Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

COUPLING OPTIMIZATION AND STATISTICAL ANALYSIS WITH SIMULATION MODELS

Benjamin Thengvall
Fred Glover
Dave Davino

OptTek Systems, Inc.
2241 17th Street
Boulder, CO 80302, USA

ABSTRACT

Simulation optimization has become commonplace in commercial simulation tools, but automated statistical analysis of the impacts of varying input parameters is much less common. In this paper we explore how both optimization and statistical analysis can be coupled with simulation models to provide key insights for decision makers. A manufacturing example is provided to illustrate the results of multi-objective optimization and post-optimization statistical analysis of the simulation runs. We demonstrate how automated statistical analysis can provide analysts with valuable information on variable sensitivities and good and bad regions of the decision trade space.

1 INTRODUCTION

Most organizations fail to take full advantage of their simulation models. Even though large amounts of time and money are invested creating a simulation tool and populating it with validated data, a large part of the valuable knowledge that the model may yield is generally overlooked.

Simulation analysts who can access such knowledge are exceedingly valuable to their organization and become highly sought-after resources. Combining optimization and statistical analysis techniques with a simulation model is the key to unlocking this knowledge. Optimization techniques can be used to execute a simulation model many times, varying the input parameter values, to determine the best input values to achieve desired system outputs. The results of these simulation runs can then be explored with statistical techniques to better understand the system modeled by the simulation. Essential optimization and analysis questions that can be answered for simulation models by combining these techniques include:

- Optimization
  - What combinations of input parameters lead to the best and worst performance of the system?
  - What are the best tradeoffs between multiple competing objectives?
- Analysis
  - Which input parameters have the greatest influence on the system being modeled and which have the least?
  - Are there good or bad regions of the input parameter space that can be defined by a subset of input parameters with restricted ranges?
  - Are some areas of the parameter trade space more robust to parameter variation than others?

To derive the greatest benefit from a simulation model, an analyst should apply both optimization and the statistical analysis techniques. Combining these techniques can provide answers to these essential questions and key insights for decision makers. More importantly, they increase the organizational return on investment from simulation studies. Answering these types of questions additionally provides model validation and builds stakeholder confidence.

2 OPTIMIZATION AND STATISTICAL ANALYSIS IN COMMERCIAL SIMULATION PACKAGES

Over the past two decades optimization tools in commercial simulation packages have become widespread and are relatively easy to use, even if not all practitioners exploit them. Commercial simulation packages also have analysis tools that explore the variability uncovered through simulation replications (or Monte Carlo runs) for a single set of input parameters. However, the analysis of all simulation runs resulting from an optimization run is less commonly available, at least in an automated, easy to digest way.

The underlying statistical techniques discussed in this paper are not new. However, in many tools today, to perform variable sensitivity and good and bad region analysis across simulation runs executed with different combinations of input parameter changes, analysts have to use multiple tools, or perform the simulations and then piece together the results of various statistical techniques. Thus, these types of valuable simulation analyses are done infrequently, and are often performed only by technical consultants and advanced users. To perform these analyses, users of discrete event simulation packages export their simulation results and then use specialized statistical tools like JMP, SPSS, or R for analysis. Users of spreadsheet-based Monte Carlo simulations have more statistical analysis tools at their disposal, but even for these users, gaining insights across all simulation runs is not an automated process.

The critical goals of identifying good and bad regions of a parameter trade space, and of discovering robust solutions, are sometimes pursued by more advanced analysts through generation of a response surface approximation by coupling design of experiments with simulation. This approximate response surface is then explored through various stochastic optimization techniques (Samuelson 2011). Such an approach generally relies on moving from tool to tool for the different steps in the process: generating the design of experiments, executing the simulations, and performing the stochastic optimization. This type of process has the conspicuous shortcoming of frequently oversimplifying complex response surfaces, which can entail a costly loss of valuable insights.

If these statistical analyses across simulation runs are integrated into commercial simulation products and automated, then key results can be presented as a direct outcome of simulation optimization runs, providing more information and allowing more robust analysis by all simulation users. It is our belief that just as simulation optimization has become commonplace, these types of automated analyses across different input parameter sets following optimization runs will also become a standard feature of commercial simulation packages in coming years. In the remainder of the paper a concrete example will be provided and results presented to show the benefits of integrated optimization and analysis with a simulation model.

3 AN ILLUSTRATIVE EXAMPLE

To illustrate emerging simulation analysis features to overcome the limitations in many current systems, we will look at a simulation manufacturing example and show tables and graphs produced by SimWrapper™. This tool embodies technologies which we expect will become standard in commercial simulation packages over the next decade, just as simulation optimization features have been added to most packages in the past 15 years.

SimWrapper does not provide simulation capabilities itself. It wraps (integrates with) simulation models to provide complementary optimization and analysis features. SimWrapper embeds OptQuest®, a state-of-the-art black box optimizer incorporated in many commercial simulation tools, and OptAnalysis™, a library of statistical and data mining techniques. Among its varied features, it allows users to create binary, integer, continuous, enumerated, choice, and location decision variables from simulation input parameters. Users also specify outputs of the simulation model to be used in constraints and objectives to optimize.

Following a batch run, design of experiments run, or optimization run, automated simulation analysis results are displayed for the user in a series of tables and two- and three-dimensional graphs that summarize and elucidate the results.

Although the remainder of the paper follows a manufacturing example, the optimization and analysis techniques discussed are equally applicable in other domains. The optimization techniques treat the model as a "black box", requiring only that each set of input values yields corresponding output values. Likewise, the statistical techniques presented require only rows of data containing dependent and independent values, which is what the input and output results from each run of a simulation model are. This pairing of optimization and statistical analysis techniques with simulation models can be used beneficially anytime a complex system is being modeled and many combinations of input parameters are being considered.

We will consider an example of an organization that has a simulation model of its manufacturing environment which includes the key input parameters and performance metrics shown in Figure 1.

Simulation Inputs Parameters to Change	Performance Metrics
Number of production lines	Throughput
Machine configurations and performance	Work in progress
Job scheduling policies	Inventory
Batch size (or bounds on size)	Fixed and variable cost
Queue lengths (or bounds on lengths)	Utilization
Staffing – shift lengths and start times, overtime	Job tardiness

Figure 1: Simulation parameters and key performance metrics for a manufacturing example.

Managers often do not know the breadth of information that can be garnered from a simulation model and consequently are prone to ask questions that are too narrow. For instance, a manager hoping to increase throughput may ask a simulation analyst to run a small parametric study and report results from executing changes relative to the current batch size of -20%, -15%, -10%, -5%, 0%, +5%, +10%, +15%, and +20%. Suppose, however, the key drivers of throughput in this manufacturing environment are job scheduling policies and staffing. The manager may tweak the batch sizes based on the simulation runs, but will fail to discover the most influential throughput improvements because the question posed was not broad enough, and the analyst simply ran the requested simulations. The routine parametric study of batch sizes gives no chance to uncover the important influence of scheduling and staffing policies.

Decision Variables	Type	Values
Number of production lines	Integer	Min: 1, Max: 5
Machine configuration	Choice	C1, C2, C3, or C4
Job scheduling policies	Choice	P1, P2, or P3
Batch size	Integer, Stepped	Min: 10, Max: 50, Step: 5
Queue lengths	Integer, Stepped	Min: 5, Max: 25, Step: 5
Number of shifts	Integer	Min: 1, Max: 3
Voluntary overtime	Boolean	True, False

Figure 2: Decision variables defined for a manufacturing example.

To expand the example, consider the situation where our manufacturing firm has experienced a significant increase in orders and is struggling with job tardiness. An analyst has been asked to explore options for reducing job tardiness and their cost ramifications. The analyst identifies seven simulation inputs that may be significant and defines potential ranges on their values as shown in Figure 2.

4 SIMULATION OPTIMIZATION

The next step for an analyst is to define one or more objectives and perform an optimization of the simulation model. In this example, one approach would be to minimize job tardiness at a fixed cost level. However, more information can be gained via simulation optimization by solving for two objectives simultaneously, minimizing job tardiness and minimizing fixed and variable costs, and then exploring the non-dominated efficient frontier of possibilities.

Our analyst runs a simulation optimization solving for both objectives simultaneously and obtains the graph shown in Figure 3 summarizing the results of 144 simulation runs. The dark blue line is the non-dominated efficient frontier showing the optimal tradeoffs between cost and tardiness. The grey points are other dominated solutions, and the lighter blue lines highlight the best efficient frontiers found with fewer simulation trials.

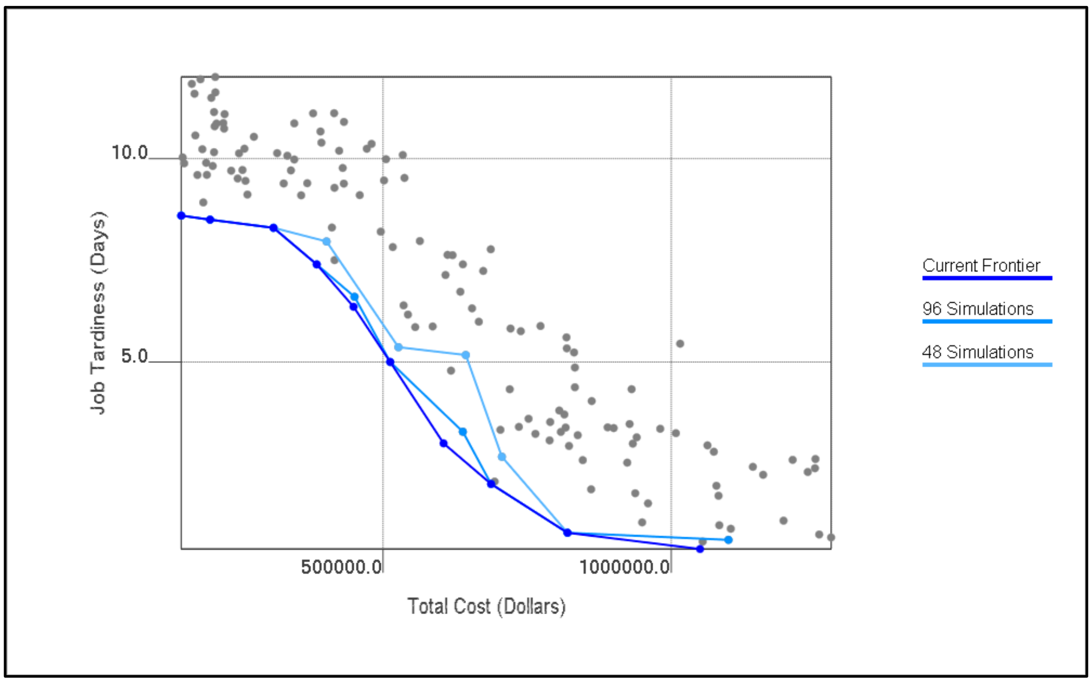

Figure 3: Plot of all simulation runs highlighting the non-dominated efficient frontier.

In this example, configurations that increase costs at the lower end yield minimal benefits in reducing job tardiness. Then at a point lying roughly between the costs of $310,000 and $685,000, there is a significant decrease in job tardiness per unit cost with tardiness dropping from 8.3 days to 2 days. The slope then flattens again, and further gains in job tardiness are increasingly expensive. Using these already executed 144 simulation runs, the analyst can also answer key simulation analysis questions.

5 SIMULATION ANALYSIS

Before sharing these results with supervisors the analyst would like to understand which simulation inputs have the biggest impact on the objectives, along with the changes that must take place in the manufacturing environment to achieve the solutions found at different points along the curve. To see how variations in the simulation inputs impact the total cost and job tardiness the analyst can examine the sensitivities of these variables to the different objective values. SimWrapper includes several routines which automatically identify influential variables and quantify how each relates to each objective of interest.

5.1 Variable Sensitivities

Figures 4 and 5 show the variable sensitivity summaries for the job tardiness objective and the total cost objective, respectively. The tables show results from least-squares regression analysis, variable effects analysis, mutual information, and regression tree analysis as well as an overall influence rank for each variable. The composite influence rank is an aggregation of the individual analysis results and ranges from 0 to 100. The larger the influence rank, the more impact the variable has on the objective.

	Influence Rank	R² Value	MI	Variable Effects	Tree Level
	Job Tardiness (Days) ▼	Job Tardiness (Days)	Job Tardiness (Days)	Job Tardiness (Days)	Job Tardiness (Days)
Number of Production Lines	63	0.76	42.70	-10.11	1
Number of Shifts	50	0.52	35.68	-7.96	
Machine Configuration	29		18.48		3
Job Scheduling Policy	25		17.88		
Queue Length	22	0.02	41.03	-0.76	3
Batch Size	7	0.02	5.83	-1.44	
Voluntary Overtime	5	0.02	2.72	-0.95	

Figure 4: Variable sensitivities table for the job tardiness objective.

	Influence Rank	R² Value	MI	Variable Effects	Tree Level
	Total Cost (Dollars) ▼	Total Cost (Dollars)	Total Cost (Dollars)	Total Cost (Dollars)	Total Cost (Dollars)
Number of Production Lines	72	0.88	60.80	1002095.06	1
Number of Shifts	51	0.55	38.65	757378.87	
Machine Configuration	26		16.93		3
Job Scheduling Policy	20		14.06		
Queue Length	16	0.00	31.86	11250.78	
Voluntary Overtime	13	0.10	8.65	200234.04	4
Batch Size	6	0.02	3.35	145648.35	

Figure 5: Variable sensitivities table for the total cost objective.

5.1.1 Least squares regression

A linear regression finds a linear fit between an output or objective and one (or more) decision variables (Neter, Wasserman, and Kutner 1990). Regression results include an R^2 value which ranges from 0 to 1 and measures the goodness of a linear fit, with values closer to 1 indicating a stronger linear relationship. In our manufacturing example, the R^2 score for number of production lines and the total cost is 0.88, indicating a strong linear relationship. Adding production lines is a significant capital investment.

5.1.2 Variable effects analysis

The variable effects score shown is based on the idea of isolating the "main effects" of a variable (Montgomery 1997). These effects provide a measure of how much a model output of interest changes as the variable value moves from its lower bound to its upper bound within its range. High absolute value indicates a high level of influence on the objective value with the sign indicating the direction of influence (e.g., with a negative sign higher variable values result in lower objective values) and the magnitude of the score provides the approximate amount of change in the objective as the variable value moves from its lower to upper bound. Looking at the variable effects column of Figure 4 we can see that increasing the number of shifts from 1 to 3 reduces the job tardiness by almost 8 days on average.

5.1.3 Mutual information

The mutual information value for a variable represents how much the variable and objective "move together." It's a measure of correlation that can pick up nonlinear dependencies (Cover and Thomas 1991). In Figures 4 and 5 a normalized mutual information value is reported that ranges from 0 to 100, with higher values indicating stronger relationships. The mutual information score complements the regression and variable effects scores by identifying non-linear relationships. Scanning the R^2 and mutual information columns in Figure 4 the queue length variable has an R^2 score of only 0.02, but a mutual information score of 41.03. From Figure 6 we see that queue length and job tardiness have a strong nonlinear relationship.

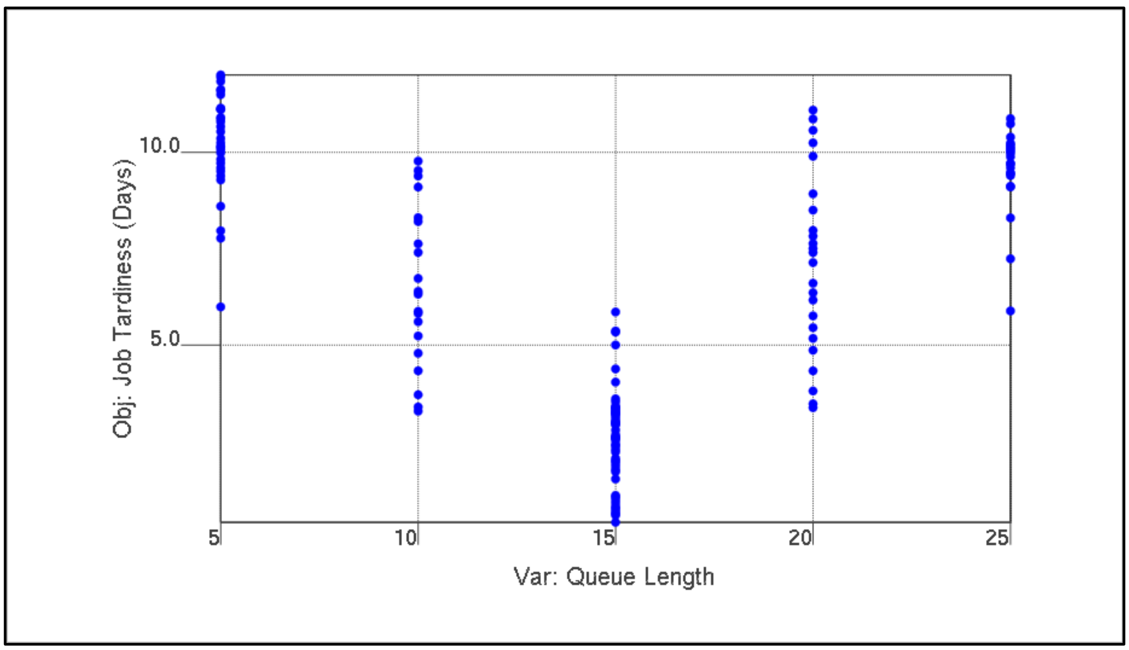

Figure 6: A plot showing the relationship between queue length and job tardiness.

5.1.4 Regression tree analysis

Regression tree analysis (Breiman et al. 1984) provides the tree level in the variable sensitivity tables. A regression tree segments the decision space, identifying combinations of variable value ranges that yield good and bad simulation run outcomes based on an objective. The root node represents all simulation runs, and then successive branches are created with child nodes containing subsets of the simulation runs with decision variable values in different ranges.

A variable branched on closer to the root indicates higher influence. Regression trees are another method that can identify nonlinear variable influence. Figure 7 shows a regression tree for the job tardiness objective. The tree splits multiple times on number of production lines and then splits on machine configuration or queue length depending on the node. The two leaf nodes at the lower left show that for four or five production lines, using machine configuration C2 rather than configuration C1, C3, or C4 yields average job tardiness of 2.2 days versus 3.9 days.

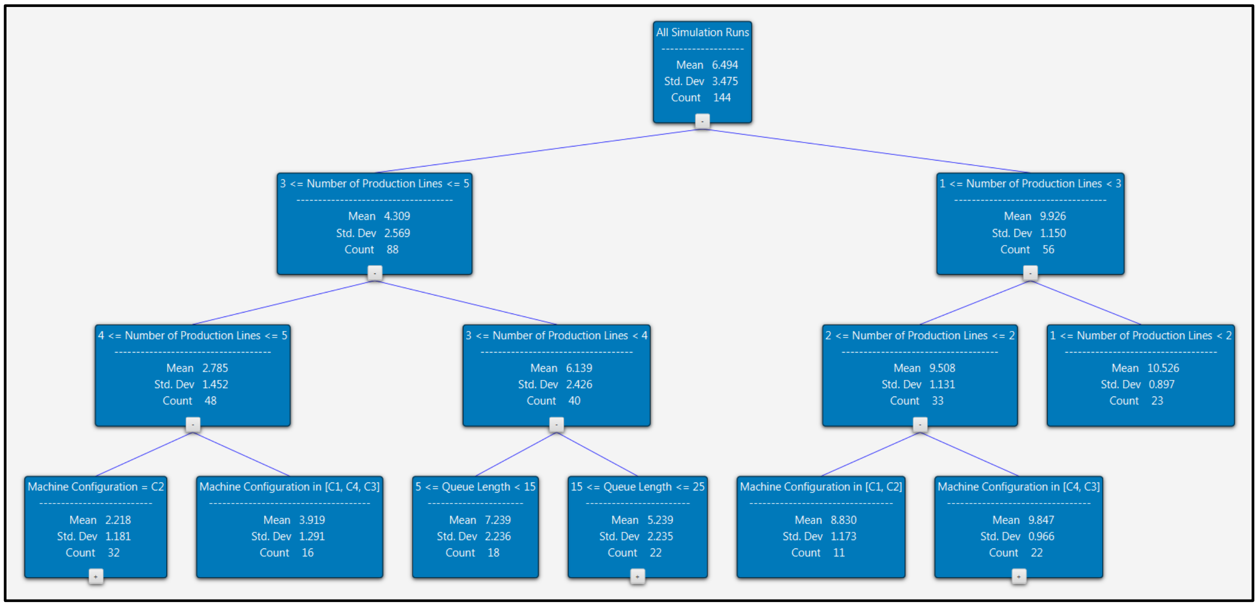

Figure 7: A regression tree for the job tardiness objective.

5.2 Good and Bad Regions

Identifying good and bad regions of the decision variable trade space is also exceedingly useful for analysts. Combinations of decision variable ranges define subsets of simulation runs (regions) that are associated with either high or low objective values. Statistical measures for a region can be provided as one measurement of the robustness of results found in a region. These good and bad regions are an important complement to influential variable analysis. They highlight for an analyst how variables interact in combination in specific ranges to influence an objective of interest. Sometimes the optimal set of input values represents an outlier surrounded by solutions with much worse objective value. Decision space analysis allows the analyst to avoid these solutions, and to select robust solutions with confidence.

Figure 8 shows two good regions for the job tardiness objective, the first defined just by having four or five production lines, and the second by having five production lines and using machine configuration C2. The second good region identified has a smaller standard deviation than the first, indicating less variability across the simulation runs that share these characteristics. One bad region is identified defined by one production line and the use of job scheduling policy P1 or P2.

All Simulation Runs

Simulation Run Statistics	Count	Job Scheduling Policy	Machine Configuration	Number of Production Lines	Queue Length	Analysis Type
Mean: 6.49 Std: 3.47 Min: 0.40 Max: 12.01	144 (100%)	P1 P3 P2	C4 C3 C1 C2	Min: 1.00 Max: 5.00	Min: 5.00 Max: 25.00	

Good Regions

Simulation Run Statistics	Count	Job Scheduling Policy	Machine Configuration	Number of Production Lines	Queue Length	Analysis Type
Mean: 2.79 Std: 1.45 Min: 0.40 Max: 5.88	48 (33%)			Min: 4.00 Max: 5.00	Min: 10.00 Max: 25.00	PRIM Analysis
Mean: 1.62 Std: 1.09 Min: 0.40 Max: 4.33	16 (11%)		C2	Min: 5.00 Max: 5.00		Tree Analysis

Bad Regions

Simulation Run Statistics	Count	Job Scheduling Policy	Machine Configuration	Number of Production Lines	Queue Length	Analysis Type
Mean: 10.82 Std: 0.79 Min: 9.82 Max: 12.01	17 (11%)	P1 P2		Min: 1.00 Max: 1.00		PRIM Analysis

Figure 8: Good and bad regions for the job tardiness objective.

The results shown in Figure 8 are derived from regression tree analysis and Patient Rule Induction Method (PRIM) analysis. PRIM is a machine learning method (Friedman and Fisher 1999) which is specifically designed to identify good and bad regions of the decision space. Both the regression tree and the PRIM analysis result in multiple good regions, but In Figure 8 only the top region identified by each technique is displayed.

6 CONCLUSION

We have illustrated the benefits of coupling both optimization and statistical analysis with simulation models to provide key insights for decision makers. Utilizing optimization and statistical analysis together provides answers to the "what's the best?" questions of the optimization realm, while the complementary simulation trade space analysis provides answers to questions like "which inputs are influential?" and "how can the system be characterized?" which rely on statistical analysis of many executed simulation runs. Combining optimization and statistical analysis with simulation enables managers and analysts to delve into the crucial factors that influence the functioning of their organizations, and gain greater insights than those available through simulation alone.

REFERENCES

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. *Classification and Regression Trees*. New York: Wadsworth.

Cover, T. and J. Thomas. 1991. *Elements of Information Theory*. New York: John Wiley & Sons.

Friedman, J. and N. Fisher. 1999. “Bump Hunting in High-dimensional Data.” *Statistics and Computing* 9: 123-143.

Glover, F., J. P. Kelly and M. Laguna. 1996. “New Advances and Applications of Combining Simulation and Optimization.” In *Proceedings of the 1996 Winter Simulation Conference* edited by J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 144-152. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Montgomery, D. 1997. *Design and Analysis of Experiments*. New York: John Wiley & Sons.

Neter, J., W. Wasserman and M Kutner. 1990. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL:.Richard Irwin.

Samuelson, D. 2010. “When Close is Better than Optimal.” *ORMS-Today* 37(6):144-152.

AUTHOR BIOGRAPHIES

BENJAMIN THENGVALL is Vice President of Government Solutions at OptTek Systems, Inc. He earned his M.S.E. and Ph.D. in Operations Research and Industrial Engineering from the University of Texas at Austin. Prior to joining OptTek, he was an Operations Research Scientist at Navitaire, Inc., a leading operations research technology firm serving the airline industry. His primary interests have been in mathematical modeling, real-time optimization, transportation and scheduling, agent-based and discrete-event simulation, and simulation optimization and analysis. He has taught at the Peter Kiewit Institute at the University of Nebraska and at the McCombs School of Business at the University of Texas. His email is thengvall@opttek.com.

FRED GLOVER is the Chief Technology Officer in charge of algorithmic design and strategic planning initiatives for OptTek Systems, Inc. He also holds the title of Distinguished Professor of the University of Colorado System. Dr. Glover is a leading figure in the field of metaheuristics, whose name he coined in the 1980s – an area that is now the subject of numerous books and international conferences, focusing on the development of models and methods enabling the solution of complex Thengvall and Glover nonlinear and combinatorial problems that lie beyond the ability of classical optimization procedures. He has authored or co-authored more than four hundred published articles and eight books in the fields of mathematical optimization, computer science and artificial intelligence, with particular emphasis on practical applications in industry and government. Dr. Glover is the recipient of the distinguished von Neumann Theory Prize, an elected member of the National Academy of Engineering, and has received numerous other awards and honorary fellowships, including those from the American Association for the Advancement of Science (AAAS), the NATO Division of Scientific Affairs, the Institute of Operations Research and Management Science (INFORMS), the Decision Sciences Institute (DSI), the U.S. Defense Communications Agency (DCA), the Energy Research Institute (ERI), the American Assembly of Collegiate Schools of Business (AACSB), Alpha Iota Delta, and the Miller Institute for Basic Research in Science. His email is glover@opttek.com.

DAVID DAVINO is Vice President of Client Relations and is the primary interface for all OptTek commercial clients in providing excellent customer care. Dave has 25+ years of account management and client support experience in the high-tech arena. Prior to OptTek, Dave managed strategic accounts for Deutsche Telekom North America and provided business, marketing, and operational support for a variety of software development teams at Bellcore/Telcordia. Dave holds an MBA in Finance from Seton Hall University as well as a BA in Economics from Bucknell University. His email address is davino@opttek.com.