Automated vs. 'Hand' Calibration of System Dynamics Models
An Experiment With a Simple Project Model

James M. Lyneis and Alexander L. Pugh, III
Pugh-Roberts Associates
41 William Linskey Way
Cambridge, MA 02142

Summary

Introduction

A necessary part of any system dynamics analysis is the estimation of parameter values which best correspond to the real system. The method of estimating parameters for a typical system dynamics model usually involves two steps (with potentially multiple iterations in the second!):

1. Make an *a priori* estimate based on direct observation, educated guesses by managers, or similar parameters in other settings (Graham [1980] refers to this as data "at or below the level of aggregation); and
2. Revise those estimates in the process of calibrating the model to aggregate data.

The calibration of the model (i.e., the adjustment of parameters in order to improve the correspondence of simulated and data) is typically done "by hand." The entire parameter estimation process therefore relies on the expertise and experience of the modeler.

The "hand" method of parameter estimation has been criticized for a number of reasons:

1. Given the complexity of high-order, multi-loop, non-linear feedback systems, parameter estimation and calibration are oftentimes very difficult; the process relies considerably on the experience and intuition of the modeler -- it appears to be "art" rather than "science";
2. The process and results are not replicable -- different modelers are likely to come up with different parameterizations (this is particularly true where the calibration process leads to the modification or addition of model structure);
3. The modeler cannot be certain that the final calibration is the best that can be achieved (i.e., that it is not a local optimum); and
4. Hand-calibration makes the generation of sensitivity analyses and confidence bands more cumbersome and less robust (in those cases where the procedures require re-calibrating the model to the revised set of parameters).

A number of statistically-based, automated parameter estimation/calibration approaches are common in other disciplines: econometrics (e.g., ordinary least squares, nonlinear least squares); engineering-based full-information maximum-likelihood estimation (and its derivatives); and nonlinear optimization algorithms. In addition to offering to overcome some or all of the

above shortcomings, automated approaches offer an additional benefit: the use of automated techniques may allow those with little experience to set up and calibrate a model. This would be most appropriate for "packaged" models, that is, models with generic applicability such as project models which have been set up to allow easy adaptation to new situations. An inexperienced modeler should be able to set up a generic project model for a specific project, and have the software automatically calibrate the model to available data.

This paper discusses a series of experiments in which the same model is first calibrated by several professionals, and then by nonlinear optimization software. The model was a relatively simple model of the design phase of a design and build project, based on the rework cycle [Cooper, 1994], with dynamic productivity and quality driven by four effects: experience of staff, prior work quality, organizational size, and customer disruption. In total, the model contains 10 levels and 70 active equations. Synthetic "data" were produced by taking simulated output from the calibrated model and applying noise to staffing, initial issues, drawing revisions, and scheduled completion date to produce four "hard" data streams, and bias to the effects on productivity and quality to produce "management observations" for eight effects. Consequently, the structure of the model and the system are identical.[1] This removes one source of error in estimating parameters.

Results of the Experiments -- "Hand" v. Automated Calibration

Hand-Calibration

The "hand" experiment was conducted as a part of internal staff development at Pugh-Roberts in which an experienced modeler worked with a more junior staff person. Five teams took part in the exercise. They had about one and one-half hours to calibrate the model. Table 1 provides summary calibration statistics for the five teams.[2] Given the limited time available, and the fact that the primary purpose of the exercise was training, the hand-calibrators did remarkably well. The best three teams came within 10-12% of the hard data (given the noise in the data, the best possible fit is around 6%). They were further off for the observations because fitting this data typically occurs later in the calibration process, and time ran out.

Table 1 Calibration Statistics for Hand-Calibration Teams

	Initial	Team 1	Team 2	Team 3	Team 4	Team 5
Hard Data	26.08	10.10	12.15	12.88	19.18	18.97
Observations	48.79	12.26	17.22	24.28	42.74	23.48
Aggregate	31.76	10.64	13.42	15.73	25.07	20.10

[1] In practice, discrepancies between real life and simulated life can occur for three reasons: (1) there are errors in measuring the values of "real life" (e.g., because of sampling or accounting errors, or because of inconsistencies in which groups of people charge to a project); (2) the model does not correspond exactly to the real system (e.g., attrition rates may be represented as a constant fraction, when in fact they vary); and (3) there are differences between the inputs which affect real life and simulated life (e.g., variations in the impact of weather on productivity).

[2] Statistics given are weighted values of mean absolute percent error statistics to the "hard" and "soft" data streams [see Reichelt, et. al 1996 for further discussion.]

Some additional observations:

1. Experienced calibrators followed very similar paths in calibrating the model. While a certain element of "art" will always remain, hand-calibration follows a logical, predictable, and transferable sequence.
2. The parameters developed by the hand calibrators were remarkably close to each other. The coefficient of variation for the key parameters was 10-14% (the automatic calibration software was only marginally more consistent, ranging from 7-12%). Hand-calibration is replicable.
3. Experienced calibrators achieved relatively consistent results quickly. In the time allotted, the calibrators performed between 13 and 25 simulations. Another hour or so of work would have produced very polished results.

Automated Calibration

In the "automated" component of the experiment, nonlinear optimization algorithms were applied to the same model and starting point.[3] Numerous experiments were conducted, using different error calculations and components, different starting conditions, and different calibration "handles." In addition, various tests were conducted using the optimization algorithms to "polish" the hand calibrations, and following an "expert system" approach. The bottom line is that, with the right conditions, automated calibration can work. For example, from three different starting conditions, the DYNAMO software produced the improvement in aggregate fit statistics given in Table 2.

Table 2 Fit Statistics Before and After for Different Starting Conditions

	Start 1		Start 2		Start 3	
	Before	After	Before	After	Before	After
Hard Data	26.084	5.857	52.174	5.355	40.157	5.388
Observations	48.789	10.927	95.394	10.294	119.862	11.503
Aggregate	31.76	7.125	62.979	6.59	60.083	6.917

However, there are a number conditions necessary for the software to work:

1. *The choice of error calculation is critical.* We tested three different statistics: modified mean absolute percent error (MAPE); modified root mean square percent error (RMSPE), and average absolute error as a percentage of the mean of the data (AAE) (see Reichelt, et. al [1996]). AAE was the only statistic for which the software was able to reliably calibrate the model. For the other statistics, the software reached a local optimum far from the true optimum, and was not able to break out.

[3] The automated calibration package incorporated in Professional DYNAMO (developed by Alexander L. Pugh, III) was used for these experiments. Additional experiments were conducted with the software incorporated into VENSIM (developed by Robert Eberlein).

2. *The software must be given an error function which includes all of the data.* In order to reliably calibrate the model, the error function must include every data stream for which there is hard or qualitative data. While a hand-calibrator can mentally include and exclude specific data streams as the calibration progresses, this is not possible for automated calibration (unless an "expert" system approach is adopted). Unfortunately, the more components included, the less important any one becomes, and the more compromises/tradeoffs there are. While the AAE statistic did not have a problem with this, it may help to explain while MAPE and RMSPE had a hard time breaking out of local optima.
3. *The software must be given a complete set of tuning handles, and cannot compensate for or detect incorrect/incomplete structure.* Again, a hand calibrator can focus in on different elements of behavior, and the calibration handles pertinent to those behaviors, as the calibration progresses. The software, however, must be given all the possible handles at the beginning. Unfortunately, including all possible handles dramatically increases the number of simulations required to calibrate the model (the DYNAMO software took 900-1000 simulations to calibrate this simple model). Without a complete set of calibration handles, in effect the model structure is incomplete. The software attempts to tune the model anyway, but cannot do so accurately. Only visual inspection can determine the nature and cause of the problem. Clearly human interaction is required.

Conclusions and Future Research

Hand calibration works, and is less of an art and more replicable than might be expected. Moreover, it produces results which are as close to the true values as automated calibration, and are typically close enough to make no significant difference to the outcome of policy interventions.

However, automated calibration offers great promise. While it has proven, with the proper error function and sufficient parameter "handles," are able to calibrate a model from scratch, the computational requirements may prove prohibitive for realistic models until computer power increases further. Therefore, in the short-term it can most effectively be used to "polish" the efforts of hand-calibration, and for recalibration during would-have sensitivity analysis. In the intermediate term, an expert system approach offers to reduce the computational requirements and thereby speed the overall calibration process.

The combination of the best parameter estimation practices from system dynamics (i.e., use of a priori estimation and qualitative information), with statistical approaches, offers to improve the efficiency and reliability of system dynamics practice. The next steps in this research are to begin applying the software on larger models, and to address the tolerance and expert system issues.

References

(please refer to complete paper)