Proceedings of the 1995 Winter Simulation Conference
ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman

AMPLITUDE SELECTION IN TRANSIENT SENSITIVITY ANALYSIS

Douglas J. Morrice

MSIS Department
The University of Texas at Austin
Austin, Texas 78712-1175

Sheldon H. Jacobson

Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061-0118

ABSTRACT

Most research in computer simulation sensitivity analysis can be classified as steady state sensitivity analysis. This paper considers an approach to *transient sensitivity analysis*, i.e., an approach to study the transient behavior of a performance measure in response to changes in one or more input factors. We examine the issue of factor magnitude change (or *factor amplitude*) selection and how the choice of factor amplitudes impacts the overall analysis. We also consider a control variate variance reduction scheme designed to improve performance when small amplitude changes are used. A tandem queue example illustrates the results.

1 INTRODUCTION

Sensitivity analysis is an important area of research in computer simulation of stochastic discrete event dynamic systems. Most of the research that has been done on this topic can be classified as steady state sensitivity analysis because it considers the long-run or steady state impact of input factor changes on an output performance measure. This paper considers an approach to sensitivity analysis introduced in Morrice and Gupta (1994) and developed in Morrice (1995). The approach analyzes the short-term or transient behavior of a performance measure resulting from changes in one or more input factors. Since the approach studies transient behavior, it is referred to as *transient sensitivity analysis*.

Morrice (1995) provides a methodology for changing factors simultaneously during a single set of simulation runs. Such simultaneous variation offers potential computational efficiencies relative to other methods that require multiple sets of simulation runs for transient sensitivity analysis information. However, simultaneous variation during a single run can also be seriously constrained by such things as the stability of the system (for example, traffic intensity being less than one) and the requirement for a factor to remain positive (for example, service rate in a queueing model). Constraints of this type have direct impact on the *factor amplitude*, i.e., the magnitude by which factors can be changed.

This paper focuses on the selection of factor amplitudes for the method proposed by Morrice (1995). In addition, a control variate variance reduction scheme developed by Jacobson (1993) is considered for cases when small amplitudes are selected. The remainder of the paper is organized in the following manner. Section 2 contains model assumptions and necessary background material. Section 3 illustrates the amplitude selection problem and discusses issues related to this problem. Section 4 provides an example and section 5 contains concluding remarks.

2 MODEL ASSUMPTIONS AND BACKGROUND

Using the modeling assumptions of Morrice (1995), consider a simulation model with continuous input factors, $X_k(t)$, $k = 1, \ldots, K$, and a scalar output response $Y(t)$ for $t = 0, \ldots, N-1$. The quantity t is an observation index. Examples include the simulation time clock and a job (or customer) sequence number (Hazra, Morrice, and Park 1995). The sample size is represented by N. Although the exact functional relationship between the $X_k(t)$ and $Y(t)$ is unknown, we assume that this relationship is approximately described by the following metamodel:

$$Y(t) = \sum_{k=1}^{K} \sum_{r=0}^{q_k} h_k(r) X_k(t-r) + \varepsilon(t) \qquad (1)$$

The quantity q_k is a positive integer that represents the lag length in the k-th memory filter. The quantity $h_k(r)$ is the r-th coefficient (unknown) in the k-th

memory filter. It satisfies the property,

$$\sum_{r=0}^{\infty} |h_k(r)| < \infty$$

The term $\{\epsilon(t)\}$ is included to model uncertainty. It is a zero-mean, covariance stationary, random process with autocovariance function $\gamma(|i-j|) = \text{Cov}(\varepsilon(i), \varepsilon(j))$.

The inner summation of unknown coefficients in (1) models the dynamic or transient relationship between $Y(t)$ and the $X_k(t)$. It is included to mimic behavior often found in, for example, queueing models of manufacturing systems: namely, the current value of the output is dependent upon the current and past values of the input factors. Transient sensitivity analysis is designed to estimate the transient relationship between $Y(t)$ and the $X_k(t)$, i.e., the $\{h_k(r)\}$.

In Morrice's procedure, the factors are varied according to

$$X_k(t) = x_k + a_k \sum_{l=1}^{n_k} \cos(2\pi\omega_{kl}t), \qquad (2)$$

for $k = 1, 2, \ldots, K$, within a run of the simulation model. The quantity x_k is a fixed value set at the beginning of the simulation run (*nominal level*), a_k is the oscillation amplitude (range over which the factor is changed) and $\omega_{kl} = v_{kl}/N$ for $v_{kl} \in \{1, \ldots, \lfloor N/2 \rfloor\}$ is the oscillation frequency (rate at which the factor is changed). The values for x_k, a_k, and ω_{kl} are selected by the user. The ω_{kl} must be chosen uniquely for each $X_k(t)$ because the impacts of factors on $Y(t)$ are distinguished by frequency.

Upon substituting (2) into (1), the metamodel relationship can be rewritten as

$$Y(t) = \beta_0 + \sum_{k=1}^{K}\sum_{r=0}^{q_k} \beta_k(r) \sum_{l=1}^{n_k} \cos(2\pi\omega_{kl}(t-r)) + \varepsilon(t). \qquad (3)$$

where

$$\beta_0 = \sum_{k=1}^{K}\sum_{r=0}^{q_k} h_k(r) x_k$$

and for $k = 1, 2, \ldots, K$

$$\beta_k(r) = a_k h_k(r). \qquad (4)$$

Employing a well-known trigonometric identity,

$$\begin{aligned}\cos(2\pi\omega_{kl}(t-r)) &= \cos(2\pi\omega_{kl}t)\cos(2\pi\omega_{kl}r) \\ &\quad + \sin(2\pi\omega_{kl}t)\sin(2\pi\omega_{kl}r),\end{aligned}$$

(3) can be rewritten as

$$\begin{aligned}Y(t) &= \beta_0 + \sum_{k=1}^{K}\sum_{l=1}^{n_k}[A_k(\omega_{kl})\cos(2\pi\omega_{kl}t) \\ &\quad + B_k(\omega_{kl})\sin(2\pi\omega_{kl}t)] + \varepsilon(t) \qquad (5)\end{aligned}$$

where

$$A_k(\omega_{kl}) = \sum_{r=0}^{q_k} \beta_k(r)\cos(2\pi\omega_{kl}r)$$

and

$$B_k(\omega_{kl}) = \sum_{r=0}^{q_k} \beta_k(r)\sin(2\pi\omega_{kl}r).$$

Morrice (1995) uses the following procedure to get an estimate of the $\{h_k(r)\}$:

1. Make M simulation runs of length N observations.
2. On each run vary $X_k(t)$, $k = 1, 2, \ldots, K$, according to (2).
3. Average the N observations across the M runs.
4. Transform the resulting N observations to the frequency domain.
5. Fit a polynomial regression model to (5) in the frequency domain using the weighted least squares procedure described in Morrice and Bardhan (1995). The regression model produces estimates of the functions $A(\omega_k)$ and $B(\omega_k)$.
6. Apply an inverse cosine and sine transform to $A(\omega_k)$ and $B(\omega_k)$ to get estimates of the $\{\beta_k(r)\}$.
7. Rescale the $\{\beta_k(r)\}$ to get the $\{h_k(r)\}$.

3 AMPLITUDE CONSIDERATIONS

Varying the factors simultaneously according to (2) can be challenging in practice due to practical constraints on the system being modeled. For example, if factor $X_k(t)$ is constrained to positive values, then x_k, a_k and the $\{\omega_{kl}\}$ must be chosen to satisfy this constraint. Another possible constraint on (2) is the stability of the system being modeled. For example, in a queueing system, arrival and service rates must yield a traffic intensity that is less than one for the system to be stable. The queueing literature provides some limited theoretical guidance on how to choose factor levels according to (2) depending on how t is chosen (Whitt 1991 and Morrice, Gajulapalli, and Tayur 1994). In general, simulation can be used to empirically verify the stability of the system.

The quantity x_k is usually chosen as the center point of the experimental region. Morrice (1995) provides some discussion on the selection of the $\{\omega_{kl}\}$, but additional research is required on this topic. This paper focuses on the issue of the selection of the $\{a_k\}$.

The selection of a_k is governed by conflicting objectives. Jacobson (1993) notes this for harmonic gradient estimates. On the one hand, the quantity a_k determines the strength of the signal associated with $X_k(t)$. A signal must be chosen strong enough to ensure that it is not masked by random noise in the simulation output if the $\{h_k(r)\}$ are nonzero. This requirement supports choosing a_k as large as possible. On the other hand, a_k cannot increase without bound due to the positivity and stability constraints mentioned above. In fact, such constraints can become more restrictive as additional cosinusoidal terms are added to (2). For example, as more terms are added to (2), a positive factor level can be ensured by reducing a_k. Having the flexibility to include additional terms in (2) is desirable because more observations of the $A_k(\omega_k)$ and $B_k(\omega_k)$ functions may be necessary for the regression analysis. This type of flexibility dictates choosing a_k as small as possible.

A key to satisfying both requirements is to reduce the variance in the noise process. With a reduced variance, a weaker signal (i.e., a signal with a reduced amplitude) becomes sufficient to overcome the noise masking problem. As a result, more flexibility is provided for adding more terms to (2). Morrice (1995) considers a crude form of variance reduction by averaging observations across independently seeded simulation replicates. The next section considers a control variate variance reduction scheme suggested by Jacobson (1993).

4 EXAMPLE

To illustrate the effect of choosing smaller amplitude values, an example from Morrice (1995) is used. Consider a three work station assembly line model where the interarrival times and the service times on all three work stations are exponential. Buffer capacity between each work station is assumed to be infinite and jobs are processed on a first-come, first-serve basis. The four factors in this model are the mean interarrival time, X_1, mean service time for the first station, X_2, mean service time for the second station, X_3, and mean service time for the third station, X_4. Only factors X_1 X_2, and X_4 will be changed to illustrate transient sensitivity analysis. The output response, Y, is the waiting time in the system and its expected value is the performance measure of interest.

The following three scenarios are used to illustrate transient sensitivity analysis:

1. The quantity X_1 is changed from 9 to 11, X_2 is fixed at 6, and X_4 is fixed at 6.

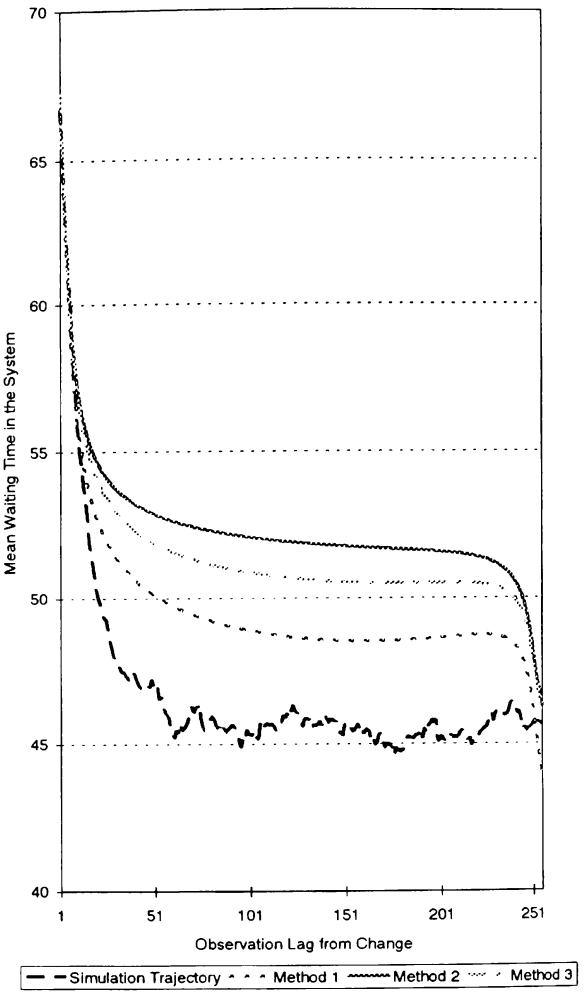

Figure 1: Mean Waiting Time Trajectory After Changing the Mean Interarrival Time

2. The quantity X_1 is fixed at 10, X_2 is changed from 5 to 7, and X_4 is fixed at 6.

3. The quantity X_1 is fixed at 10, X_2 is fixed at 6, and X_4 is changed from 5 to 7.

In all scenarios, X_3 is set at 7. All three scenarios are assumed to be in steady state before the changes are made. The transient behavior in $E[Y]$ is its readjustment to steady state after the changes are made.

Scenarios 1, 2, and 3 are depicted in figures 1, 2, and 3, respectively. The simulation trajectory in figure 1 is generated by averaging the system waiting time of 800 service completions over 5000 runs. The system is started empty and idle and the mean interarrival time is changed from 9 to 11 after the 525th arrival. The number 525 is used because it is determined, by inspection, that the initial transient in the mean waiting time has ended by customer ser-

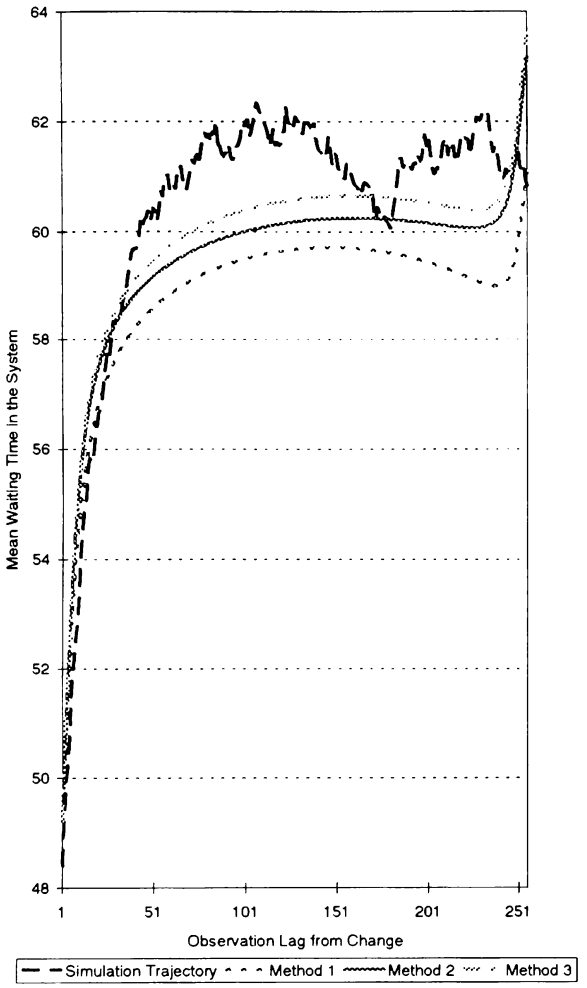

Figure 2: Mean Waiting Time Trajectory After Changing Mean Service Time on the First Station

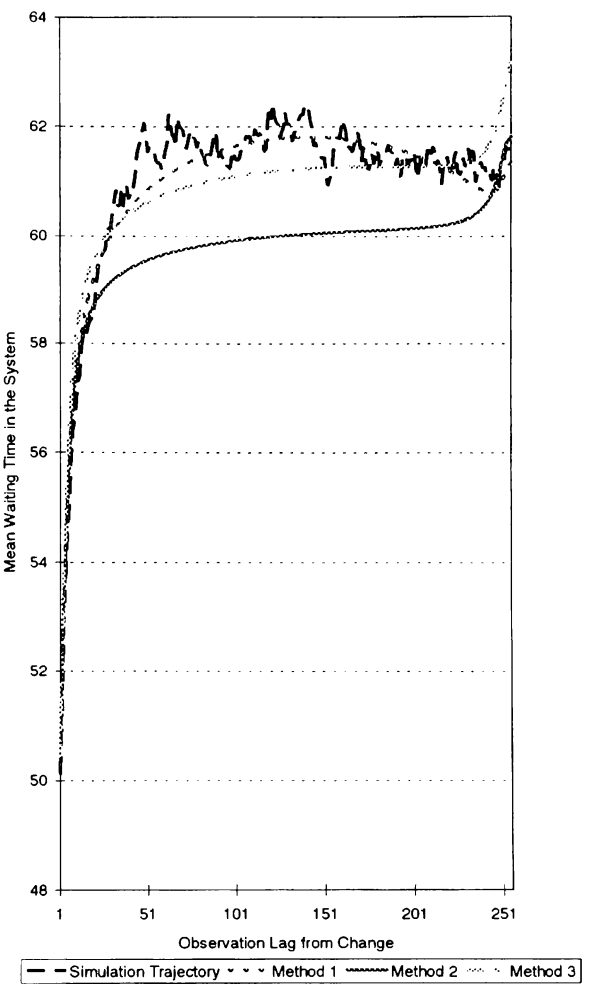

Figure 3: Mean Waiting Time Trajectory After Changing Mean Service Time on the Third Station

vice completion 500. The simulation trajectories in figures 2 and 3 are generated in a similar manner except that they are based on 5000 runs of 510 service completions and the factor change is made after the 250th service at the corresponding work station. As before, 250 is chosen, by inspection, to be a point beyond the transient period attributable to the initial conditions. For all scenarios, the simulation trajectories are used as benchmarks.

The results for Method 1 are from Morrice (1995). The results for this method in all three figures are generated by two sets of simulation runs. On the first set, the system waiting time from 4596 service completions are averaged across 500 runs. In order to mitigate the effects of an initial transient, each run is warmed up for 500 customer service completions with the factors set at the following nominal values: $X_1(t) = 10, X_2(t) = 6, X_3(t) = 7, X_4(t) = 6$. Then, the first 500 observations are discarded. The value 500 was chosen from the time domain simulation results for the simulation trajectories. After the 500-th service completion, the index t in (2) is reset to zero and for the remainder of each run the factors are varied according to:

$$X_1(t) = 10 + \sum_{\omega_1 \epsilon S_1} \cos(2\pi\omega_1 t),$$

$$X_2(t) = 6 + \sum_{\omega_2 \epsilon S_2} \cos(2\pi\omega_2 t),$$

and

$$X_4(t) = 6 + \sum_{\omega_4 \epsilon S_4} \cos(2\pi\omega_4 t).$$

where

$$S_1 = \{5, 505, 1005, 1505, 2005\},$$

$$S_2 = \{9, 509, 1009, 1509, 2009\},$$

$$S_4 = \{11, 511, 1011, 1511, 2011\},$$

and the elements of each frequency set are divided by 4096. Only five cosine terms are chosen for each parameter so that $X_2(t)$ and $X_4(t)$ are guaranteed to be positive. In order to obtain more observations for the $\{A_k(\omega_k)\}$ and $\{B_k(\omega_k)\}$, an additional set of runs is required. The second set of runs consists of another 500 independently seeded runs of 4596 service completions. At the 501-st observation, t is reset to zero and the factors are varied according to:

$$X_1(t) = 10 + \sum_{\omega_1' \epsilon S_1'} \cos(2\pi\omega_1' t),$$

$$X_2(t) = 6 + \sum_{\omega_2' \epsilon S_2'} \cos(2\pi\omega_2' t),$$

and

$$X_4(t) = 6 + \sum_{\omega_4' \epsilon S_4'} \cos(2\pi\omega_4' t)$$

where

$$S_1' = \{255, 755, 1255, 1755\},$$

$$S_2' = \{259, 759, 1259, 1759\},$$

$$S_4' = \{261, 761, 1261, 1761\},$$

and the elements of each frequency set are divided by 4096. Morrice (1995) transforms the data from each set of runs to the frequency domain separately. The two resulting data sets are combined into one data set in the frequency domain. Then the last three steps in the procedure described in section 2 are used to get estimates of the $\{\beta_k(r)\}$.

The purpose of Method 2 is to eliminate the need for a second set of simulation runs. Method 2 has exactly the same run time set-up as the first set of runs on Method 1 except that the factors are varied according to

$$X_1(t) = 10 + 0.5 \sum_{\tilde{\omega}_1 \epsilon \tilde{S}_1} \cos(2\pi\tilde{\omega}_1 t),$$

$$X_2(t) = 6 + 0.5 \sum_{\tilde{\omega}_2 \epsilon \tilde{S}_2} \cos(2\pi\tilde{\omega}_2 t),$$

and

$$X_4(t) = 6 + 0.5 \sum_{\tilde{\omega}_4 \epsilon \tilde{S}_4} \cos(2\pi\tilde{\omega}_4 t)$$

where

$$\tilde{S}_1 = \{5, 255, 505, 755, 1005, 1255, 1505, 1755, 2005\},$$

$$\tilde{S}_2 = \{9, 259, 509, 759, 1009, 1259, 1509, 1759, 2009\},$$

$$\tilde{S}_4 = \{11, 261, 511, 761, 1011, 1261, 1511, 1761, 2011\},$$

and the elements of each frequency set are divided by 4096. Notice that reducing the amplitude to 0.5 for each factor facilitates choosing more cosine terms for $X_2(t)$ and $X_4(t)$ and thus eliminates the need to do the second set of runs required by Method 1.

For all three scenarios, Method 2 provides comparable results to Method 1. In particular, relative to Method 1, Method 2 provides slightly degraded results for scenarios 1 and 3, but slightly improved results for scenario 2. These results are encouraging because Method 2 requires half the number of observations of Method 1.

Method 3 relies on two sets of runs. The first set are those from Method 2. The second set has exactly the same experimental setup as the first set of runs except that the factors are fixed at their nominal levels. This is called a *control run*. Common random number streams are used between the two sets of runs. Then the performance measure from the second set of runs is used as a control variate for the first set of runs. This is implemented by simply taking a difference between the data series from the first and second sets of runs. The resulting data series is then used in the transient sensitivity analysis procedure described in section 2 to produce the results for Method 3.

The figures illustrate that the control variate scheme does provide some improvement in the results. In particular, Method 3 provides uniform improvement over the results provided by Method 2.

5 CONCLUSIONS

In this paper, we have considered issues governing amplitude selection in transient sensitivity analysis. We have illustrated empirically that using smaller amplitudes offers potential for a more efficient experimental design. A key component of this strategy is the use of variance reduction techniques to reduce noise that often masks low amplitude signals in the simulation output.

Future research includes exploring the use of other types of variance reduction schemes. For example, we will consider using a control variate from a control run that oscillates factors according to (2) with an amplitude of $\{-a_k\}$ (Jacobson 1993).

ACKNOWLEDGEMENTS

The first author acknowledges the financial support of the CBA/GSB Faculty Research Committee of the College of Business Administration, The University of Texas at Austin. The second author acknowledges

the support from the NSF (DMI-9409266) and the AFOSR (F49620-95-1-0124).

REFERENCES

Hazra, M. M., D. J. Morrice, and S. K. Park. 1995. A simulation clock-based solution to the frequency domain experiment indexing problem. Working Paper, Department of Management Science and Information Systems, The University of Texas at Austin, Austin, Texas.

Jacobson, S. H. 1993. Variance and bias reduction techniques for the harmonic gradient estimator. *Applied Mathematics and Computation* 55:153–186.

Morrice, D. J. 1995. An experimental approach to transient sensitivity analysis. Working Paper, Department of Management Science and Information Systems, The University of Texas at Austin, Austin, Texas.

Morrice, D. J. and I. R. Bardhan. 1995. A weighted least squares approach to computer simulation factor screening. *Operation Research* (to appear).

Morrice, D. J. and A. Gupta. 1994. Transient sensitivity analysis in computer simulation. In *Proceedings of the 1994 European Simulation Multiconference*, ed. A. Guasch and R. Huber, 858–862. Society for Computer Simulation International, Barcelona, Spain.

Morrice, D. J., R. S. Gajulapalli, and S. R. Tayur. 1994. A single server queue with cyclically indexed arrival and service rates. *Queueing Systems: Theory and Applications* 15:165–198.

Whitt, W. 1991. The pointwise stationary approximation for $m_t/m_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37:307–314.

AUTHOR BIOGRAPHIES

DOUGLAS J. MORRICE is an assistant professor in the Department of Management Science and Information Systems at The University of Texas at Austin. He received his undergraduate degree in Operations Research at Carleton University in Ottawa, Canada. He holds an M.S. and a Ph.D. in Operations Research and Industrial Engineering from Cornell University. His research interests include discrete event and qualitative simulation modeling and the statistical design and analysis of large scale simulation experiments. He is a member of the The Institute for Operations Research and Management Science (InfORMS). He is currently the Secretary for the InfORMS College on Simulation.

SHELDON H. JACOBSON is an Assistant Professor in the Department of Industrial and Systems Engineering at Virginia Polytechnic Institute and State University (Virginia Tech). Before joining Virginia Tech, he served for five years on the faculty in the Department of Operations Research at Case Western Reserve University. He has a B.Sc. and M.Sc. in Mathematics from McGill University, and a Ph.D. in Operations Research from Cornell University. He has served as the Advanced Tutorial Track Coordinator at both the 1994 and the 1995 Winter Simulation Conferences. He also served as the Doctoral Colloquium Coordinator at both the 1993 and the 1994 Winter Simulation Conferences. At present, he is the Treasurer for the InfORMS College on Simulation. His research interests include simulation optimization and sensitivity analysis, frequency domain approaches to analyzing simulation outputs, and issues related to the complexity of analyzing structural properties of discrete event simulation models.