Proceedings of the 1990 Winter Simulation Conference
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

SEQUENTIAL BAYESIAN ANALYSIS OF SIMULATION EXPERIMENTS

Jorge Haddock
Thomas R. Willemain

Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

ABSTRACT

This research incorporates prior information in the analysis of simulation experiments to reduce the number of observations needed to estimate delay before service in a queuing system with a specified precision. The research extends to simulation of queuing systems a previous successful application of sequential Bayesian analysis to interrupted time series experiments. The results include methodologies that use sequential Bayesian methods for analyzing batch means as well as observations of individual entities within a single simulation run.

1. INTRODUCTION

Suppose there is an expert with knowledge about how the performance of a system would change as result of design changes. If a valid simulation model of the process began to produce estimates of delay that are consistent with the expert's judgement, then the simulation could terminate relatively quickly. In this case, the expert's judgement would serve as a partial substitute for computation. (However, if the simulation results were to surprise the expert, he would have to make more computations and revise his opinion.)

Accurate priors should be able to substitute for significant amounts of computation. We test this proposition by developing sequential Bayesian methods for analysis of batch means and for estimation of the probability of long delays using data from individual customers.

2. LITERATURE REVIEW

Little published work deals with the application of Bayesian methods to queuing systems. A few pioneering papers applied Bayesian methods to analytical queuing theory, notably McGrath, Gross and Singpurwalla [1987] and McGrath and Singpurwalla [1987]. These papers showed how to convert uncertainty about input parameter values into uncertainty about state distributions and performance measures. The only paper about Bayesian simulation addressed the same problem. Glynn [1986] considered the computational issues involved in translating priors on input parameters into Bayesian confidence intervals for simulation output. Glynn did not address, but called for research on, the important problem we are exploring: exploitation of a practitioner's priors on outputs rather than inputs.

The present research extends to simulation analysis the results of Willemain and Hartunian [1982] on the Bayesian analysis of interrupted time series experiments. That work showed that, even evaluating the results in a non-Bayesian framework, the Bayesian approach to combining data with judgement can yield significant improvements in efficiency.

Willemain and Hartunian considered the evaluation of interventions designed to reduce the (Poisson) rate of serious crimes. Their work examined the interrupted time series design, consisting of a baseline data collection period followed by an experimental period. They noted that the baseline period could be significantly shortened if the baseline results unfold in a way consistent with the evaluator's prior beliefs about the existing crime rate. Holding total resources constant, the sequential Bayesian estimates of the reduction in crime rate produced Mean-Square-Error (MSE) up to 25% lower than conventional estimates based on equal allocations of resources to the baseline and experimental phases. While this work did not focus on queuing systems, it did demonstrate the potential efficiencies of sequential Bayesian methods.

The other finding of Willemain and Hartunian with significance for the simulation research concerned the sensitivity of the efficiency gains to the bias and uncertainty of the evaluator's prior distributions. The degree of improvement depended on the strength and accuracy of the evaluator's joint prior for the baseline crime rate and the proportional reduction in the rate. With diffuse priors, gains were weaker but less sensitive to inaccuracies. With firm priors, gains were greater but more sensitive to inaccuracies in the prior estimate of the baseline rate, though quite insensitive to the prior for proportional reduction in the rate. Accordingly, one of the primary issues in our simulation research is whether sequential Bayesian estimates can be successful using priors with only modest accuracy.

3. SEQUENTIAL BAYESIAN ANALYSIS

3.1 Overview

We develop the Bayesian methods in the context of the M/M/1 queuing system. One method develops sequential Bayesian estimates of mean delay from aggregated data, using averages computed from batch means. The other method develops estimates of the probability of long delay from disaggregated data, using the (correlated) delay data of individual customers.

3.2 Sequential Bayesian Estimates from Aggregated Data

Consider the problem of estimating the mean customer delay in a queuing system. In the conventional approach, the estimate is calculated using averages computed either from several independent replications or from a single long run divided into batches. The independent averages or batch means form the basis for a confidence interval using Student's t-distribution. If the simulation is non-terminating there may be an additional step of deleting initial data from the transient period. In the sequential versions of these methods, which are taken as the standards against which to compare sequential Bayesian methods, the first step is to determine a standard for proportional uncertainty in the estimate. Then the

methods use some data, check whether the precision standard has been satisfied, and, if not, gather and analyze more data. Banks and Carson [1984] described the sequential independent replications method, while Law and Kelton [1982] described the sequential batch means method.

Sequential Bayesian methods have the potential advantage of substituting a prior distribution for part of the computation. However, they run the risk that the possibility that the analyst's priors may either be vague and unhelpful, or firm but wrong (and therefore require more sample values before converging on a good estimate). This paper presents results that show the net effect of sequential Bayesian methods can be positive.

Sequential Bayesian estimation from aggregated data would proceed as follows.

1. Establish a precision standard for the estimator. This standard is expressed in terms of the relative half-width of a 90% highest density region or credible interval (CI), the Bayesian analog of a confidence interval.
2. Elicit a prior distribution. For estimates of mean delay, the prior must be a joint distribution for the mean and standard deviation of delay. Assuming that the batch means are independent and Normal, there is a conjugate prior: the Normal/Chi-squared, also known as the Normal/Inverted Gamma [Lee 1989]. This prior has four parameters: two relating to the presumed values of mean and variance and two relating to the strength of belief in these presumed values. In stochastic service systems, the mean and variance tend to be positively associated, so an issue for further study is how to reduce the need for four prior parameters to a more manageable two. We present the results of one such selection below.
3. Run the simulation to obtain a few sample data, i.e., the means of some number of batches, say three or more. From these, compute the sample mean and variance of the new batch means.
4. Update the posterior distribution. Since the Normal/Chi-squared is a conjugate prior, the updating calculations are straight forward.
5. Evaluate the convergence of the posterior distribution. The posterior is characterized by the relative half-width of a 90% CI.
6. If the relative half-width of the CI satisfies the precision criterion of step 1, go to step 7; otherwise, go to step 3.
7. Report results and stop.

3.3 Sequential Bayesian Estimates from Disaggregated Data

The method described above concerns substitution of sequential Bayesian estimates for conventional estimates when each datum is the average delay of many simulated customers. We also present a more innovative approach, based on continuous updating of estimates as each simulated customer finishes processing.

Continuous updating promises to be even more efficient, further avoiding the simulation of more customers than needed to achieve a given level of precision in estimates of system performance. However, continuous updating also has problems. One is the added computational overhead of the updating calculations. Another is the difficulty of expressing the likelihood function simply, given that successive delays are highly correlated at the level of individual customers.

To cope with the problem of developing a likelihood function, we restrict our choice of performance measures to an important special case: the probability that queuing delay before service exceeds a specified level. This restriction reduces the data set to a binary sequence, no matter what the nature of the queuing system. Then we assume that the binary sequence can be adequately modeled by a first-order Markov process. A Markov model provides a feasible route to a likelihood function since it expresses the dependence in the data in a compact and manageable way.

The inspiration for this approach comes from the work of Kedem [1980], Daley [1968] and Stanford et al. [1987]. Kedem shows that any "clipped" or "hard-limited" stationary sequence (such as the binary sequence of long delays) can be approximated by a Markov process of some order. In particular, Kedem shows that a clipped AR(1) process can reasonably be modeled by a first-order Markov process. Plots of the results of Daley and Stanford et al. on serial correlations of delay in M/M/1 and GI/M/C systems, respectively, show that the stochastic process of delay values in those systems can indeed be plausibly approximated by a simple AR(1) model (i.e., the log autocorrelogram is essentially linear). This suggests that the stochastic process of occurrences of long delays can profitably be modeled as a first-order Markov process. West and Mortera [1987] showed that a second-order Markov process generated by clipping an MA(1) process was well-approximated by a first-order Markov model.

The Markov modeling approach developed here contrasts with the two alternatives we have identified in the literature. Law [1983] suggested applying the sequential batch means method to the binary data. Fishman and Moore [1979] developed a method based on the theory of recurrent states. It would be useful to compare these methods with the Bayesian approach we outline next.

Let d_i represent the delay of the i^{th} customer and x_i the value of the delay clipped at level T, i.e.,

$$x_i = 1 \text{ if } d_i > T$$
$$= 0 \text{ otherwise.}$$

The performance measure we wish to estimate is the probability of a long delay

$$\pi = Pr[x_i = 1]$$

We will assume the sequence of binary values $\{x_i\}$ to be a stationary, first-order Markov process characterized by transition probabilities

$$p = Pr[x_i=1 \mid x_{i-1} = 0]$$
$$q = Pr[x_i=0 \mid x_{i-1} = 1]$$

Given values of p and q, the performance measure is

$$\pi = p/(p+q)$$

The basic data for this process are the counts of state transitions. Define indicator random variables to denote the four types of transitions

$$\delta_{jk} = 1 \text{ if } x_i = k \text{ and } x_{i-1} = j \quad j,k \in \{0,1\}$$
$$= 0 \text{ otherwise.}$$

After each transition, the Bayesian analysis converts a joint prior distribution for p and q into a joint posterior distribution, using the likelihood of the observed transition

$$\text{Posterior}(p,q) \propto \{ (1-p)\delta_{00}+p\delta_{01}+q\delta_{10}+(1-q)\delta_{11} \} \times \text{Prior}(p,q)$$

Given the joint posterior distribution of p and q, one can derive the cumulative distribution function (CDF) of the performance measure

$$Pr[\ \pi \le \pi_0\ ] = Pr[\ p \le q\ \pi_0/(1-\pi_0)\ ]$$

Given the CDF, one can compute the 90% CI for π. Comparing the CI to the precision standard provides the stopping rule for the sequential analysis.

An important practical issue is how to specify the joint prior distribution for the Markov parameters p and q. In practice, the system expert providing the prior is unlikely to think in terms of p and q; rather, he will think directly in terms of π. It is possible to convert estimated quantiles for π into a smooth joint prior for p and q. However, for this research, we adopt the expedient of expressing the joint prior more simply as

$$\text{Prior}(p,q) = \text{Prior}(p|q) \times \text{Prior}(q)$$

where Prior (q) is uniformly distributed over (0,1) and Prior $(p|q)$ is uniformly distributed between the lines $q\pi_H/(1-\pi_H)$ and $q\pi_L/(1-\pi_L)$.

4. EXPERIMENTS

4.1 Preliminary Results

An empirical test of sequential Bayesian methods is presented in this section. We generated ten datasets containing customer delays before service in an M/M/1 queuing system with utilization 0.5. We use these datasets to compare the conventional sequential batch means procedure against a Bayesian sequential analysis of the same batch means (the aggregated analysis). We also used them to test the Markov approximation in making a sequential Bayesian estimate of the probability of long delays (the disaggregated analysis).

A thorough empirical test would involve much more experimentation, including larger sample sizes, analysis of systems more complex than the M/M/1 queue, wider variation of parameter values such as system utilization, and exploration of alternative prior distributions. Nevertheless, this limited test serves to demonstrate the concept and point the way for future experimentation.

4.2 Aggregated Analysis

We generated ten test datasets for an M/M/1 queue operating at 50 percent utilization. For this system, the known steady-state delay was 0.5. Each dataset consisted of 40 batches created according to the procedure of Law and Carson [1979], resulting in batch sizes ranging from 40 to 80 customers. For each dataset, we determined two performance measures for the 90 percent confidence intervals: whether the confidence interval did include the known value of 0.5, and the relative half-width of the confidence interval. Averaging over the 10 datasets provided a characterization of the performance of the various confidence interval procedures.

Using the ten sets of 40 batch means, the Law and Carson procedure produced 90 percent confidence intervals that included the true value in nine of ten cases. The average relative half-width over the ten simulators was 0.143. Normally, a value of 0.143 would be insufficiently precise, and more batches would be needed. However, we were able to reduce the relative half-width using the original 40 batches by as much as 50 percent while maintaining coverage, using priors of sufficient accuracy and precision.

As noted above, the Bayesian analysis requires specification of four prior parameters. We believe that the system expert could specify a normal prior for the mean delay given the variance, but would find it difficult to specify a prior for the variance itself. We experimented with a heuristic that takes advantage of a property of GI/G/C queuing systems [Köllerström, 1974]: in heavy traffic, the distribution of delay is approximately exponential, so the variance is roughly the square of the mean. Accordingly, we set the prior estimate of the variance to the square of the prior estimate of the mean, and we set the degree of confidence in the estimate of the variance equal to half the confidence in the estimate of the mean.

This heuristic reduced the problem of specifying a prior from a four-parameter to a two-parameter problem. Then we tested the sensitivity to those parameters by varying the degree of bias and imprecision in the prior for the mean delay. Figure 1 shows the results of the sensitivity analysis. When the prior was perfectly accurate and highly confident (upper left of graph), the Bayesian CI was over 50 percent narrower than that produced by Law and Carson's procedure. At each level of confidence, adding bias to the prior had only a small effect on relative half-width, though at some point the bias was great enough to ruin the coverage (right side of the graph). As long as increased bias was accompanied by a reduction in confidence, it was possible to achieve notable reductions in relative half-width (e.g., when the prior mean was 0.57 — representing a 14 percent positive bias, a confidence parameter of 100 resulted in an 18 percent improvement in the precision of the estimate.

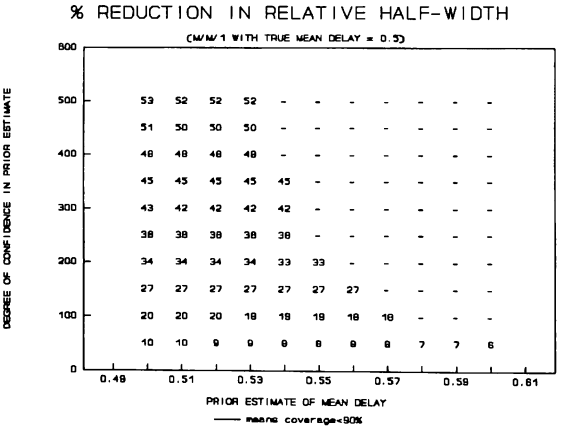

Figure 1. Sequential Bayesian Analysis of Batch Means

This analysis is clearly not exhaustive, but it does show : (1) that sequential Bayesian estimation can produce dramatically more precise estimates without sacrificing coverage; (2) that is is possible to create workable prior distributions; and (3) that even quite imperfect priors can still lead to notable improvements in precision.

4.3 Disaggregated Analysis

Our first step was to test the validity of the assumption that the binary sequence of clipped delay times could be considered a first-order Markov process. For this investigation, we generated an additional five datasets drawn from an M/M/1 queue with utilization 0.9. We chose this heavily loaded system because the problem of serial correlation worsens with utilization. We analyzed each series using three threshold values for delay: 2, 5 and 10; for comparison, the mean delay in this system is known to be 8.1.

We made an informal, graphical test of the hypothesis that the binary sequence is first-order Markov. We did this by estimating the second-order transition probabilities, P_{ki0}, where i and k are either 0 or 1. In a first-order process, P_{000} and P_{100} both equal the same number, while P_{010} and P_{110} both equal some other number. The results of this test are shown in Figure 2. The conclusion is that the process is actually of at least second order, since $P_{000} \neq P_{100}$ and $P_{010} \neq P_{110}$. However, the first-order approximation may still work reasonably well, since the bulk of the transitions are of the form $0 \rightarrow 0 \rightarrow 0$ and $1 \rightarrow 1 \rightarrow 1$.

5. CONCLUSIONS AND EXTENSIONS

Simulation is a proven, viable tool for the analysis of large, complex systems. One of its major drawbacks, however, is the cost of the experimentation. Even with the development of faster and cheaper computers, the cost issue will probably always exist. Just as the advent of supercomputers has not met the demands of weather forecasters or particle physicists, so simulators keep ratcheting up their ambition for modeling and analyzing more complicated systems in more time-critical applications.

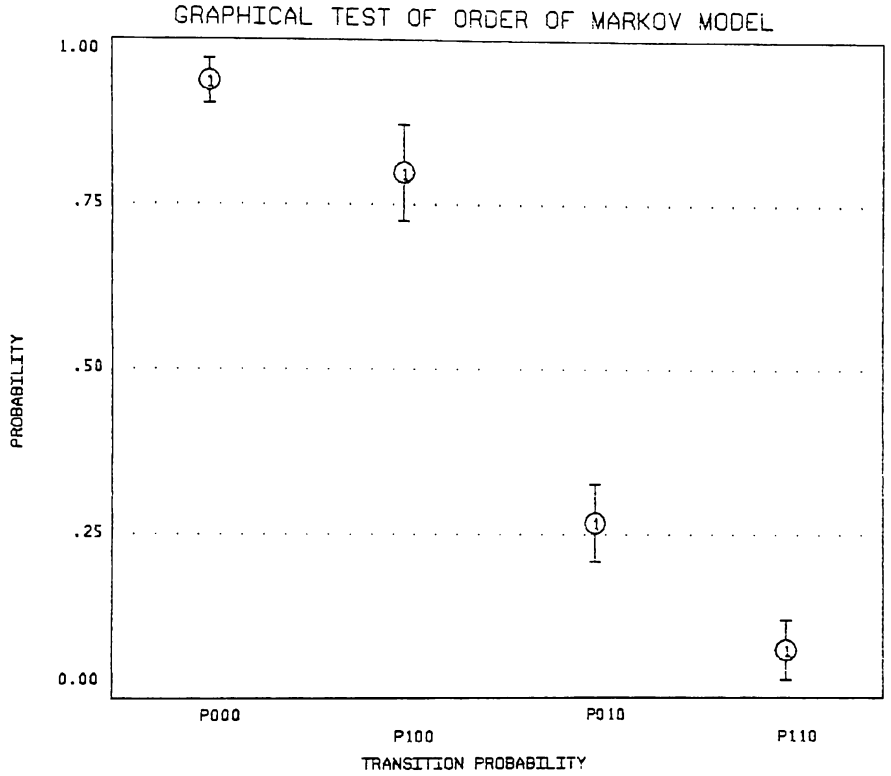

Figure 2. Graphical Test of Order of Markov Process

Our second step was to execute the sequential Bayesian analysis on two delay datasets from the M/M/1 queuing system with utilization 0.5. We chose as a threshold value the 75th percentile of the known delay distribution. Thus, the estimates of the probability of long delay should be 0.25. Figure 3 shows the 90% confidence interval for the Bayesian estimate as a function of the number of customers processed for one of the two datasets. (Note that the prior distribution is neither accurate nor precise.) After 2500 customers, the 90% confidence interval centered on 0.254 with a half-width of 0.022. Figure 3 also shows the simple empirical estimate that one would get by treating the binary sequence of delays as a Bernoulli sequence. Ignoring the serial correlation among delays would have resulted in using this simple estimate of 0.257 with a half-width of only 0.015. This result would surely be too optimistic, and we prefer the Bayesian analysis with its wider but more realistic confidence interval. For comparison, the second dataset yielded a Bayesian 90% confidence interval of 0.243 with a half-width of 0.020; the simple empirical estimate was 0.238 with a half-width of 0.014. In both cases, the Bayesian confidence intervals were about 45 percent wider.

Work by Willemain and Hartunian [1982] established that sequential Bayesian methods can improve the efficiency of resource-constrained interrupted time series experiments. Bayesian methods are just beginning to appear in the queuing theory and simulation literature. We expect sequential Bayesian methods to play a role in pushing out the "edge of the envelope" in simulation. Although sequential Bayesian procedures run some risks (i.e., additional effort in the case of wrong prior information), the savings are potentially significant. The empirical results reported here strongly suggest that the sequential Bayesian approach is worth pursuing. In the analysis of aggregated data from the M/M/1 queue, we found that excellent priors could decrease the relative half-width of confidence intervals for mean delay by 50 percent, and moderately good priors could provide gains of about 20 percent. In the analysis of disaggregated data, the Bayesian method produced confidence intervals about 45 percent wider than those ignoring serial correlation. Subject to extensive empirical testing of coverage, we argue that this is a step in the direction of more valid inference.

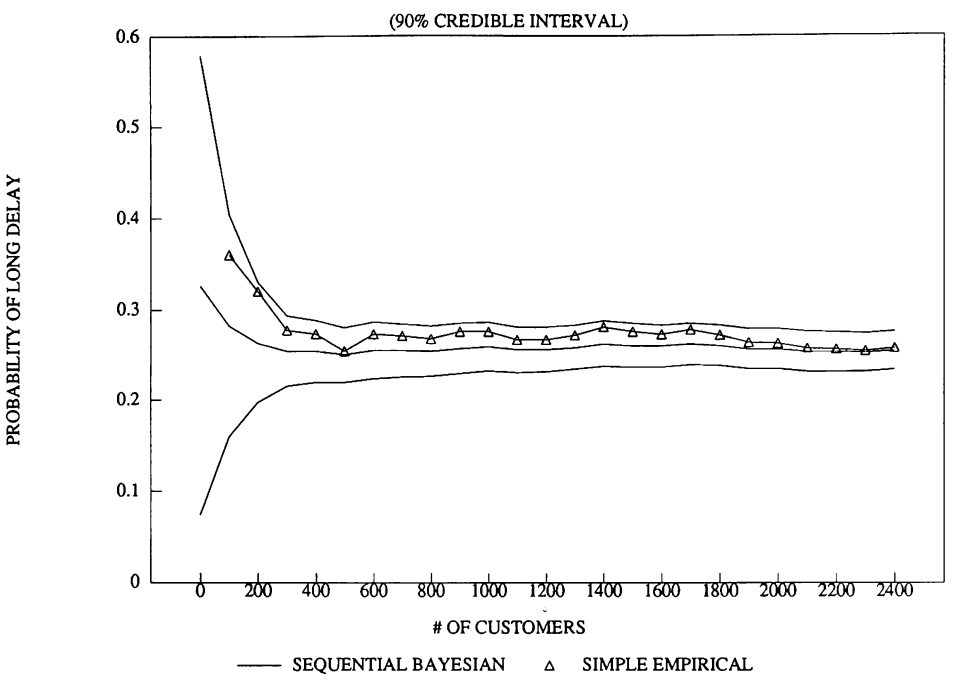

Figure 3. Sequential Bayesian Analysis of Individual Delays

We expect that there are a number of obvious extensions of this work. The analysis of aggregated data should apply to independent replications as well as batch means, and to total time in system as well as delay before service. The analysis of individual customer data should apply to the probability of loss in systems with finite queue capacity as well as to the probability of long delays in systems with unlimited queue capacity.

Another possible extension of the work would use Bayesian methods to cope with start-up bias in non-terminating simulations. A prior distribution for queue length might be used in either of two ways. One approach, analogous to an idea by Kelton [1988], would reduce the warmup period by using prior information to start the simulation at different initial conditions. The other approach would use the prior mean to determine the deletion point.

REFERENCES

Banks, J. and J. S. Carson (1984), *Discrete-Event System Simulation*, Prentice-Hall, New York, NY.

Daley, D. J. (1968), "The Serial Correlation Coefficients of Waiting Times in a Stationary Single Server Queue," *Journal of the Australian Mathematical Society*, 8, 693-699.

Glynn, P. W. (1986), "Problems in Bayesian Analysis of Stochastic Simulation," *Proceedings of the 1986 Winter Simulation Conference*, 52-59.

Kedem, B. (1980), *Binary Time Series*, Marcel Dekker, New York, NY.

Kelton, D. (1988), "Random Initialization Methods in Simulation," University of Minnesota Supercomputer Institute Report UMSI 88/110.

Köllerström, J. (1974), "Heavy Traffic Theory for Queues With Several Servers", *Journal of Applied Probability*, 11, 544–552.

Law, A.M. (1983), "Statistical Analysis of Simulation Output Data", *Operations Research*, 31:6, 983-1029.

Law, A. M. and J. S. Carson (1979), "A Sequential Procedure for Determining the Length of Steady-State Simulation," *Operations Research*, 27:5, 1011-1025.

Law, A. M., and Kelton, W. D. (1982), *Simulation Modeling and Analysis*, McGraw-Hill, New York, NY.

Lee, P.M. (1989), *Bayesian Statistics: An Introduction*, Oxford University Press, New York, NY.

McGrath, M. F., D. Gross and N. D. Singpurwalla (1987), "A Subjective Bayesian Approach to the Theory of Queues I — Modeling," *Queuing Systems*, 1, 317-333.

McGrath, M. F. and N. D. Singpurwalla (1987), "A Subjective Bayesian Approach to the Theory of Queues II — Inference and Information in M/M/1 Queues," *Queuing Systems*, 1, 335-353

Stanford, D. A., B. Pagurek and C. M. Woodside (1987), "The Serial Correlation Coefficients of Waiting Times in the Stationary GI/M/m Queue," *Queuing Systems*, 2, 373-380.

West, M. and J. Mortera (1987), "Bayesian Models and Methods for Binary Time Series", In *Probability and Bayesian Statistics*, R. Viertl, Ed., Plenum Press, New York, NY.

Willemain, T. R. and N. S. Hartunian (1982), "The Design of Time Series Comparisons Under Budget Constraints," *Evaluation Review*, 6:2, 537-557.