How robust are conclusions from a complex calibrated model, really? A project management model benchmark using fit-constrained Monte Carlo analysis[*]

Alan K. Graham
Alan.Graham@PAConsulting.com

Jonathan Moore
Jonathan.Moore@PAConsulting.com

Carol Y. Choi
Carol.Choi@PAConsulting.com

PA Consulting Group
One Memorial Drive
Cambridge, Massachusetts
02142, USA
617-225-2700

Abstract

System dynamics-based simulation models are useful for analyzing complex systems characterized by both large parameter spaces and pervasive nonlinearity. Unfortunately, these characteristics also make confidence intervals for model outcomes difficult to assess. Standard Monte Carlo testing with a priori *realistic parameter variations produces simulated behavior that is* a posteriori *improbable, rendering simple Monte Carlo approaches inappropriate for establishing confidence intervals.*

This paper gives a case study of a model used to forecast completion of design and construction of a large defense program, and proposes a more correct Monte Carlo process, the fit-constrained Monte Carlo analysis. A confidence interval for outcome is computed, using Monte Carlo trials and discarding combinations that do not achieve an acceptable fit of simulated behavior to historical data. For this case, the experiment confirmed the intuitive view that a well-formulated closed loop model calibrated against sparse but widespread data and an appropriate statistical fit criterion can create tight confidence intervals on some model outcomes. By contrast, conventional (non-fit constrained) Monte Carlo results give substantially misleading implications for a confidence interval. The correlations between model parameters and outcomes are also explored, but they do not reveal significant issues with the method or results.

Keywords—Monte Carlo, fit-constrained parameters, historical data, *a posteriori*, system dynamics, confidence interval, outcome, calibration, project, program

[*] Acknowledgements: The authors are grateful to their PA colleagues that contributed to this analysis, including Marlene Casciano, Ken Cooper, Keith Eubanks, Tamara Kett, Jim Lyneis, Evonne Meranus, Tom Mullen, Kim Reichelt, and Craig Stephens. A partial report on this work was presented earlier as (Graham, Choi and Mullen 2002).

1. Introduction: The value of reducing uncertainty

To some degree, all quantitative models are imperfect. Everyone is therefore uncertain, to some degree, about the outcomes computed from those models. When taking action based on those models, we would like to understand the ramifications of that uncertainty. To put a point on one issue that will return again when the case application is described, the models and outcomes being addressed here (and described further below) are dynamic systems whose outcomes have high stakes in the real world, which makes understanding uncertainty a matter of considerable importance. A few examples are predicting commercial aircraft demand (Lyneis 2000), optimizing a marketing strategy (Graham and Ariza 2001) and managing design and construction of major development programs like ships, automobiles, and airplanes (Cooper 1980, Cooper and Mullen 1993, Graham 2000). These are situations where the difference between good answers and not-so-good answers has a major impact, typically hundreds of millions or billions of dollars in value. It is therefore appropriate to devote significant time and effort to quantify confidence in these models.

1.1 Strategies for dealing with uncertainty vary widely

Although the totality of techniques for dealing with uncertainty is too broad to survey here, in the practice of using quantitative models we can observe a spectrum of approaches:

1. *Ignore uncertainty.* This is a common approach, where an estimate is implicitly (or sometimes explicitly) treated as accurate, or at least well within the relevant tolerances. One variation is to acknowledge that an output is not certain, and leave it to consumers of the analysis to apply their own judgment of uncertainty and its consequences.
2. *Use arbitrary uncertainties.* Moving slightly forward is to explicitly recognize the uncertainty and supply a judgment of range. Analysts often provide such ranges, stating a result of "x" within an uncertainty of, say, +/- 10%. Sometimes the range of uncertainty provided is not purely arbitrary as it is based partially on the analyst's experience in similar circumstances. An approach common in dynamic modeling is to conduct limited sensitivity tests, and characterize uncertainty on that basis.
3. *Compute confidence intervals to verify robustness of conclusions.* If one can compute the confidence intervals for the model outcomes, and demonstrate that they make little difference to the desirability of actions being considered, all is well.
4. *Choose actions to deal with uncertainties.* If one computes the confidence intervals of outcomes and discovers that the choice of actions in fact depends on factors that are at present unknown, there are still possibilities. Sometimes one's choices can be recast in terms of real options (Copeland et al. 1995, Luehrman 1998ab). Adaptive control is a somewhat broader theoretical framework for dealing with uncertainty, specifically the mathematics of designing control systems to identify changing conditions and formulate an optimal response (Schweppe 1973, Feng and Lozano 1999) or at least a robust response (Rohrer and Sobral 1965). To the authors' knowledge, however, these theoretically elaborate and computation-intensive techniques have not been materially applied to actual corporate management systems, although Ackoff's (1981) Interactive Planning methodology for corporate strategy makes use of adaptive control *concepts*.

System Dynamics has long been practiced as a fundamentally deterministic method, usually staying between levels one and two above. While for some problems this is an acceptable simplification, for many problems it is not. In terms of establishing confidence in model outcomes, the further one can move through this sequence, the better. Of course, there are challenges and costs as one progresses, and the path becomes increasingly difficult when the analysis is based on a complex nonlinear model. The most pivotal step in understanding uncertainty and its consequences is computing confidence intervals.

1.2 Neither standard econometric computation nor standard Monte Carlo are appropriate for quantifying confidence in outcomes of typical complex dynamic models

If one extrapolates from the world of Ordinary Least Squares (OLS) regression modeling, establishing confidence intervals may seem straightforward. There are well-known formulas for parameter and forecast confidence intervals (Theil 1971). However, the basic formulas rest on assumptions which are nearly always impossible to satisfy in a commercial modeling setting. As a terminological note relative to econometric conventions, econometricians such as Theil distinguish between a *confidence interval* (the range within which a parameter estimate will differ from the theoretical "real" value) and a *prediction interval* (the range within which a prediction will differ from the theoretical "real" condition). Because the outcome of a model analysis may be a prediction, but may also be a "what if" or a strategy conclusion, we prefer, rather than "prediction interval" to use the more informal and general "confidence interval for the outcome".

Individually, the difficulties with standard statistical techniques are sometimes surmountable, but the complexity of correctly dealing with them simultaneously increases rapidly. In a large-scale dynamic model, although econometric methods may be used at times for individual equations, they are virtually never suitable for assessing confidence in overall results. While System Dynamics modelers can and do deal with the problems above, with methods consonant with the relevant theory (maximum-likelihood estimate for dynamic systems (Graham 2002, Schweppe 1973, Peterson 1980), the methods are mostly not econometric.

The usual approach to analytical intractability of uncertainty is Monte Carlo simulation: Just take a few thousand samples–enough to show statistical significance–and one can know the distribution of the outcome (Hammersley and Handscomb 1964, Rubinstein 1981, Fishman 1996). Indeed, simple Monte Carlo analysis has been intermittently explored within the System Dynamics community for some time (Phillips 1980, if not earlier). However, for dynamic models, there are again problems.

In any modeling effort involving calibration to real data, at least some of the original parameter values will have been selected in part for their consistency with observed data on the real system's behavior. Consequently, randomly-selected parameter variations around that original set–even relatively minor variations that are *a priori* plausible–may produce model behavior inconsistent with known real behavior. Such parameter sets are clearly *a posteriori* implausible. To run a Monte Carlo exercise that produces confidence intervals in outcomes, we must somehow constrain the combination of parameters selected to those that produce behavior consistent with known facts.

For large nonlinear dynamic models, no closed-form calculation is possible that would translate the *a posteriori* constraint that combinations of parameters produce realistic behavior into *a priori* constraints that yield only combinations of parameters that fit the observed data—linear methods (e.g. Morgan 1966) break down.

In other problem domains, there has been work using Markov chains and other methods to create Monte Carlo analyses under *a posteriori* constraints (Metropolis et al. 1953, Hastings 1970, Geman and Geman 1984, Gilks et al. 1995). However, given that we have been doing the first explorations of fit-constrained Monte Carlo in the domain of dynamic models of large programs, we elected to use a simple brute-force approach. We pick parameter sets according to *a priori* knowledge, conduct the simulation trials, and discard the parameter sets and simulations that did not adequately fit the real data. As we shall see later on, this screening process makes a substantial difference in the variability of outcomes and the confidence that modelers can have in results.

It should be mentioned that there is a related thread of methodological inquiry, with search-based sensitivity analysis, e.g. (Wong 1980, Miller 1997), that partially address the flaw of sensitivity analysis that results in *a posteriori* unrealistic behavior (e.g. Vermuelen and DeJongh 1977). These methods find maximum and minimum possible outcomes, constrained by fit (either a hard or soft constraint). But these maxima and minima are only loosely and conceptually linked to the likelihood of those outcomes—the searches provide an upper bound on the width of a confidence interval. So these methods, while identifying interesting insights, will not directly identify a confidence interval. There have been theoretical inquiries within the authors' company on search-based approaches to confidence intervals since 1989, but we did not articulate a practically-implementable approach (fit-constrained Monte Carlo) until 1995.

Section 2 describes the modeling methodology used, as it pertains to the outcomes confidence interval problem. Section 3 characterizes the particular development program model and the Monte Carlo experiment. Section 4 gives the outcome confidence interval, and Section 5 concludes.

2. Modeling Methodology

Although the model used in the analysis cannot be fully disclosed here, due both to space constraints and commercial confidentiality, we can summarize the construction, validation, and broad characteristics of the model as they pertain to understanding the confidence interval problem. Summary descriptions of this series of models are available (Cooper 1980, 1993, Cooper and Mullen 1993, Graham 2000), and roughly similar published models appear in (Abdel-Hamid and Madnick 1991) and (Ford 1995). The construction and testing of the model and its predecessors follow the broad outlines of System Dynamics practice (Forrester 1961, Graham and Alfeld 1976, Graham 1980, Forrester and Senge 1980, Richardson and Pugh 1981, Sterman 2000).

2.1 A Series of System Dynamics models

The theory of cause and effect in large development programs embodied by the particular System Dynamics model discussed here is well validated. This model is one of a series of models of more than 100 large, complex programs, developed either to advise management on development strategy, or quantify consequences of actions for dispute resolution. They run the gamut of large-scale engineering, including major construction like nuclear reactors or the Channel Tunnel, development and production of aircraft (the F/A-18 E/F Hornet, for example), naval vessels, missiles and satellites, and developing large software programs (telephone switching, air traffic control, and air defense systems) and automobiles. The range of project performance runs from extremely successful (award-winning in some cases) to terminally unsuccessful (where the modeling was used to diagnose problems for either disputes or lesson-learning). In the dispute setting, the structure and behavior of these models has been critiqued by external academic modeling experts several times. In aggregate, this modeling activity represents many dozens of person-years of refining hypotheses (embodied in simulation models) about the cause and effect structure of complex development and construction projects, in many variations.

2.2 Data sources both qualitative and quantitative

Sources of calibration data include program data systems such as labor cost account records, progress reports and contractual documents. Where numerical data are not readily available, for example in the estimation of work quality loss due to protracted use of employee overtime (in general, "drivers"), structured interviews are used to obtain time-series for data based on first-hand knowledge. Although these data are treated as generally less accurate than quantitative time series, they have proven quite useful in understanding program behavior and calibrating the models. The interviews are conducted in accordance with the knowledge elicitation protocol described in (Ford and Sterman 1998). The phases of the elicitation process are:

1. Position - establish the context and goals of the session by providing multiple examples and operational descriptions focused on one non-linear relationship at a time,
2. Description - allow interviewees to visually record, graph and explain their recollections,
3. Discussion - compare, test, understand and refine the descriptions of individuals, subset groups, and/or prior groups sessions.

Interview sessions will usually be repeated at least once, when simulations reveal whatever inconsistencies exist among descriptions of cause and effect, recalled behavior, "hard data", and simulated behavior, in the usual hypothesis-testing cycle of the scientific method.

2.3 Hypothesis-testing

The initial hypothesis–the model structure and its parameters–come from both prior work and initial interviews that "rough out" the primary variations for the particular project (number and relationship of program phases, any major exogenous events, constraints on the program, etc.) The values of the *a priori* set parameters are originally determined through 1) interviews with program engineers and managers, 2) modeler's experience with similar design and construction programs in both the same and different industries, and 3) comparison of values against models of similar programs in the same industry from an internal database.

Most of the modeling effort is spent in the hypothesis—test—reformulation cycle of the scientific method, also known as calibration. Calibration refines many parameter values throughout the model (within bounds of *a priori* plausibility), sometimes refines elements of cause and effect structure, and often detects flaws in the measurement, interpretation (relative to model variable definitions) and aggregation of data. These flaws are detected because simulation and calibration provide a consistency check between model structure, data and *a priori* parameters.

At the beginning of the calibration process, mismatches between simulation and data are detected by visual inspection of time plots. As model behavior gets closer to the data, an explicit objective function, the "Average Absolute Error" (AAE) is introduced. (Lyneis *et al.* 1996) discusses this choice. AAE takes the absolute value of difference between the simulated and data values on a point-by-point basis, as a percentage of the mean of the data, and then takes the average of those values over each of the points in time.

The typical program is executed in multiple phases of design and construction effort, both in parallel and in series. Within each phase of work, we measure AAE on two of the three categories of calibration data: 1) cumulative progress achieved by quarter and 2) spend rate of direct labor resources by quarter. We use the third, more qualitative, data source ("drivers") for visual comparison only, due to the lower accuracy and imprecision of scales.

Independent AAE statistics are computed for the work progress and labor data series for each phase of the project, as well as the overall AAE statistic, which is a straight average of the individual series AAE statistics.

Calibration is considered good when:

1. Simulated progress values have AAE < 10%
2. Simulated staff profiles have AAE < 15%
3. The overall AAE statistic < 10%

These calibration standards arise from empirical observation (Lyneis and Reichelt 1996) of what is typically achievable with good calibration effort. Calibration substantially better than these standards has generally proven impossible without descending to a wholly inappropriate level of detail and exogenous inputs that add no predictive value.

3. Quantifying Confidence Intervals on the Athena Program

3.1 The Athena Program and its Model

The "Athena" program is a large-scale, complex defense development program employing advanced technologies in an evolving, competitive environment. Like most large development projects, the program experienced delay and disruption due to design changes and unexpected customer requests, technological problems and staffing difficulties. Finding the program over budget and behind schedule, the Athena program managers asked us to provide a mid-program estimate of the program completion date. For commercial reasons, it was important to quantify a statistical confidence interval on the program completion date estimate.

The dynamic simulation model built to analyze the Athena program is typical of the project models described above, containing several interrelated design and construction phases (typically between 5 and 15). Activity in each phase is primarily characterized by a rework cycle (Cooper 1980, 1993) in which technical work depends on the progress and quality of upstream work-phase products in order to make progress, and progress includes discovery of designs thought complete but in fact needing more work and to some extent, inadvertent creation of more such rework. The model consists of more than 300 non-linear ordinary differential equations and over 1000 *a priori* set parameters and initial conditions. The model is actually much more compact than these numbers might imply because equations and parameters are often subscripted to execute corresponding calculations for each of the design and construction phases.

The Athena model is calibrated against more than 1,000 data points, many in the form of quarterly time series data. About 100 data series, comprising about 80% of the data points, are used specifically for model calibration purposes. The remaining data is input to the model as initial setup conditions or exogenous drivers of behavior**.** The 100 calibration data series generally fall into three categories: 10% of the data represents quarterly spend of labor resources, 10% of the data represents progress achieved by quarter, and 80% of the data represents on-site personnel's best estimates of factors affecting productivity and quality performance over time. In total, the model tracks over 12 distinct AAE statistics as well as an overall AAE statistic.

3.2 Characterizing Uncertainties

To prepare for the Monte Carlo simulations, we categorized the 1000+ parameters in the Athena model, based on the degree of uncertainty in the *a priori* values, as described in Table 1. The parameters in the categories with very low uncertainty were not varied in the Monte Carlo experiment. The parameters in the five categories with moderate uncertainty (delays, levers, normals, weights and tuners on tables) were varied in the sensitivity tests. Slightly more than 50% of the parameters fell into this category.

We used a triangular distribution over a bounded range for randomly varying the parameter values. This choice allows the base value to be the most likely value (unlike the uniform distribution) while simultaneously confining the test values to lie strictly within a predefined range (unlike the normal distribution).

3.3 Monte Carlo Sensitivity Trials

We used a new simulation software package, Jitia (Eubanks and Yeager 2001), a successor to DYNAMO (Richardson and Pugh 1981) to perform the Monte Carlo sensitivity trials on the Athena model. We purposely began with very wide *a priori* parameter ranges to ensure we were not entrenched in a local optimum. Although we thought it unlikely that very different sets of parameter values could produce good historical fit, we did not want to ignore the possibility. Not surprisingly, we found that we had to perform the sensitivity trials several times under increasingly restricted ranges of variation before successfully identifying a statistically significant set of even moderately *a posteriori*-plausible results.

Initial parameter boundaries permitted trial values to range from a small fraction to several times the hand-calibrated parameter values. Prior to the first set, we were uncertain as to the yield that would be produced given the wide *a priori* parameter ranges. We found that with such a broad *a priori* range, even after running more than 10,000 simulations, none met the requirements of good historical fit to data on all work-phases. Of the randomly selected parameter sets, nearly all produced simulated results that were wildly dissimilar to program performance, rendering them clearly inapplicable to the task of determining a confidence interval for the outcomes of the Athena program.

Moderate Uncertainty	Very Low Uncertainty
Delays – time delays (e.g., Normal amount of time to gain authorization to hire employees) **Levers** – definitions of relative strengths of relationships (e.g., maximum overtime as a percentage of total hours under normal conditions) **Productivity and Quality Normals** – assumed normal values for productivity of staff and quality of work (Sterman 2000, 525-9, Graham and Alfeld 1976, 123-126). **Weights** – relative weights to average two or more variables (e.g., a downstream work phase may place equal weight on the availability of outputs from two upstream work phases) (Sterman 2000, 535-6) **X / Y Lookup Tables** – nonlinear relationships between an independent and dependent variable (e.g., the % reduction in productivity when proceeding with an incomplete design package). For ease and simplicity, calibration coefficients were associated with all table relationships. These coefficients allow the strengths of table relationships to vary without changing every point in the X / Y lookup table. (Sterman 2000, Ch. 14, Graham and Alfeld 1976, 154-7)	**Definitions** – definitional parameters (e.g., the number of quarters per year) **Links** – Boolean links to identify relationships between model phases (e.g., link to identify that work products from sector A were needed to make progress in sector B but not in sector C) **Initial Conditions** – values for levels at the start of the simulation (e.g., the initial value for the amount of fabrication work complete when the program began) **Program Targets** – parameters specifying the planned targets for the program (e.g., initial labor-hour budgets and schedules) **Simulation Settings** – inputs necessary for model construction but not relevant to simulation behavior and outputs (e.g., length of simulation)

Table 1. Characterization of parameters

We then performed several iterations of multi-thousand simulation sets, each time tightening the range over which we allowed the inputs to vary. Our working hypothesis was that the wider ranges allowed too many implausible input sets to be chosen. The more plausible sets had values near the base case values. By tightening the ranges, we would eventually narrow down on a range that both ensured the occasional selection of plausible sets while still allowing enough freedom for the selection of non-base case values.

Finally when we reduced the input range to 90% to 110% of the hand-calibrated parameter values and ran more than 50,000 simulations, less than 0.2% of them, 99 simulations, were within the minimum required AAE statistic of 10% on progress and 15% on labor data on all phases of work. We were at first mildly surprised that none of the Monte Carlo simulations produced an overall AAE statistic that fit the data better than the base case model. But of course, the hand-calibrated parameter values resulted from thousands of simulations during calibration that in effect have already come close to optimizing the fit.

4. Results

The simulated program lengths and associated error statistics, both relative to the base case (100%, 100%), are illustrated in Figure 1.

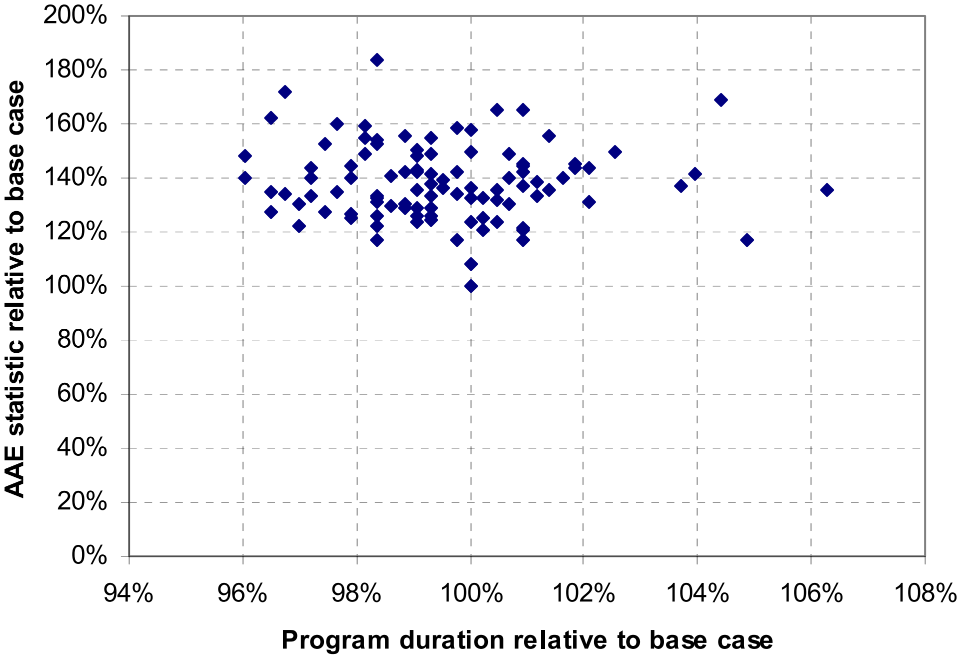

Figure 1. Scatter plot of simulated program durations vs fit (AAE) statistic

These results provide some validation of the empirically-chosen standards for adequate AAE fit. These standards are apparently sufficiently stringent to constrain the range of outcomes to a relatively narrow band.

4.1 Confidence Intervals

The 99 simulations meeting the minimum fit to data requirement produce overall AAE statistics between 6% and 10%. The mean of the distribution of outcomes (shown in Figure 2) essentially equals the base case estimate, with a standard deviation of 1.9% of the nominal program length. All program completion dates in the sample fall between plus and minus 6.5% of the base case completion date. Of the sample program completion dates, 90% fall between minus 3.5% and plus 4% of the mean.

Of course, it is always possible that some external action could create longer delays (e.g. workers could go on strike, the customer could introduce an unanticipated major specification change, etc.). Absent such an external event, the Monte Carlo analysis shows that the project history and the internal dynamics of rework discovery and completion are calibrated sufficiently accurately to yield a completion forecast with a pleasingly narrow confidence interval.

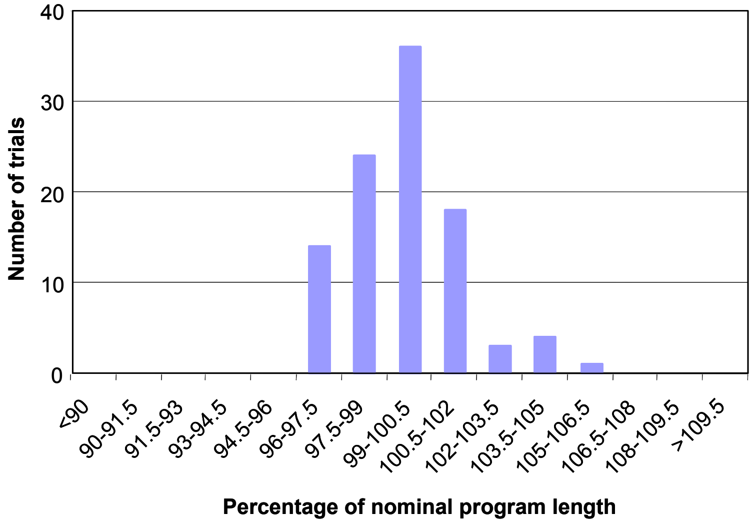

Figure 2. Range of likely program lengths as indicated by the 99 simulations with AAE statistics

In contrast, the distribution of the unconstrained results is much flatter and wider. Figure 3 shows the fit-constrained distribution (also shown in Figure 2), superimposed on results from the same analysis but without fit constraints applied. (Which means that the parameter variations that created the unconstrained case were already significantly narrowed from *a* priori ranges.)

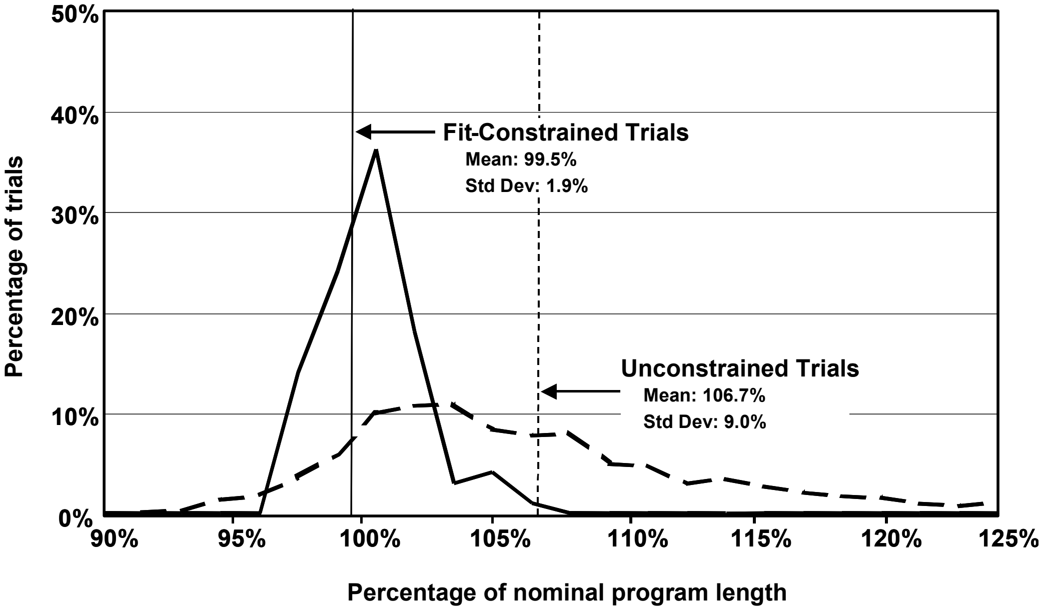

Figure 3. Comparison of program lengths as indicated by fit-constrained simulations vs. unconstrained simulations

Recall that most parameters were varied in a triangular distribution between 90% and 110% of the nominal value. The distribution of fit-constrained outcomes clearly has a smaller percentage standard deviation than its input parameter variations, whereas the non-fit constrained trials have a much greater standard deviation. The fit-constraining selection systematically eliminates large variations in fit, which are also clearly eliminating large variations in outcome.

The mean of the unconstrained outcomes is 106.7% of the nominal program length, more than 3 standard deviations away from the constrained mean. The standard deviation of the unconstrained outcomes is 9%, more than four times the standard deviation of the fit-constrained results. Before this investigation, the theoretical flaws in conventional (non-fit constrained) Monte Carlo were known, so misleading results were a theoretical possibility. But this comparison shows that, at least for this case, conventional Monte Carlo analysis is significantly misleading with regard to the confidence interval for the outcome.

4.2 Probing the selection process

Recall that practical limitations forced us to systematically constrain the *a priori* range of all parameters in order to increase the yield of plausible parameter sets. We did so by uniformly narrowing each sampling distribution. Having done so, it is desirable to understand in more detail the consequences and implications, because in future work, it should be possible to adopt more efficient algorithms for identifying *a posteriori* plausible trials.

4.2.1 Searching for under-constrained parameters

How completely does the fit-constrained Monte Carlo constrain parameter values? To begin understanding this question, we can look at how many parameters it is that the fit constraints actually impact during the trial selection process. We compared the 99 *a posteriori* samples against the corresponding *a priori* distributions for each parameter varied during the trials, using a chi-squared statistic. The numbers of parameters for which the equal-distributions hypothesis is rejected are displayed in Figure 1 below.

Chi-squared statistical significance	**Number of parameters**	
	Absolute	**Percentage**
95%	41	7.8%
99%	18	3.4%
99.9%	6	1.1%

Table 1: Parameters for which the equal-distributions hypothesis is rejected

At all three levels of significance, the number of rejections is greater than expected in the absence of any filtering, but not excessively so. The number of parameters showing significant changes in distribution is still close to what would occur by chance alone. In brief, the situation is modestly reassuring: There is no clearly-differentiated handful of parameters (at least by this test) around which the fit-constraint selection process revolves. The selection process appears to impact parameter values broadly.

Note that the test for a change to a parameter's distribution with variations of many other parameters is much stronger than a standard single-variable sensitivity test, for the test here allows the possibility of combinations of parameters being impacted by the filtering process.

We have examined the impact of fit-constraint, and we can now go on to examine the results, in term of characterizing the impact on the outcome of each parameter versus the impact on fit. Does the picture of broadly-based impacts from large numbers of parameters continue to hold? We tested the significance of two correlations for each parameter; each determined using a univariate linear regression:

1. The correlation between the parameter and the outcome, and
2. The correlation between the square of the deviation of the parameter from its base value and the goodness-of-fit statistic.

The choice of variables in item 2 above is motivated by the process that often selected the base case parameter values: minimizing the goodness-of-fit statistic. If a given parameter p is at a minimum with respect to fit, under variations in the value of p about its base value p_0, we would expect the slope of the regression relationship to be zero, and the second derivative to be positive (for upward curvature in both directions), i.e. the goodness-of-fit statistic should vary as $A(p - p_0)^2$, where A is positive for each parameter. If this is true we should see many parameters with positive correlation between fit and the squared error, and no negative correlations.

The results of these tests are displayed in Figure 4. Both axes show one-sided significance measures for the fitted correlations – more precisely, they show the associated Student's t-distribution cumulative density functions. Perfectly positive correlations would have a significance of 1.0. Perfect negative correlations would have a significance of 0.0. Zero observed correlation would have a significance of 0.5. Under the standard linear regression no-correlation null hypothesis, the observed significances would be drawn from a uniform random variable on the interval [0,1].

For each parameter, the vertical axis indicates the significance of its correlation with outcome, and the horizontal axis indicates the significance of the correlation with fit of its squared deviation from its base value. For convenience, the lines for 95% and 99% significance are shown. The parameters are distributed remarkably evenly with respect to their correlations with fit and outcome. The one exception is that, as expected, many more parameters show a positive correlation between the square of their deviations from their base values and the goodness-of-fit statistic than show a negative correlation. This is consistent with the selection of the base case values to minimize the goodness-of-fit statistic.

The distribution is also consistent with the hypothesis that there are no parameters that do impact the simulation outcome but at the same time do not impact the fit to historical data. In other words, this test fails to identify a significant incidence of parameters, beyond what one would expect by chance in a sample of several hundred parameters, that have a significant impact on outcome but are unconstrained by (i.e. do not impact) fit and thus perhaps the confidence interval.

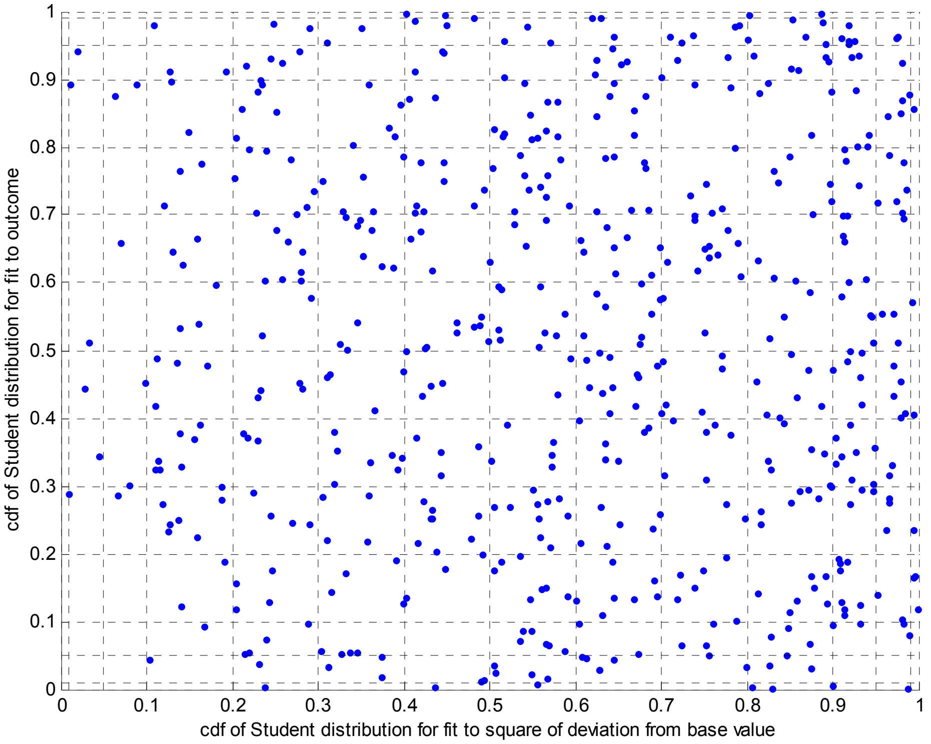

Figure 4: Student's t-distribution of cumulative density function values for regressions relating parameter values to outcomes (horizontal axis) and goodness of fit (vertical axis). The unlabelled grid-lines show values of 0.01, 0.05, 0.95 and 0.99.

4.2.2 The role of combinations of parameters

One hypothesis for why so few trials resulted in an acceptable fit is that good calibration requires certain pairs of parameter values to be appropriate to one another: In that case, changing one parameter without changing the other will cause the simulation to be rejected due to poor fit. Consequently, if this hypothesis were true, we should see a correlation between variations of parameter pairs in the sample of trials that do still meet the fit criterion.

As a particular example, we know that the normal productivity and quality parameters are usually set as residuals, to whatever value, when combined with values of the driving multipliers for productivity and quality, will recreate the known behavior. Therefore, we might expect that if a variation in some parameter requires an offsetting variation in normal productivity or quality, only the Monte Carlo trials where the normals also have the needed change–a tiny, improbable minority of trials–will have sufficiently good fit. In that case, we should see correlation between variations of the parameter in question and the variations of some normal productivity or quality parameter. Indeed, this was our going-in expectation, for normal quality and normal productivity, at least for "rough tuning" are conventionally thought to trade off against one another.

We examined three different sets of parameter pairs, in all cases excluding correlating a parameter with itself:

1. The distinct pairs of all the parameters
2. The distinct pairs of the normal quality and productivity parameters, and
3. The pairs where one parameter was a normal productivity or quality and the other was not.

For each set, we analyzed the constituent parameter pairs using univariate linear regression analysis, deriving one-sided significance measures for the fitted slopes, as described in Section 4.2.1 above. For each of the three sets, neither visual inspection nor a chi-squared test showed significant deviations from the null hypothesis of a uniform distribution. Therefore, the hypothesis that a significant fraction of the parameter pairs exhibited collinearity because of the fit-constrained selection is rejected. At least for the fully-calibrated model, our preconception about offsetting normal parameter values was incorrect.

In the case of the Athena project data and model, the general question of parameter identifiability and uniqueness of confidence interval seems answered, within the limits of this approach. (See Fisher 1966 for discussion relative to econometric models.) The statistical analysis fails to identify sets of interdependent parameters with significant correlations above chance levels. Moreover, if a large fraction of the sampled parameter sets had near-optimal fit but widely different outcomes, the graphs of outcome versus fit would have had a flat, level bottom, which Figure 1 does not show.

One fact that the authors did not initially appreciate is that the number of parameters, and combinations of parameters, constitutes a large enough statistical universe where even at higher significance levels, some outcomes will still happen by chance. The simple univariate tests described here imply that if identifiability and uniqueness problems do in fact exist, their magnitude is small enough that more elaborate tests, such as multivariate regression on fit and outcome, are needed to detect them.

On the other hand, caveats about method do not imply that the “truest” confidence interval would be wider than is reported here. Presumably, if the conceptual and technical problems of including “soft” recollections time series in the fit function were overcome, the “wiggle room” for parameter values would be even smaller, and the confidence interval would be even narrower.

5. Conclusions

Our work has repeatedly subjected us to intense queries from interested parties as to how confident they should be in model outcomes. Previously the best answers were qualitative statements of accuracy based upon extensive experience calibrating and using similar models (level 2 in the sequence described earlier). Faced with the need to quantify uncertainty in more rigorous ways, we have adopted fit-constrained Monte Carlo trials as a practically useful and analytically sound method of quantifying confidence in outcomes. The results confirm the long-standing belief within the field that the scientific method (iterative calibration to qualitative and quantitative information) creates a relatively tightly-constrained confidence interval for the outcome. The results also confirm that the existing qualms about conventional Monte Carlo analyses are correct: without consideration of whether parameters create realistic behavior, conventional (non-fit constrained) Monte Carlo analyses can yield substantially misleading results.

The effort required to carry out this analysis was significant. It took over a calendar month to carry out the whole Monte Carlo-based analysis, after the point at which we had provided a point-forecast of the base case. Such analysis would seem suitable only for problems where considerable effort is justified, and decisions can be deferred for several weeks without great loss. In the near term, however, improved software support and algorithms that are more efficient have the potential to reduce the required analysis time drastically. We believe that an analysis of the scale presented here should require only a few days rather than several weeks. If so, this approach would become applicable to a wide range of problems.

References

Abdel-Hamid, T. and S.E. Madnick 1991. *Software Project Dynamics: An Integrated Approach*. Prentice-Hall.

Ackoff, Russell 1981. *Creating the Corporate Future*. New York: John Wiley, New York.

Cooper, Kenneth G. 1980. Naval Ship Production: A Claim Settled and a Framework Built. *Interfaces* **10**(6), 20-36.

Cooper, Kenneth G. 1993. The Rework Cycle: Benchmarks for the Program Manager. *Project Management Journal,* March 1993.

Cooper, Kenneth G. and Thomas W. Mullen 1993. Swords and Plowshares: The Rework Cycle of Defense and Commercial Software Development Projects. *American Programmer* **6**(5).

Copeland, T., T. Koller, and J. Murrin 1995. *Valuation: Measuring and Managing the Value of Companies*. New York: John Wiley.

Eubanks, C. Keith and Larry F. Yeager 2001. An Introduction to Jitia: Simulation Software Designed for Developing Large System Models. *Proceedings of the 2002 International System Dynamics Conference*. Atlanta, Georgia.

Feng, G. and R. Lozano 1999. *Adaptive Control Systems*. Oxford, UK: Butterworth-Heinemann.

Fisher, Franklin M. 1966. *The Identification Problem in Econometrics*. New York: McGraw-Hill.

Fishman, G.S. 1996. *Monte Carlo: Concepts, Algorithms and Applications*. Springer Verlag.

Ford, David N. 1995. *The Dynamics of Project Management: An Investigation of Projects' Process and Coordination on Performance*. Cambridge, Mass.: Massachusetts Institute of Technology PhD dissertation, Department of Civil and Environmental Engineering.

Ford, David N. and John D. Sterman 1998. Expert Knowledge Elicitation for Improving Mental and Formal Models. *System Dynamics Review* **14**(4), 309-340.

Forrester, Jay W. 1961. *Industrial Dynamics*. Waltham, Mass.: Pegasus Communications, 1961.

Forrester, Jay W. and Peter M. Senge 1980. Tests for Building Confidence in System Dynamics Models. *TIMS Studies in the Management Sciences* **14**, 209-228.

Geman, S. and D. Geman 1984. Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images“, *IEEE Trans. Pattn. Anal. and Mach. Intell.* **6**, 721-741.

Gilks, W.R., S. Richardson and D.J. Spiegelhalter 1995. *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Graham, Alan K. 1980. Parameter Estimation in System Dynamics Modeling. *Management Science* **14**, 125-142.

Graham, Alan K. 2000. Beyond PM101: Lessons for Managing Large Development Programs. *Project Management Journal* **31**(4), 7-18.

Graham, Alan K. 2002. On Positioning System Dynamics as an Applied Science of Strategy. *Proceedings of the 2002 International System Dynamics Conference.* Palermo, Italy, forthcoming.

Graham, Alan K. and Carlos A. Ariza 2001. Dynamic, hard and strategic questions: Using Optimization to Answer a Marketing Resource Allocation Question. *Proceedings of the International System Dynamics Conference.* Atlanta, Georgia.

Graham, Alan K. and Louis Edward Alfeld 1976. *Introduction to Urban Dynamics*. Waltham, Mass.: Pegasus Communications.

Graham, Alan K., Carol Y. Choi and Thomas W. Mullen 2002. Using Fit-Constrained Monte Carlo Trials to Quantify Confidence in Simulation Model Outcomes. *Proceedings of the 2002 Hawaii Conference on Complex Systems* (HICCS).

Hammersley, J. M. and D. C. Handscomb 1964. *Monte Carlo Methods*. Chapman and Hall.

Hastings, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika* **57**, 97-109.

Luehrman, T. A. 1998a. Investment Opportunities as Real Options: Getting Started on the Numbers. *Harvard Business Review* July-August 1998, 51-67.

Luehrman, T. A. 1998b. Strategy as a Portfolio of Real Options. *Harvard Business Review*, September-October 1998, 89-99.

Lyneis, James M. 2000. System Dynamics for Market Forecasting and Structural Analysis. *System Dynamics Review* **16**(1), 3-25.

Lyneis, James M. and Kimberly Sklar Reichelt 1996. Calibration Standards. Cambridge, Mass.: *PA Consulting Group Internal Document*. April 11, 1996.

Lyneis, James M., Kimberly Sklar Reichelt, and Carl G. Bespolka 1996. Calibration Statistics for Life-Cycle Models. *Proceedings of the 1996 System Dynamics Conference*.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller 1953. Equations of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, **21**, 1087-1091.

Miller, J. H. 1997. Active Nonlinear Tests (ANTs) of Complex Simulation Models. *Management Science* **44**(6), 820-830.

Morgan, B. S. Jr. 1966. Sensitivity Analysis and Synthesis of Multivariable Systems. *IEEE Transactions on Automatic Control* **AC11**, 506-512.

Peterson, David W. 1980. Statistical Tools for System Dynamics. In (Randers 1980).

Phillips, W. G. B. 1980. Monte Carlo Tests of Conclusion Robustness. In (Randers 1980)

Randers Jorgen, ed. 1980. *Elements of the System Dynamics Method*. Cambridge, Mass.: MIT Press.

Richardson, George P. and Alexander L. Pugh III 1981. *Introduction to System Dynamics Modeling using Dynamo*. Waltham, Mass.: Pegasus Communications.

Rohrer, R. A. and M. Sobral, Jr. 1965. Sensitivity Considerations in Optimal System Design. *IEEE Transactions on Automatic Control* **AC10**, 43-48.

Rubinstein, R.Y. 1981. *Simulation and the Monte Carlo Method.* Engelwood Cliffs, NJ: John Wiley and Sons.

Schweppe, F.C. 1993. *Uncertain Dynamic Systems*. Engelwood Cliffs, NJ: John Wiley & Sons.

Sterman, John D. 2000. *Business Dynamics: Systems Thinking for a Complex World.* Irwin/McGraw-Hill.

Theil, Henri 1971. *Principles of Econometrics.* New York: John Wiley & Sons, Section 3.6.

Vermuelen, P. J. and D.C.J. de Jongh 1977. Dynamics of Growth in a Finite World: A Comprehensive Sensitivity Analysis. *Automatica* **3**(1), 77-84.

Wong, Cecelia S. Y. 1980. Criterion Sensitivity Analysis: A New Approach to Parameter Sensitivity. *Applied Mathematical Modeling* **4**(1), 7-15.

Back to the Top