_N = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (X_n - \bar{X}_N)^2}$$

We then consider the standard t confidence interval (with theoretical coverage $1 - \alpha$) whose relative half width is

$$\frac{t_{N-1}(1-\alpha/2)\, s_N}{\bar{X}_N N^{1/2}}$$

where $t_N(x)$ is the inverse cdf of a t distribution with N degrees of freedom. The first time this half width is less than or equal to the target value, C, the procedure stops. The method has the desirable property of realizing a result whose accuracy, as measured by the confidence interval, is (at least) plus or minus a prespecified percent. For example, if C = .1, the quantity is "known" to some reasonable degree to within plus or minus 10%. To correct for the poor small sample coverage three techniques have been proposed: not starting the procedure until a moderate value of N, say N = 10 or 20; using multipliers larger than

Table 1: Usual t-Multiplier

$N_0 = 2$

		C			
		.2	.1	.05	.025
	.5	.905(.871,.931)	.943(.914,.962)	.960(.934,.976)	.943(.914,.962)
σ/μ	.3	.910(.877,.935)	.900(.865,.927)	.908(.874,.933)	.943(.914,.962)
	.1	.938(.908,.958)	.908(.874,.933)	.908(.874,.933)	.920(.888,.944)

$N_0 = 10$

		C			
		.2	.1	.05	.025
	.5	.925(.894,.948)	.933(.902,.954)	.928(.896,.950)	.948(.920,.966)
σ/μ	.3	.943(.914,.962)	.935(.905,.956)	.940(.911,.960)	.958(.931,.974)
	.1	.955(.928,.972)	.955(.928,.972)	.948(.920,.966)	.958(.931,.974)

$N_0 = 20$

		C			
		.2	.1	.05	.025
	.5	.938(.908,.958)	.938(.908,.958)	.945(.917,.964)	.953(.926,.970)
σ/μ	.3	.955(.928,.972)	.938(.908,.958)	.943(.914,.962)	.960(.934,.976)
	.1	.940(.911,.960)	.965(.941,.980)	.948(.920,.966)	.960(.934,.976)

$t_{N-1}(1-\alpha/2)$

but which converge to that value as $N \to \infty$, and adding a few extra samples after the criteria is initially passed. We investigated the first two of these procedures.

In a first set of experiments we studied the procedure using the usual t-multiplier (N-1 degrees of freedom) and considering three starting values: $N_0 = 2$, 10 and 20. In the case of $N_0 = 2$ there is no starting delay because a minimum of two samples is required to calculate the sample standard deviation and generate a confidence interval. The behavior of the procedure is a function of the coefficient of variation of the X's, σ/μ, and the target accuracy, C. We considered the values $\sigma/\mu = .5, .3, .1$ and $C = .2, .1, .05, .025$. To estimate the coverage, 400 replications were made under each set of conditions. The results are given in Table 1. 95% confidence intervals on each of the values are included in parenthesis. The confidence intervals were generated using the normal approximation to the binomial (See equation (21.11.8) of Hald 1952). Notice that the coverage is not satisfactory for $N_0 = 2$ but that it is only slightly low for $N_0 = 10$ and 20. Hence providing a delay appears reasonably effective and $N_0 = 10$ would be recommended on grounds of efficiency.

To try to raise this low coverage we conducted a second set of experiments where we applied a correction to the t-multiplier proposed by Edelman. Based on the earlier work by Woodroofe, Edelman proposed increasing the t-multiplier by decreasing the number of degrees of freedom by 5. The results of these experiments are given in Table 2. Here only the cases $N_0 = 10$ and 20 are considered. (The procedure requires N_0 to be at least 7.) Notice in Table 2 that the coverage is too high for large values of C when the coefficient of variation is small. The problem is that in these cases the procedure is not really sequential because by the time N_0 observations have occurred the relative half width is, with a large probability, already below the target. Hence it is basically a non-sequential procedure and the inappropriately inflated t-multiplier causes the high coverage.

To attempt to correct for this high coverage we modified Edelman's procedure so that at the initial test we used the standard t-multiplier and in subsequent tests we used the inflated multiplier. These results are given in Table 3. Notice that for this procedure the coverage is uniformly correct. Hence this procedure with $N_0 = 10$ seems the best of those examined.

In the next section we will propose a ranking and selection procedure based on multiple confidence intervals. For this procedure to be valid, sequentially generated confidence intervals need to be valid. We have considered only one of a large class of procedures for generating confidence intervals. This ques-

Table 2: Edelman's Correction to the t-Multiplier

$N_0 = 10$

		C			
		.2	.1	.05	.025
	.5	.933(.902,.954)	.933(.902,.954)	.928(.896,.950)	.948(.920,.966)
σ/μ	.3	.968(.944,.982)	.943(.914,.962)	.940(.911,.960)	.958(.931,.974)
	.1	.988(.969,.995)	.975(.953,.987)	.955(.928,.972)	.943(.914,.962)

$N_0 = 20$

		C			
		.2	.1	.05	.025
	.5	.938(.908,.958)	.940(.911,.960)	.945(.917,.964)	.953(.926,.970)
σ/μ	.3	.965(.941,.980)	.938(.908,.958)	.943(.914,.962)	.960(.934,.976)
	.1	.945(.917,.964)	.968(.944,.982)	.953(.926,.970)	.963(.938,.978)

tion of sequential validity is an important question for them all.

3 RANKING AND SELECTION PROCEDURES FOR MULTIPLE MODELS

We next consider the case where the experimenter has a finite number of models and wishes to compare them with regard to some performance measure. Again we assume this performance measure is the expected value of some output process. This is the situation discussed in the ranking and selection literature. We will take a different approach centered around sequentially determined confidence intervals and the Bonferroni inequality. There are many ranking and selection goals. Most can be put in this framework. As an illustration we consider the following: the experimenter wishes to make the necessary runs so that either

1. a reliable decision can be made as to which model is best and the performance of the best model characterized to some specified accuracy or,
2. after eliminating as many models as possible a subset remains which, even though specified to the desired accuracy, cannot be reliably ranked.

Presumably in this latter case, because of the accuracy requirements on the results, for all practical purposes the subset of models remaining are equivalent. The method which we propose requires only that valid confidence intervals can be sequentially generated whose expected width goes to zero as the amount of data increases without bound. If we generate K confidence intervals with coverage $1 - (\alpha/K)$ then by the Bonferroni inequality the K confidence intervals will be jointly valid with probability $1 - \alpha$. Hence, any non-overlapping of confidence intervals will imply a ranking with probability $1 - \alpha$. The procedure is extremely straight forward. It has two parameters: the joint confidence level, $1 - \alpha$, and the accuracy, C. First, initial confidence intervals are

Table 3: Modified Edelman's Procedure

$N_0 = 10$

		C			
		.2	.1	.05	.025
	.5	.943(.914,.962)	.960(.934,.976)	.968(.944,.982)	.963(.938,.978)
σ/μ	.3	.960(.934,.976)	.935(.905,.956)	.960(.934,.976)	.935(.905,.956)
	.1	.950(.923,.960)	.940(.911,.960)	.955(.928,.972)	.950(.923,.960)

$N_0 = 20$

		C			
		.2	.1	.05	.025
	.5	.950(.923,.960)	.953(.926,.970)	.955(.928,.972)	.953(.926,.970)
σ/μ	.3	.963(.938,.978)	.950(.923,.968)	.960(.934,.976)	.943(.914,.962)
	.1	.950(.923,.968)	.945(.917,.964)	.953(.926,.970)	.958(.931,.974)

generated, one for each model. These and subsequent intervals are with confidence $1-(\alpha/K)$ where K is the number of models. Suppose the best model is that with the smallest expected value for the output variable. Consider the model with the smallest point estimate. Consider its confidence interval. Get more data for all the models whose confidence intervals overlap its confidence interval (it overlaps itself) and whose relative half widths are greater than or equal to C. Then recompute the confidence intervals and repeat until there is no more data called for. At each repetition all models are considered not just the ones for which data has been added. You will end up with a subset of "best" models whose confidence intervals overlap the confidence interval for the model with the smallest point estimate and whose performance is known to an accuracy C. Furthermore, if the system displays the complete set of confidence intervals in a range plot as the method proceeds, the experimenter will see a meaningful history of the progress and be able to intervene in a reasonable fashion.

To illustrate the procedure we simulated the behavior of an independent replications study of 5 models. We assumed that the replications for each model produced a series of i.i.d. normally distributed random variables with the means and standard deviations indicated in Table 4. The mean of each sequence is the performance measure of interest. Model 3 has the smallest performance measure and hence is the best model. To obtain the initial confidence intervals a sample of size 10 was taken. All confidence intervals were generated using the standard t-multiplier. At each stage one replication was added for each model meeting the criterion described above. The accuracy parameter was C = .1. Altogether the procedure went through 131 stages. Only Model 3 remained at the end. Hence only it had the maximum of 140 samples. Figure 1 contains a series of 6 equally spaced snapshots of the progress of the procedure. It is a series of snapshots of what might appear in a window of the experimenter's display. Model 3 was seen as the best all the way. The confidence interval for Model 5 never overlapped that for Model 3 and hence no more samples were taken for Model 5 after the initial 10.

Table 4: Model Characteristics

Model	Mean	Standard Deviation
1	20	10
2	7	3
3	5	2
4	9	4
5	30	12

This procedure is easy for the experimenter to understand. The interpretation of the intermediate and final results are straight forward. There is an effective combination of selection criteria and accuracy criteria. This is important since, a priori, the experimenter has some idea how closely his model approximates the systems modeled and hence what differences between models are likely to be significant. The procedure can easily be extended to other ranking and selection criteria. At each stage more data is taken on any model for which a decision about the criteria cannot be made and which has not yet reached the accuracy goal. If you wish to identify the best k of the K models, you would take additional data on those models whose confidence intervals overlap some one of the confidence intervals corresponding to the k smallest point estimates and have not met the accuracy goals. If you wish to rank all the models you would take more data for any model whose confidence interval overlapped any other confidence interval and have not met the accuracy goals. This approach is discussed in a somewhat similar fashion in Section 10.3 of Law and Kelton (1991).

4 CHARACTERIZING THE BEHAVIOR OF A MODEL OVER A SPECIFIED RANGE OF A SINGLE PARAMETER

We next consider the case where the experimenter is interested in characterizing the expectation of a response variable which is a function of a single continuous model parameter. We assume he is interested in this function over some range, i.e. over an interval. As in the earlier discussion we wish to develop procedures which will try to obtain rough but important information first. This information will get more and more refined until some final goal is reached. In this case a reasonable goal would be the estimation of the function to some resolution and accuracy. Further, it might be desirable to obtain an estimate which was continuous and had a continuous first derivative. As before we would like to provide the experimenter with a graphical picture of the progress of the procedure in one or more windows of his display so that he can appreciate what has been learned and, if desirable, interact with the process.

As an example suppose the procedure tries to approximate the function by a polynomial over the range specified. Then, in line with the philosophy we are proposing, the procedure would start by fitting a straight line. It would implement an appropriately powerful test of the straight line fit. If this assumption were accepted it would test the accuracy

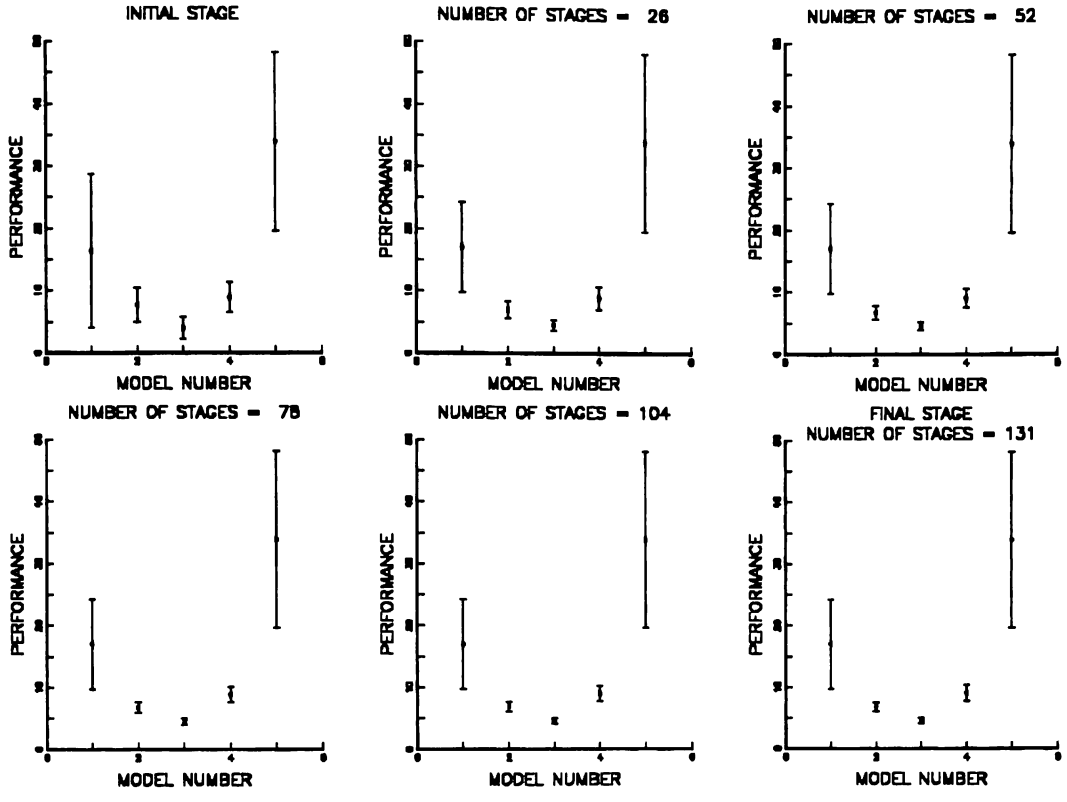

Figure 1. Six Snapshots of a Model Selection Procedure

criteria. If the accuracy criteria were met it would quit. If the straight line fit were rejected it would move to higher degree polynomials, adding data and checking the fit until either it decided it could not fit a polynomial or the resolution limits had been reached and all the confidence intervals met the accuracy criteria. If it were unable to fit a polynomial in this global fashion it could go on to try to fit a richer class of smooth functions using a technique such as local polynomial fitting (see e.g. Cleveland and Devlin 1988). If it remained unsure of the fit of a smooth curve it would present the results as a set of point estimates and confidence intervals meeting the resolution and accuracy goals. As a final display it would either show the final fitted function with its confidence intervals or the set of point estimates and confidence intervals with the point estimates connected by straight lines.

To illustrate this approach we consider the model depicted in Figure 2. It is a closed queueing model of a simple computer system. We consider the case where the number of customers is 100. It has six queues. There is an infinite server queue representing the terminal think time. There is a queue in front of the cpu and queues io1, io2, io3, io4 in front of storage devices. All these latter queues have a FCFS discipline. The service times for all 6 queues are sequences of i.i.d. exponential random variables. The routing from the cpu to the terminals and the storage devices is controlled by a sequence of i.i.d. discrete random variables with probability 0.2 for each path. The expected think time at the terminals is 9. The expected service time for the storage devices is .025. We will be trying to estimate the expected overall system response time as a function of the expected service time at the cpu. The overall system response time is the time from the departure from the terminals until the return to the terminals. We let θ_1 be the expected service time at the cpu. We let R be the expected system response time. We suppose the experimenter is interested in determining the relationship over the range $.005 \leq \theta_1 \leq .025$.

We used the method of independent replications. We ran each replication for 500 system response times and discarded the first 100 to control the initial transient. The estimate was the average over the undiscarded 400 response times. We next describe a possible procedure for trying to fit a smooth curve or defaulting to a set of confidence intervals. The experimenter picks a resolution, D, which is a power

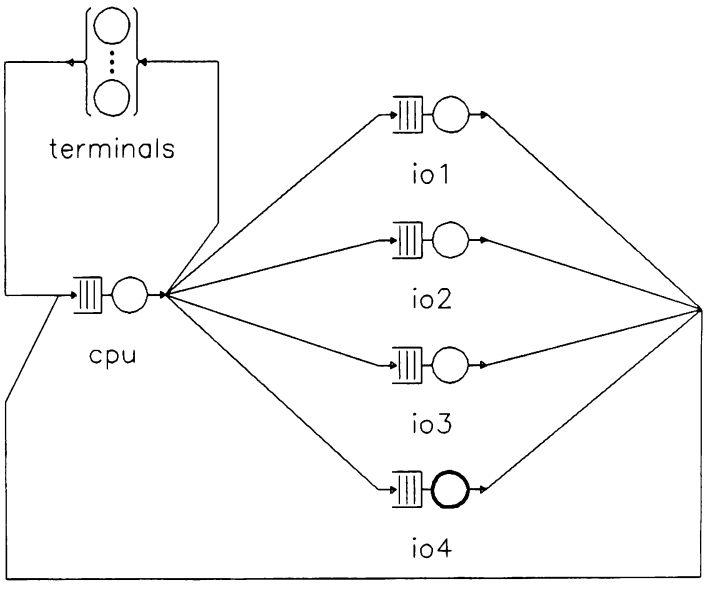

Figure 2. Model Diagram

of 2 and greater than or equal to 4. Runs will be made at, at most, D + 1 equally spaced points on the interval of interest. The experimenter also picks an accuracy, C. This is a relative half width accuracy over the interval of interest. We will assume it is a pointwise accuracy and does not hold simultaneously over the interval. If the procedure is able to fit a polynomial over the interval, the maximum value of the relative half width of the confidence interval on the fitted polynomial will be less than or equal to this accuracy. If it cannot fit a polynomial, data will be taken at D + 1 equally spaced points and the maximum relative half width over all the confidence intervals will be less than or equal to the accuracy C. In line with the goals stated earlier, the method begins by taking data at the two end points and the centerpoint of the interval. Since the assumption of equal variances is unrealistic, enough data has to be taken at each point to obtain reasonable estimates of the variances. We chose to take 10 replications initially at each point.

In Figure 3 we give a sequence of snapshots of the performance of the method for D = 8 and C = .1. These would be snapshots of the window on the display devoted to the progress of the procedure. Figure 3A shows a scatterplot of the initial data. Since we have 10 replications at each parameter point, we have a test of lack of fit. The application of weighted least squares (with the inverses of the estimated variances as weights) to the fitting of a straight line revealed a very significant lack of fit. This straight line is shown as a dashed line and a quadratic fit is shown as an undashed line. There are not enough points to check the quadratic fit so two more points were added with 10 replications each. These points are midpoint between the existing points yielding five equally spaced points in all. In Figure 3B we show a scatter plot of the data. The weighted least squares fit of the quadratic had a significant lack of fit. It is shown as a dashed line. The cubic fit passed the lack of fit test and is shown as an undashed line. To be reasonably sure of the fit, the method requires that the number of points be at least two greater than the number of coefficients of the fitting polynomial. Hence four more points were added equally spaced between the previous five. As before, 10 replications were taken at each point. In Figure 3C we show a scatterplot of the data. The cubic fit passed the goodness of fit test and is shown. Also shown are the pointwise confidence bounds on the fitted function. They are displayed as dashed lines. These confidence intervals do not pass the accuracy bounds and hence although the method has tentatively identified a fitting function it has not satisfied the accuracy criteria. The method then proceeds to add replications until the accuracy criterion are satisfied. In this experiment we added replications uniformly to each point. After adding 20 replications

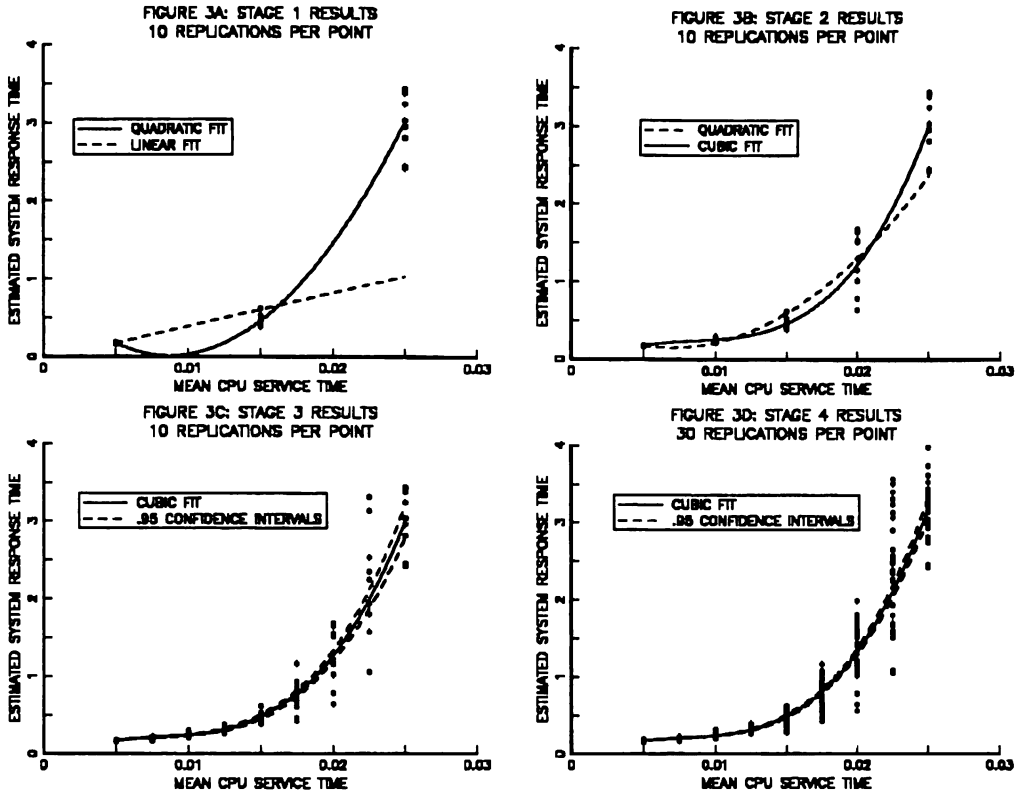

Figure 3. Sequence of Snapshots of a Single Continuous Parameter Procedure

at each point the accuracy criterion was passed and the cubic fit still passed the goodness of fit test. The final results are presented in Figure 3D.

5 CHARACTERIZING THE BEHAVIOR OF A MODEL OVER A REGION SPANNED BY TWO OR MORE PARAMETERS

In a parallel fashion we now consider the case where an experimenter is interested in characterizing the expectation of a response variable which is a function of two continuous model parameters: θ_1 and θ_2. We assume that he is interested in its behavior over a rectangular region. As before, he would like to fit a smooth function; however, if that is not possible, the default will produce an array of point estimates and confidence intervals. Again his requirements are stated in terms of resolution and accuracy. The resolution goal is given as a pair D_1 and D_2 where the D's are powers of 2 greater than or equal to 4. Runs will be made on a rectangular grid with the range of θ_1 divided into at most $D_1 + 1$ equally spaced intervals and the range of θ_2 divided into at most $D_2 + 1$ equally spaced intervals. The accuracy goal is a constant, C. The estimates will have confidence intervals with a relative half width less than or equal to C. As before, the search for a smooth function will take the form of the attempt to globally fit a bivariate polynomial.

We consider the simple computer system model described in Section 4. The two parameters studied were θ_1, the expected think time at the terminals, and θ_2, the expected service time at the cpu. We considered the region $3 \leq \theta_1 \leq 15$ and $.005 \leq \theta_2 \leq .025$. The procedure control parameters were $D_1 = D_2 = 4$ and $C = .05$. In Figure 4A we show a scatter plot of the initial set of data. 10 replications were taken at 9 points on a rectangular grid covering the rectangular region and splitting the range in half along each axis. We also show the fit of a bivariate quadratic. It was fit using weighted least squares with the inverses of the sample variances as weights. This bivariate quadratic did not pass the goodness of fit test. There were not enough points for the bivariate cubic so 10 replications were added at 16 additional points on a rectangular grid which covered the entire region and divided the range along each axis into 4 equal parts. Thus, at this stage there were 10 replications at each of 25 points. In Figure 4B we show the scatter plot of this data along with the fit of a

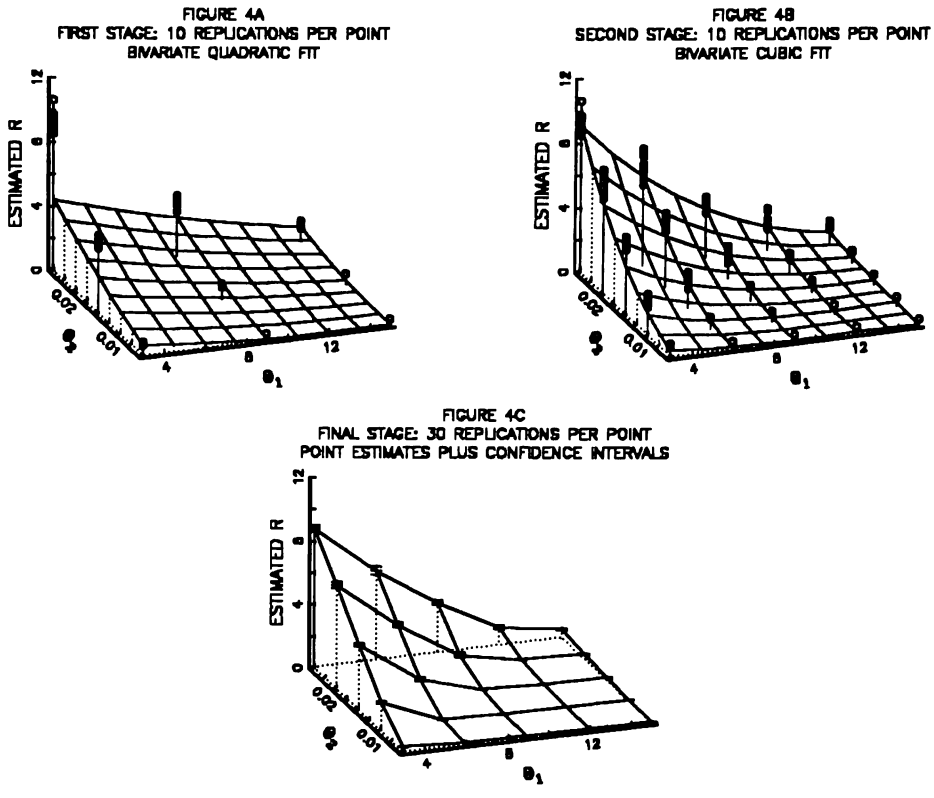

Figure 4. Sequence of Snapshots for a Procedure with Two Continuous Parameters

bivariate cubic. The bivariate cubic did not pass the goodness of fit test and we were not able to fit a bivariate quartic because of colinearity between the vectors representing the different terms. Hence the fitting of a smooth function by the global fitting of bivariate polynomials was unsuccessful. Finally, enough replications were added so that at each point the estimates met the accuracy criteria. The resulting estimates and confidence intervals are plotted in Figure 4C. Thus Figure 4 represents a sequence of snapshots of the progress of the procedure. The experimenter could at any point have intervened and modified the conditions and/or goals.

As the number of parameters increases the situation, of course, becomes much more complicated with a much greater variety of possible statistical metamodels and a greater availability of sophisticated techniques. Procedures have to be developed which

1. proceed in a sequential fashion from simple to more complex models,
2. continue to use the existing data while augmenting it with new data taking advantage of the ready availability of additional data in the simulation environment,
3. at each stage both estimate and test the model being considered,
4. at each stage provide the user with summary information so that he can intervene if desirable,
5. if unsuccessful at fitting a model, provide good summary information.

6 SUMMARY

With the proliferation of powerful scientific-engineering workstations, simulation practitioners are soon going to be challenged by a plethora of computing resources. To use these resources effectively, high level output analysis interfaces are going to have to be built. These interfaces need to exploit the unique features of the simulation environment: the fact that additional data is readily available and that this data can be immediately incorporated into the analysis and presented to the user. Thus many of the classical considerations of statistics involving making optimum use of limited data become eclipsed by questions of what additional data is

needed to make the necessary decisions and how to present the results as the data is being acquired so as to maximize the power of the combination of practitioner and machine. In this paper we have proposed sample procedures for a few of the situations arising in practice.

REFERENCES

Anscombe, F.J. 1953. Sequential estimation. *JRSS Series B* 15: 1-21.

Cleveland, W.S. and S.J. Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83: 596-610.

Edelman, D. 1990. The 5 degree of freedom rule of thumb for fixed width confidence intervals for a normal mean. Technical Report, Statistics Department, Columbia University.

Hald, A. 1952. *Statistical theory with engineering applications.* New York: John Wiley & Sons, Halsted Press.

Law, A.M. and W.D. Kelton. 1991. *Simulation modeling and analysis.* Second Edition, New York: McGraw-Hill.

Nadas, A. 1969. An extension of a theorem of Chow and Robbins on sequential confidence intervals for the mean. *Annals of Math. Statist.*, 40: 667-671.

Starr, N. 1966a. The performance or a sequential procedure for the fixed width interval estimation of the mean. *Annals of Math. Statist.*, 37: 36-50.

Starr, N. 1966b. On the asymptotic efficiency of a sequential procedure for estimating the mean. *Annals of Math. Statist.*, 37:1173-85.

Woodroofe, M. 1986. Asymptotic optimality in sequential interval estimation. *Advances in Applied Math.* 7: 70-79.

AUTHOR BIOGRAPHIES

EDWARD A. MACNAIR is a Research Staff Member in the Computer Science Department at the IBM Thomas J. Watson Research Center. His research interests are performance modeling tools and simulation output analysis. He is a member of ORSA, TIMS, and ACM, and was *Proceedings* Editor for the 1989 Winter Simulation Conference.

PETER D. WELCH is a Research Staff Member in the Computer Science Department at the IBM Thomas J. Watson Research Center. His research interests are graphical-statistical software and simulation output analysis. He is a member of ORSA.