Philip Heidelberger
Peter D. Welch

Adaptive Spectral Methods for Simulation Output Analysis

This paper addresses two central problems in simulation methodology: the generation of confidence intervals for the steady state means of the output sequences and the sequential use of these confidence intervals to control the run length. The variance of the sample mean of a covariance stationary process is given approximately by $p(0)/N$, where $p(f)$ is the spectral density at frequency f and N is the sample size. In an earlier paper we developed a method of confidence interval generation based on the estimation of $p(0)$ through the least squares fit of a quadratic to the logarithm of the periodogram. This method was applied in a run length control procedure to a sequence of batched means. As the run length increased the batch means were rebatched into larger batch sizes so as to limit storage requirements. In this rebatching the shape of the spectral density changes, gradually becoming flat as N increases. Quadratics were chosen as a compromise between small sample bias and large sample stability.

In this paper we consider smoothing techniques which adapt to the changing spectral shape in an attempt to improve both the small and large sample behavior of the method. The techniques considered are polynomial smoothing with the degree selected sequentially using standard regression statistics, polynomial smoothing with the degree selected by cross validation, and smoothing splines with the amount of smoothing determined by cross validation. These techniques were empirically evaluated both for fixed sample sizes and when incorporated into the sequential run length control procedure. For fixed sample sizes they did not improve the small sample behavior and only marginally improved the large sample behavior when compared with the quadratic method. Their performance in the sequential procedure was unsatisfactory. Hence, the straightforward quadratic technique recommended in the earlier paper is still recommended as an effective, practical technique for simulation confidence interval generation and run length control.

1. Introduction

This paper is concerned with two major problems in the statistical output analysis of single run, discrete event simulations: generating confidence intervals for the steady state mean of an output sequence and using these confidence intervals to control the length of the simulation. It discusses methods which can be incorporated into simulation packages and used by typical practitioners. Such methods must be completely automatic and have few user specified control parameters. The paper is not concerned with the problem of identifying and eliminating the effects due to initialization bias: we assume that the simulation is in steady state.

More specifically, we assume that the simulation generates a covariance stationary process $\{X(n), n \geq 1\}$ with mean $\mu = E[X(n)]$ and spectral density $p(f)$. Under general conditions (see [1]) the sample mean, $\overline{X}$, is, for large samples, approximately normally distributed with mean μ and variance $p(0)/N$, where $p(0)$ is the spectral density at zero frequency and N is the sample size. The factor $p(0)$ measures not only the variance of each individual observation but also the correlation between observations. Thus to place a confidence interval on μ it is sufficient to estimate $p(0)$. The methods developed in this paper use spectral analysis techniques to accomplish this.

The problem of run length control is addressed by defining a sequential procedure which continues the simulation until a confidence interval of desired accuracy is obtained.

There are two reasons why it is often impractical to store the entire output sequence. First, the length of the sequence required to obtain the desired accuracy is random, unknown in advance, and may be quite large. Second, in complex models there may be the need to analyze many such sequences. To avoid these problems the method operates on a set of batch means which are rebatched as N increases so as to occupy a relatively small, fixed amount of storage. With batched data the situation is conceptually the same as with unbatched data. Let B be the batch size and N_B be the number of batches ($N = BN_B$); then Variance $(\overline{X}) \approx p_B(0)/N_B$, where $p_B(f)$ is the spectral density of the batch means. Thus to generate a confidence interval it is sufficient to estimate $p_B(0)$. Furthermore, $p_B(f)$ becomes flat (*i.e.*, approaches a constant) as the batch size increases.

In [2] we described a method for estimating $p_B(0)$ through the application of polynomial regression to the logarithm of the (averaged) periodogram of the batch means. The degree of the polynomial (a quadratic was recommended) was fixed in advance and was selected based upon the results of empirical tests. This choice represented a compromise between small and large sample behavior. For small samples a quadratic is required to properly approximate log $(p_B(f))$ so as to obtain an unbiased estimate of $p_B(0)$. However, for large samples a quadratic is unnecessary due to the flattening of $p_B(f)$; a linear function and ultimately a constant is adequate, and they provide successively more stable estimates of $p_B(0)$.

In this paper we consider adaptive methods which select the degree of the polynomial according to the shape of the periodogram. The idea is to adapt to the changing shape of $p_B(f)$ and achieve both a more flexible procedure in the small sample region and improved large sample stability. The methods investigated for selecting the degree of the polynomial include standard sequential regression procedures and cross validation. Smoothing splines, a richer class of approximating functions which by their very nature are adaptive, were also considered. The amount of smoothing was chosen by cross validation. All of these methods can be completely automated and do not require a user's qualitative or graphical interpretation of the data.

Although we concentrate on the application of these methods to batched data, they are also applicable to unbatched data. With unbatched data, as the sample size N increases, the smoothing is done over intervals $(0, \varepsilon_N)$, where $\varepsilon_N \to 0$. In this case the assumption is that log $(p(f))$ can be approximated by a polynomial in the interval $(0, \varepsilon_N)$. Analogous to the batched case, $p(f)$ converges to a constant $(p(0))$ in the interval $(0, \varepsilon_N)$ as N increases. Experiments we have performed have shown the methods to be insensitive to whether or not the data are batched. Furthermore, if the data are batched, they are insensitive to the particular batching protocol.

The organization of the paper is as follows. Section 2 contains a brief review of the fixed degree quadratic method, the batching procedure, and the method of run length control. In Section 3 the adaptive procedures are described. Section 4 contains experimental results on these adaptive methods and their comparison to the quadratic method in both fixed length simulations and as applied in the run length control procedure. Section 5 summarizes the results and contains recommendations for practical applications.

We also point out that, although we are motivated by its potential use in simulation experiments, this methodology has much wider applicability. The paper addresses the general statistical problem of generating confidence intervals for the mean of a serially correlated, covariance stationary time series.

2. A fixed degree method

We assume the simulation generates a sample $X(1), \cdots, X(N)$ from a covariance stationary sequence and that we are interested in placing a confidence interval on the mean $\mu = E[X(n)]$. Let $\gamma(k)$ denote the covariance function at lag k and assume that

$$\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty \tag{1}$$

so that the process has a finite, continuous spectral density defined by (see for example [3])

$$p(f) = \sum_{k=-\infty}^{\infty} \gamma(k) \cos(2\pi f k). \tag{2}$$

For large values of N the variance of the sample mean $\overline{X}$ is approximately $p(0)/N$ (see [1]).

We consider methods of estimating $p(0)$ which do not require storing the entire sequence $X(1), \cdots, X(N)$. Suppose $N = BN_B$ and that we batch the sequence into contiguous, nonoverlapping batches of size B. Let $\overline{X}_B(m)$ denote the mean of the mth batch, *i.e.*,

$$\overline{X}_B(m) = (1/B) \sum_{j=1}^{B} X((m-1)B + j), \; m = 1, \cdots, N_B. \tag{3}$$

Notice that $\{\overline{X}_B(m), m \geq 1\}$ is also a covariance stationary

sequence with mean μ and a spectral density which we denote by $p_B(f)$ (the relationship between $p(f)$ and $p_B(f)$ is given in [2]). Since

$$\overline{X} = (1/N) \sum_{j=1}^{N} X(j) = (1/N_B) \sum_{m=1}^{N_B} \overline{X}_B(m), \tag{4}$$

the approximate variance of $\overline{X}$ is both $p(0)/N$ and $p_B(0)/N_B$ and in fact $p_B(0) = p(0)/B$.

In [2] we described a method for estimating $p_B(0)$ by fitting a polynomial to the logarithm of the averaged periodogram of the batch means. For completeness that method is outlined here. We consider such estimates preferable to classical windowed spectral estimates because the windowed estimates of $p_B(0)$ will be biased low for the type of spectra peaked at zero which are usually encountered in simulations (see [2] for additional discussion).

Let $I(n/N_B)$ be the periodogram of the batch means, *i.e.*,

$$I(n/N_B) = \left| \sum_{j=1}^{N_B} \overline{X}_B(j) e^{-2\pi i(j-1)n/N_B} \right|^2 / N_B , \tag{5}$$

where $i = (-1)^{1/2}$. Let $f_n = (4n - 1)/2N_B$ and define

$$J(f_n) = \log (\{I((2n - 1)/N_B) + I(2n/N_B)\}/2). \tag{6}$$

The quantity $J(f_1)$ is the logarithm of the average of $I(1/N_B)$ and $I(2/N_B)$, $J(f_2)$ is the logarithm of the average of $I(3/N_B)$ and $I(4/N_B)$, etc. If $0 < n, m < N_B/4$, then $J(f_n)$ has the following approximate properties (see [3] and [4]):

$$\begin{aligned} &E[J(f_n)] \approx \log (p_B(f_n)) - 0.270, \\ &\text{Variance } [J(f_n)] \approx 0.645, \\ &\text{Covariance } [J(f_n), J(f_m)] \approx 0.0, \qquad n \neq m. \end{aligned} \tag{7}$$

We estimate $p_B(0)$ by fitting a smooth function to $J(f_n)$. The sequence $J(f_n)$ is used because it has a constant variance and an approximately symmetric distribution. These points are discussed more fully in [2], and a figure illustrating these properties is given there. We empirically checked the variance of $J(f_n)$ and confirmed the theoretical variance. We also checked and confirmed the assumption that the $J(f_n)$'s are uncorrelated.

The following method for estimating $p_B(0)$ and generating a confidence interval for μ was developed in [2]. Let $g_B(f) = \log (p_B(f))$.

1. Calculate $I(n/N_B)$ for $n = 1, \cdots, 2K$ and $J(f_n)$ for $n = 1, \cdots, K$.
2. Using ordinary least squares fit a polynomial of degree d, $g(f) = \Sigma_{k=0}^{d} a_k f^k$, to $J(f_n) + 0.270$ for $n = 1, \cdots, K$.
3. Let the resulting least squares estimate of a_0 be $\hat{a}_0$. Under the assumption that $g_B(f)$ is a polynomial of degree d for $0 \leq f \leq 2K/N_B$, $\hat{a}_0$ is an unbiased estimate of $\log (p_B(0))$.
4. Estimate $p_B(0)$ by $\hat{p}_B(0) = C_1(K, d)e^{\hat{a}_0}$, where $C_1(K, d)$ is a constant chosen to make $\hat{p}_B(0)$ approximately unbiased. The function $C_1(K, d)$ is discussed in [2].
5. Finally, a confidence interval for μ is generated by assuming that

$$(\overline{X} - \mu)/(\hat{p}_B(0)/N_B)^{1/2} \tag{8}$$

has a t-distribution with $C_2(K, d)$ degrees of freedom, where $C_2(K, d)$ is also discussed in [2]. This is the distribution of the t random variable whose denominator squared has the same coefficient of variation as $\hat{p}_B(0)/N_B$.

In [2] the parameters $K = 25$ and $d = 2$ were recommended. For these parameters $C_2(25, 2) = 7$ degrees of freedom. However, if a linear fit was sufficient to produce an unbiased estimate of $p_B(0)$, then this equivalent degrees of freedom would increase to $C_2(25, 1) = 18$, and if a constant was adequate, $C_2(25, 0) = 77$. Furthermore in [2] we showed that

$$\lim_{B \to \infty} B p_B(f) = p(0),$$

so that for large batch sizes $p_B(f)$ is nearly flat and the quadratic fit is unnecessary. Thus the potential exists for increasing the stability of $\hat{p}_B(0)$ by successively removing the quadratic and linear terms from the regression as the shape of $p_B(f)$ changes. Section 3 describes several methods which attempt to achieve this.

We now briefly describe the batching and run length control procedures. The batching is done in a straightforward manner. We store between L and $2L$ batches and assume there are always a sufficient number of batches to generate K independent values of $J(f_n)$. The procedure generates an increasing sequence of batch sizes which are successive powers of two. If the current batch size is B, then enough observations are collected until $2L$ such batches are obtained. At that point the number of batches is halved by doubling the batch size and forming $\overline{X}_{2B}(1) = (\overline{X}_B(1) + \overline{X}_B(2))/2, \cdots, \overline{X}_{2B}(L) = (\overline{X}_B(2L - 1) + \overline{X}_B(2L))/2$. The subsequent observations are stored in batches of size $2B$ until further rebatching is necessary. The procedure requires at most $2L$ storage locations. In the experiments described below we chose $L = 100$ to reflect the practical need for economy of storage in simulation applications. However, as previously mentioned, the methods are insensitive to the batching scheme.

The run length control procedure operates on a relative confidence interval half-width criterion. A sequence of

checkpoints, $j_1, j_2, \cdots, j_{\max}$, is generated, where $j_{\max}$ is the maximum run length and represents a cost constraint. At each checkpoint a confidence interval is generated. If the relative half-width of the confidence interval (confidence interval width divided by $2 \mid \bar{X} \mid$) is less than a prespecified value, ε, the simulation is terminated. Otherwise it is continued to the next checkpoint. In [2] we suggested generating the checkpoints according to the formula $j_{n+1} = \min(1.5 \times j_n, j_{\max})$. These geometrically increasing checkpoints reduce the degradation in confidence interval coverage inherent in such a sequential procedure. We evaluate the adaptive methods using this run length control procedure with 90% confidence intervals and $\varepsilon = 0.05, 0.10, 0.15$, and 0.20. This range of accuracies seems reasonable for most practical applications.

3. Three adaptive methods

In Sections 1 and 2 we saw that the log of the spectrum, $g_B(f)$, begins with a shape which is characteristic of the process $\{X(n)\}$ but as B increases becomes progressively smoother and eventually flat. Because of this there is the potential to obtain both a more robust small sample and a more stable large sample estimate of $p_B(0)$ by having a fitting procedure which adapts to this changing shape. In this section we discuss three such procedures. Two of them apply polynomial regression but select the degree of the polynomial adaptively. One uses standard regression statistics, the other cross validation. The third applies smoothing splines with the amount of smoothing determined by cross validation. The third approach is appealing *a priori* not only because it is adaptive but also because it offers a class of fitting functions richer than the polynomials.

- *Sequential regression*

As described above, there is motivation to first examine the log of the averaged periodogram and determine what degree polynomial is required to adequately describe its shape and then fit that degree rather than to always use a quadratic. For small B a quadratic, or perhaps even a higher degree polynomial, is required to approximate $g_B(f)$. We experimented with the inclusion of a cubic in the adaptive polynomial procedures. However, it provided very little improvement in the small sample region and detracted significantly from the overall performance because of the large variance of the estimate it generates; $C_2(25, 3) = 3$ equivalent degrees of freedom. Hence it is not included in the polynomial procedures, and we attempt only to improve the large sample stability by using polynomials of degrees $d = 0$, 1, or 2.

The first approach to this problem is through the application of standard polynomial regression theory (see, for example, [5]). Let $\hat{g}_0(f)$, $\hat{g}_1(f)$, and $\hat{g}_2(f)$ be the least squares polynomials of degree 0, 1, and 2, respectively, and let $ss(1)$ and $ss(2)$ be the usual error sum of squares associated with the linear and quadratic terms, *i.e.*,

$$ss(d) = \|\hat{g}_d(f_n)\|^2 - \|\hat{g}_{d-1}(f_n)\|^2, \tag{9}$$

where $\| \|^2$ indicates the sum of squares of the components of the vector. In this case, since the variance of $J(f_n)$ is a known constant, 0.645, the tests are based on the statistics $ss(d)/0.645$ for $d = 1$ and 2. The statistic $ss(d)/0.645$ is used to test the hypothesis that a_d, the coefficient of the term of degree d, is equal to zero. Under the assumptions that $g_B(f)$ is a polynomial of degree $d - 1$ and the errors are approximately normally distributed, the statistic $ss(d)/0.645$ has approximately a χ^2 distribution with one degree of freedom. Let $\chi_1^2(\phi)$ denote the inverse distribution function of a χ^2 random variable with one degree of freedom.

The first adaptive method, which we call sequential regression, is a standard procedure for the selection of the degree in polynomial regression (see [6]). In this procedure there is the desire to choose as low a degree as is consistent with the data in the interest of having as simple a function as is consistent with the data. Hence the test is sequentially applied at some high significance level, ϕ ($\phi = 1 - \alpha$, where α is the probability of a Type I error). Specifically such a test takes the form:

If $ss(2)/0.645 \geq \chi_1^2(\phi)$, choose $d = 2$;

if $ss(2)/0.645 < \chi_1^2(\phi)$
and $ss(1)/0.645 \geq \chi_1^2(\phi)$, choose $d = 1$;

if $ss(2)/0.645 < \chi_1^2(\phi)$
and $ss(1)/0.645 < \chi_1^2(\phi)$ choose $d = 0$.

The parameter ϕ regulates the behavior of the procedure. For small values of ϕ the power of the tests is high against linear or quadratic alternatives, and it is relatively difficult to drop these terms. However, with a small ϕ, once $g_B(f)$ is flat, degrees $d = 1$ and $d = 2$ are selected rather frequently (with probabilities $\phi(1 - \phi)$ and $(1 - \phi)$, respectively) resulting in a large asymptotic variance for $\hat{p}_B(0)$. For large values of ϕ the linear and quadratic terms are dropped more readily. Once dropped, the probability is low that they will be reinstated into the regression. This results in a smaller asymptotic variance. We experimented with this procedure at a number of significance levels but report only the results for $\phi = 0.90$.

- *Polynomial selection with cross validation*

The previous method of polynomial degree selection is dependent upon distributional assumptions which are only approximate and contains a significance test parameter which must be set in an experimental fashion. The

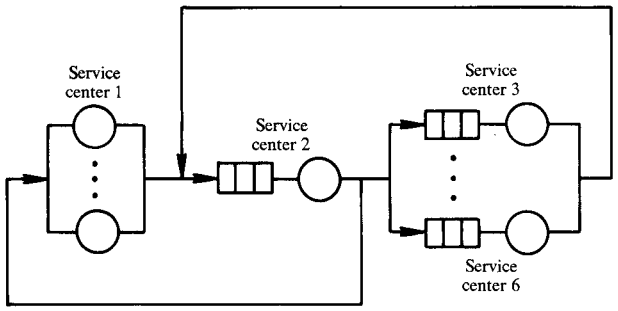


Figure 1 Closed queueing network model.

method we now describe is free of these shortcomings. It selects the polynomial degree which is in some real sense most consistent with the data. The method is known as cross validation, or PRESS (see [5, 7, 8]).

For a degree d let $\hat{g}_{d,n}(f)$ be the least squares polynomial of degree d obtained by excluding the point at f_n. The cross validation sum of squares for degree d is defined by

$$E^2(d) = \sum_{n=1}^{K} (\hat{g}_{d,n}(f_n) - J(f_n))^2. \tag{10}$$

The method of cross validation chooses from amongst a set of possible degrees that one which minimizes $E^2(d)$. We applied cross validation to select the degree where at each checkpoint in the sequential procedure the set $d = 0$, 1, or 2 was considered.

- *Smoothing splines with cross validation*

The third method applied was the method of cubic smoothing splines (see [9]) with the amount of smoothing chosen by cross validation. This method of smoothing was suggested by Wahba and Wold in [10]; see also [11–14]. The smoothing cubic spline $\tilde{g}_S(f)$ is the solution to the optimization problem

$$\text{minimize} \int_{f_1}^{f_K} (g''(f))^2 df \tag{11}$$

$$\text{such that} \sum_{n=1}^{K} (g(f_n) - J(f_n))^2 \leq S, \tag{12}$$

where g is restricted to have two continuous, square integrable derivatives. The listing of a FORTRAN program to compute $\tilde{g}_S(f)$ can be found in [15]. The parameter S controls the amount of smoothing: the larger S the smoother the function $\tilde{g}_S(f)$ and for some S_1, if $S \geq S_1$, then $\tilde{g}_S(f)$ is the least squares straight line fit to $J(f_n)$. In the notation of the two previous subsections, for $S \geq S_1$, $\tilde{g}_S(f) = \hat{g}_1(f)$. If $S = 0$, then the spline interpolates the points $J(f_n)$. The parameter S is analogous to the degree of the polynomial in polynomial regression; increasing S corresponds to decreasing d. By applying smoothing splines we have at our disposal a much richer class of smoothing functions than the low order polynomials. They can place curvature as represented by the square of the second derivative either locally or globally as needed. However, corresponding to the question of what degree polynomial is the question of what value of the smoothing parameter S. As suggested in [10] we chose S by cross validation. Let $\tilde{g}_{S,n}(f)$ be the smoothing spline for parameter S with the point f_n excluded. The cross validation error sum of squares corresponding to S is defined as

$$\tilde{E}^2(S) = \sum_{n=1}^{K} (\tilde{g}_{S,n}(f_n) - J(f_n))^2. \tag{13}$$

We choose the spline $\tilde{g}_S(f)$ whose cross validation error sum of squares is minimized. This is the third adaptive method. It will choose, on the average, a smoother and smoother curve as the batch size increases.

We actually implemented an approximation to the method of cross validation described above. We considered the set of values $\{S: S = kS_2, k = 0.40, 0.45, 0.50, \cdots, 1.30\}$, where S_2 is the error sum of squares of the quadratic, *i.e.*

$$S_2 = \sum_{n=1}^{K} (\hat{g}_2(f_n) - J(f_n))^2. \tag{14}$$

The smoothing spline from amongst this set which minimized $\tilde{E}^2(S)$ was chosen. This corresponds to a range of functions about the quadratic extending on the one hand to the linear and on the other to functions much less smooth than the quadratic. The actual estimate of $p_B(0)$ was obtained by letting $\hat{p}_B(0) = \exp\{\hat{g}_B(0)\}$, where $\hat{g}_B(0) = \tilde{g}_{S^*}(f_1) - f_1\tilde{g}'_{S^*}(f_1)$, and S^* was the value of S chosen by the cross validation. The intercept $\hat{g}_B(0)$ is thus obtained by the linear extrapolation of $\tilde{g}_{S^*}(f)$ from f_1 using the derivative of $\tilde{g}_{S^*}(f)$ at f_1.

4. Experimental results

- *Models studied and description of experiments*

The choice of models and output sequences studied in this paper directly reflects our interest in the performance modeling of computer systems. We conducted experimental studies on models of the general form shown in Fig. 1. They are simple closed queueing network models of interactive computer systems. The parameters of these models are the same as in [2], and a detailed description of them may be found there. We considered two models of this type, Models A and B, and for each model we studied a waiting time sequence at a congested queue and the sequence of system response times. The spectra of these processes are shown in [2].

Table 1 Fixed sample size simulation results for response time process in Model A, $\mu = 41.2$.

Run length		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.62	0.78	0.92	0.80	0.80	0.88
500	Width	7.53	10.38	15.59	13.86	14.43	14.98
	Var (width)	1.17	3.02	17.97	30.11	26.12	29.52
	Coverage	0.60	0.78	0.94	0.88	0.86	0.88
750	Width	6.60	10.07	13.75	12.17	12.41	12.54
	Var (width)	0.62	3.73	13.32	16.26	16.25	14.21
	Coverage	0.64	0.82	0.96	0.90	0.92	0.86
1125	Width	5.90	9.56	12.34	11.12	11.46	10.50
	Var (width)	0.39	1.95	7.99	8.88	9.36	8.67
	Coverage	0.66	0.90	0.92	0.90	0.92	0.86
1687	Width	5.38	8.85	10.42	9.06	9.57	8.92
	Var (width)	0.22	2.05	7.15	4.03	6.85	5.32
	Coverage	0.68	0.88	0.88	0.84	0.84	0.84
2530	Width	5.02	7.39	7.69	7.15	7.16	7.23
	Var (width)	0.21	2.25	3.87	3.78	3.27	4.10
	Coverage	0.78	0.94	0.94	0.88	0.92	0.90
3795	Width	4.50	6.22	6.60	6.00	6.22	6.06
	Var (width)	0.17	1.02	1.97	2.27	2.44	2.89
	Coverage	0.82	0.90	0.88	0.86	0.90	0.86
5692	Width	4.04	5.00	5.35	4.63	4.88	5.12
	Var (width)	0.13	0.43	1.09	0.89	0.86	1.22
	Coverage	0.84	0.90	0.92	0.86	0.86	0.88
8538	Width	3.38	3.92	4.29	3.57	3.81	3.86
	Var (width)	0.07	0.29	1.03	0.38	0.64	1.12
	Coverage	0.92	0.94	0.94	0.94	0.94	0.92
12 807	Width	2.77	3.11	3.27	2.97	2.92	2.95
	Var (width)	0.04	0.24	0.48	0.24	0.19	0.59
	Coverage	0.88	0.94	0.90	0.90	0.92	0.90
13 500	Width	2.62	3.00	3.23	2.80	2.79	3.02
	Var (width)	0.03	0.24	0.44	0.18	0.13	0.41

For each of the four output sequences we ran 50 independent simulations, each 14 000 elements long. These sequences are identical to those considered in [2]. The first 500 elements of each sequence were removed to control the effect of initial conditions.

Tables 1–4 report the results of fixed sample size simulations. They list results for six methods of generating confidence intervals: the three adaptive procedures described in Section 3 as well as the three fixed degree methods corresponding to $d = 0$, 1, and 2. The sample sizes in these tables are the checkpoints of the sequential procedure. The fraction of the fifty 90% confidence intervals which actually contained μ is reported for each type of output sequence, checkpoint, and confidence interval method. This fraction is called a (90%) coverage, and it should be close to 0.90 if valid confidence intervals are being formed. Coverages less than 0.82 are significantly lower than 0.90 at the 0.90 level. Space considerations preclude reporting the entire coverage function (see [16]). These tables also report the means and sample variances of the confidence interval widths.

Tables 5–8 report the results of tests on the methods when operating in the run length control procedure. These tables list the coverage, mean run length, and mean relative half-width corresponding to each confidence interval method and each of the accuracy requirements, $\varepsilon = 0.20, 0.15, 0.10, 0.05$. More specifically, the coverages and relative half-widths are those of the confidence intervals with which the run length control procedure terminates, *i.e.*, those confidence intervals which either first satisfy the accuracy criterion or, if the accuracy criterion is never met, those produced at $j_{\max}$.

- *General behavior of the adaptive procedures*

The adaptive procedures generally made reasonable decisions concerning the amount of smoothing given that only

Table 2 Fixed sample size simulation results for waiting time process at Queue 2 in Model A, $\mu = 3.77$.

Run length		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.18	0.58	0.76	0.70	0.72	0.80
500	Width	0.65	2.02	2.90	2.64	2.68	3.73
	Var (width)	0.009	0.36	1.18	1.29	1.31	4.20
	Coverage	0.40	0.64	0.74	0.72	0.74	0.80
750	Width	0.74	1.84	2.61	2.29	2.45	2.98
	Var (width)	0.010	0.25	0.78	0.83	0.90	1.91
	Coverage	0.40	0.82	0.86	0.84	0.86	0.88
1125	Width	0.77	1.79	2.50	2.24	2.37	2.55
	Var (width)	0.011	0.19	0.48	0.57	0.60	1.08
	Coverge	0.40	0.82	0.94	0.90	0.92	0.90
1687	Width	0.76	1.66	2.45	2.27	2.31	2.41
	Var (width)	0.011	0.13	0.53	0.66	0.62	1.17
	Coverage	0.58	0.92	0.96	0.94	0.96	0.90
2530	Width	0.77	1.61	2.15	1.96	2.02	1.95
	Var (width)	0.012	0.15	0.36	0.39	0.43	0.50
	Coverage	0.56	0.88	0.94	0.92	0.92	0.88
3795	Width	0.76	1.52	1.81	1.63	1.73	1.59
	Var (width)	0.012	0.09	0.26	0.19	0.25	0.23
	Coverage	0.62	0.90	0.90	0.88	0.88	0.86
5692	Width	0.75	1.31	1.45	1.35	1.35	1.30
	Var (width)	0.007	0.08	0.20	0.17	0.18	0.14
	Coverage	0.68	0.94	0.90	0.90	0.90	0.84
8538	Width	0.70	1.13	1.16	1.10	1.11	1.05
	Var (width)	0.006	0.06	0.10	0.09	0.08	0.11
	Coverage	0.66	0.80	0.84	0.82	0.82	0.80
12 807	Width	0.65	0.90	0.98	0.88	0.90	0.87
	Var (width)	0.006	0.04	0.09	0.07	0.07	0.06
	Coverage	0.72	0.86	0.90	0.88	0.88	0.86
13 500	Width	0.64	0.87	0.97	0.85	0.87	0.87
	Var (width)	0.006	0.04	0.08	0.08	0.07	0.05

25 data points, $\{J(f_n), 1 \leq n \leq 25\}$, were available to them. We were particularly impressed by degree selections made by cross validation. A sample of their selections is illustrated in Figs. 2–5. For the waiting time sequence of Model B these figures contain examples at 1125 and 13 500 observations for the first 8 of the 50 replications. With a maximum of 200 batches these correspond to batch sizes of 8 and 128, respectively. Estimates of $g_B(f)$ at these batch sizes are given in Fig. 6. Notice that $g_8(f)$ is approximately quadratic and $g_{128}(f)$ is approximately linear.

Figures 2–5 contain plots of $J(f_n)$, the three least squares polynomials $\hat{g}_0(f)$, $\hat{g}_1(f)$, and $\hat{g}_2(f)$ and the selected spline $\hat{g}_{S^*}(f)$. The corresponding intercept estimates of $g_B(0) = \log(p_B(0))$ are indicated by stars. These plots also list the degrees selected by the two polynomial procedures and the coefficient k yielding the optimal S^* [$S^* = kS_2$, where S_2 is given by (14)]. Also indicated are whether or not the confidence interval produced by each fit covered μ (Cover = 1 or 0, respectively).

The underestimation of $p(0)$ using degrees $d = 0$ or 1 for small run lengths is clearly indicated in Figs. 2 and 3. The large sample variability of the estimates produced by the quadratic procedure is seen in Figs. 4 and 5.

The increased flexibility of the splines over the polynomials is evident throughout Figs. 2–5. They generally gave reasonable looking fits although they were erratic at times and sensitive to random patterns in $\{J(f_n)\}$ [see, for example, Fig. 2, replication 4, and Fig. 4, replication 1].

- *Fixed sample size behavior*

We first discuss the results of the fixed sample size experiments. Notice from Tables 1–4 that the small sample coverages for the fixed degree $d = 2$ procedure are generally higher than those for $d = 0$ and $d = 1$. For

Table 3 Fixed sample size simulation results for response time process in Model B, $\mu = 171.0$.

Run length		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.68	0.74	0.86	0.72	0.76	0.78
500	Width	33.51	36.81	46.58	35.53	40.15	49.13
	Var (width)	11.79	24.45	111.84	96.05	139.00	288.30
	Coverage	0.70	0.82	0.90	0.80	0.82	0.82
750	Width	25.92	32.86	40.79	33.78	36.75	38.80
	Var (width)	6.23	28.05	85.15	108.26	124.82	179.19
	Coverage	0.78	0.84	0.90	0.80	0.88	0.86
1125	Width	22.11	28.35	36.73	28.41	31.15	33.61
	Var (width)	3.75	24.10	64.50	80.87	89.81	113.85
	Coverage	0.72	0.88	0.96	0.86	0.88	0.96
1687	Width	18.76	24.95	32.01	26.15	27.96	27.65
	Var (width)	4.55	15.87	67.01	84.33	96.70	51.83
	Coverage	0.80	0.92	0.94	0.90	0.94	0.92
2530	Width	16.33	22.76	26.39	22.10	24.29	23.76
	Var (width)	1.87	14.20	41.62	44.70	42.10	41.44
	Coverage	0.82	0.90	0.90	0.84	0.86	0.90
3795	Width	14.37	18.62	20.81	17.62	19.68	18.67
	Var (width)	1.35	8.81	40.19	27.41	38.99	40.87
	Coverage	0.86	0.94	0.92	0.90	0.92	0.90
5692	Width	12.87	15.87	16.55	14.80	15.47	15.31
	Var (width)	1.00	6.65	23.71	16.97	18.98	19.13
	Coverage	0.92	0.94	0.96	0.94	0.94	0.96
8538	Width	10.89	12.87	13.99	11.48	12.27	12.62
	Var (width)	0.49	2.92	10.31	3.57	6.59	7.74
	Coverage	0.92	0.96	0.90	0.92	0.94	0.92
12 807	Width	9.27	10.73	11.53	9.79	10.40	10.22
	Var (width)	0.53	2.99	6.90	2.25	4.13	5.04
	Coverage	0.94	0.96	0.92	0.94	0.94	0.94
13 500	Width	9.06	10.46	11.14	9.63	10.37	9.23
	Var (width)	0.63	2.63	6.79	3.36	5.56	4.11

small samples $p_B(0)$ is severely underestimated using $d = 0$ or 1. Corresponding to the higher coverages are wider confidence intervals for $d = 2$ than for $d = 0$ or 1. The coverages for $d = 0$ are particularly low; this case corresponds roughly to the method of batch means. The method of batch means estimates the variance by fitting a degree zero polynomial to the periodogram (see [2]) rather than to the logarithm of the averaged periodogram as is done here.

For large samples the coverages for $d = 1$ and $d = 2$ are acceptable, and their mean confidence interval widths are approximately equal. For the waiting time sequences the coverages for $d = 0$ do not reach acceptable levels even by $j_{max} = 13\,500$ observations. For all run lengths the variances of the confidence interval widths produced by the quadratic are higher than those obtained from either $d = 0$ or $d = 1$. This relationship is as predicted by $C_2(K, d)$. Thus the goal of an adaptive procedure is to move to the more stable $d = 0$ or $d = 1$ estimates of $p_B(0)$ but to do so only after their bias is sufficiently low so as to produce correct coverage.

For the two adaptive polynomial regression procedures the small sample coverages are generally greater than those for $d = 0$ or 1 although not as high as the $d = 2$ coverages. The corresponding confidence interval widths are similarly ordered. These adaptive polynomial procedures produced correct large sample coverages throughout the experiments. Notice, however, that the small sample confidence interval width variances are *larger* than those of the $d = 2$ procedure. The reason for this variance increase is that for small samples the distribution of $\hat{p}_B(0)$ using an adaptive procedure is actually more spread out than that of the quadratic due to the bias in $\hat{p}_B(0)$ for $d = 0$ or 1. For large samples this effect is not so

Table 4 Fixed sample size simulation results for waiting time process at Queue 3 in Model B, $\mu = 34.22$.

Run length		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.28	0.66	0.82	0.78	0.76	0.86
500	Width	4.11	15.15	25.10	23.65	23.67	28.36
	Var (width)	0.22	11.74	57.79	72.95	72.55	131.16
	Coverage	0.26	0.64	0.88	0.84	0.84	0.86
750	Width	4.81	15.25	23.80	22.48	22.70	26.22
	Var (width)	0.39	10.60	41.42	52.49	49.72	102.76
	Coverage	0.20	0.72	0.90	0.86	0.88	0.86
1125	Width	5.25	15.37	22.94	21.09	22.07	23.56
	Var (width)	0.37	8.61	35.08	40.39	44.95	75.11
	Coverage	0.36	0.90	0.92	0.92	0.92	0.92
1687	Width	5.62	15.82	22.50	20.40	21.08	20.98
	Var (width)	0.25	10.97	27.06	35.29	33.83	39.67
	Coverage	0.44	0.88	0.92	0.90	0.90	0.90
2530	Width	6.31	15.17	18.31	16.30	16.69	16.19
	Var (width)	0.29	5.85	20.62	17.16	18.48	20.70
	Coverage	0.54	0.92	0.96	0.92	0.92	0.90
3795	Width	6.34	13.18	14.50	13.70	13.77	12.84
	Var (width)	0.20	5.32	16.05	12.88	12.21	15.81
	Coverage	0.68	0.90	0.90	0.88	0.88	0.80
5692	Width	6.49	11.18	10.73	10.64	10.47	10.18
	Var (width)	0.18	3.28	8.18	6.22	6.46	13.60
	Coverage	0.80	0.96	0.94	0.94	0.94	0.90
8538	Width	6.07	8.93	8.57	8.56	8.47	8.20
	Var (width)	0.17	2.16	3.90	3.15	3.18	5.37
	Coverage	0.80	0.88	0.88	0.84	0.84	0.86
12 807	Width	5.44	7.01	7.42	6.69	6.72	6.83
	Var (width)	0.26	1.62	3.90	3.05	3.02	2.61
	Coverage	0.72	0.88	0.88	0.82	0.80	0.84
13 500	Width	5.31	6.87	7.03	6.47	6.54	6.72
	Var (width)	0.28	1.43	2.78	2.06	2.22	2.76

Table 5 Sequential simulation results for response time process in Model A, $\mu = 41.2$.

Relative half-width		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.62	0.78	0.92	0.78	0.78	0.80
0.20	Run length	500	500	669	634	634	625
	Half-width	0.09	0.12	0.16	0.13	0.14	0.14
	Coverage	0.62	0.76	0.92	0.74	0.70	0.74
0.15	Run length	500	548	1198	804	872	852
	Half-width	0.09	0.12	0.13	0.12	0.12	0.12
	Coverage	0.62	0.86	0.84	0.80	0.80	0.76
0.10	Run length	525	2190	2727	1979	2183	2348
	Half-width	0.09	0.09	0.08	0.08	0.08	0.08
	Coverage	0.78	0.88	0.88	0.80	0.82	0.88
0.05	Run length	6352	9633	9849	7407	8270	8582
	Half-width	0.05	0.04	0.04	0.04	0.04	0.04

Table 6 Sequential simulation results for waiting time process at Queue 2 in Model A, $\mu = 3.77$.

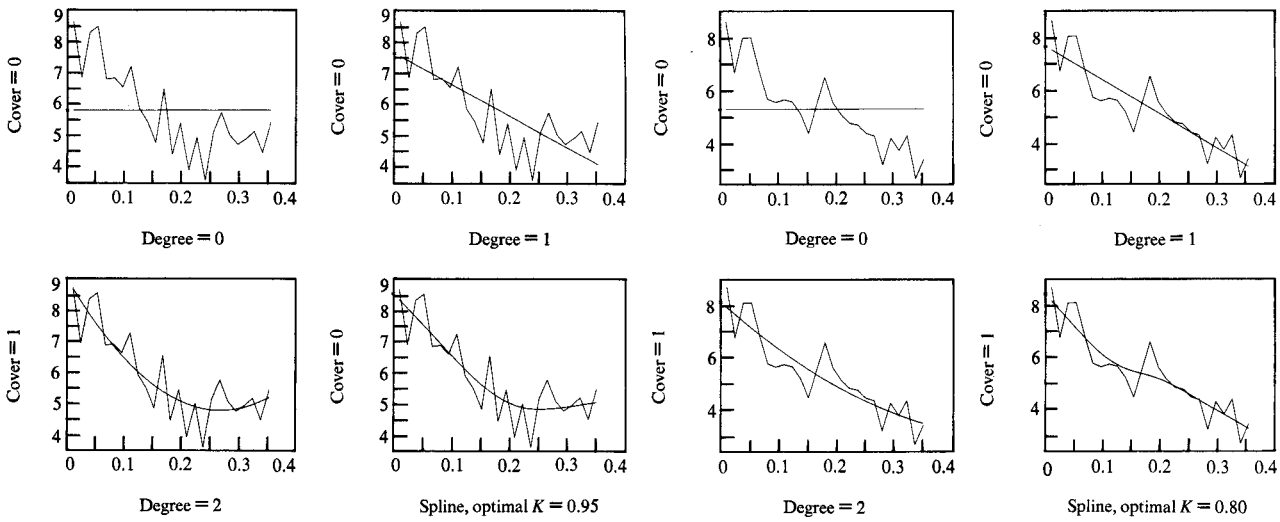
Relative half-width		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.18	0.80	0.90	0.82	0.82	0.86
0.20	Run length	500	3628	5660	4614	4510	4360
	Half-width	0.09	0.18	0.17	0.18	0.18	0.17
	Coverage	0.18	0.86	0.86	0.84	0.84	0.76
0.15	Run length	500	9432	9506	9209	9218	7847
	Half-width	0.09	0.13	0.13	0.13	0.13	0.13
	Coverage	0.36	0.80	0.84	0.82	0.78	0.76
0.10	Run length	1937	13 262	12 934	12 291	12 404	12 139
	Half-width	0.08	0.12	0.13	0.11	0.11	0.11
	Coverage	0.72	0.86	0.90	0.88	0.88	0.86
0.05	Run length	13 500	13 500	13 500	13 500	13 500	13 344
	Half-width	0.08	0.12	0.13	0.11	0.11	0.11

Table 7 Sequential simulation results for response time process in Model B, $\mu = 171.0$.

Relative half-width		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.68	0.74	0.86	0.72	0.76	0.76
0.20	Run length	500	500	510	510	510	530
	Half-width	0.09	0.11	0.13	0.11	0.11	0.13
	Coverage	0.68	0.74	0.84	0.72	0.74	0.76
0.15	Run length	500	500	590	570	558	633
	Half-width	0.09	0.11	0.12	0.10	0.10	0.11
	Coverage	0.66	0.76	0.80	0.68	0.68	0.72
0.10	Run length	520	811	1350	778	818	1144
	Half-width	0.09	0.09	0.09	0.09	0.09	0.08
	Coverage	0.76	0.84	0.84	0.82	0.82	0.80
0.05	Run length	2606	5926	6566	4809	5294	5120
	Half-width	0.05	0.04	0.04	0.04	0.04	0.04

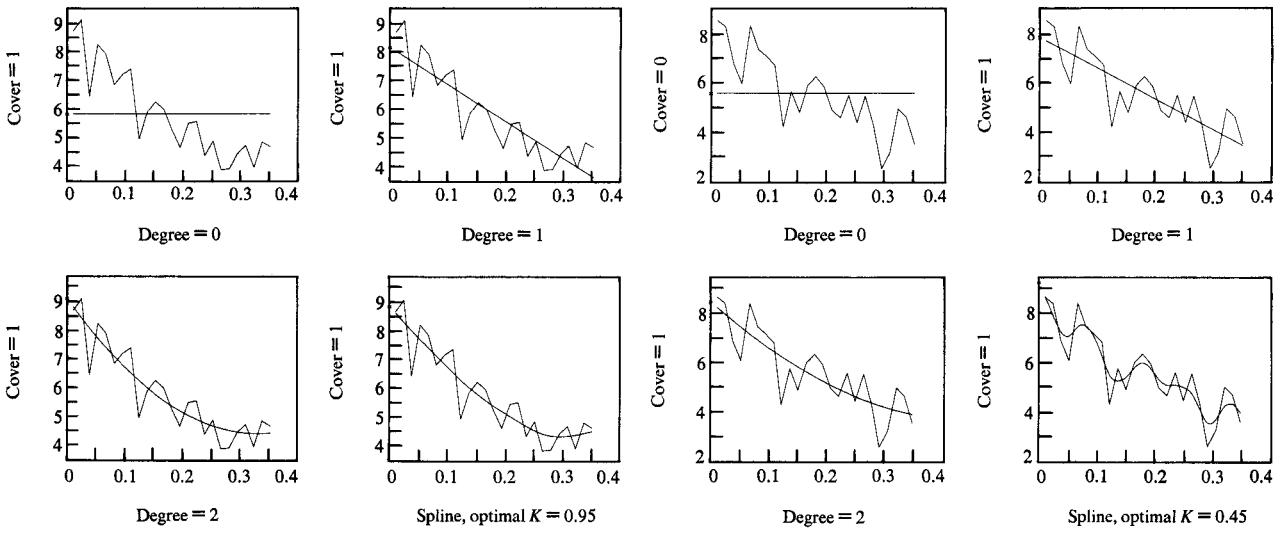
Table 8 Sequential simulation results for waiting time process at Queue 3 in Model B, $\mu = 34.22$.

Relative half-width		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
	Coverage	0.28	0.86	0.88	0.90	0.86	0.84
0.20	Run length	500	3035	4913	4404	4247	4885
	Half-width	0.07	0.18	0.17	0.17	0.17	0.16
	Coverage	0.28	0.92	0.92	0.88	0.88	0.84
0.15	Run length	500	8247	8276	7707	7465	7365
	Half-width	0.07	0.13	0.12	0.12	0.13	0.12
	Coverage	0.30	0.90	0.90	0.84	0.86	0.86
0.10	Run length	7561	12 980	12 569	12 216	12 301	11 342
	Half-width	0.07	0.10	0.10	0.09	0.10	0.10
	Coverage	0.68	0.88	0.88	0.82	0.80	0.84
0.05	Run length	12 720	13 500	13 401	13 401	13 401	13 306
	Half-width	0.08	0.10	0.10	0.10	0.10	0.10



Replication 1: cross validation selects degree = 2; sequential regression selects degree = 2.

Replication 2: cross validation selects degree = 1; sequential regression selects degree = 1.



Replication 3: cross validation selects degree = 2; sequential regression selects degree = 2.

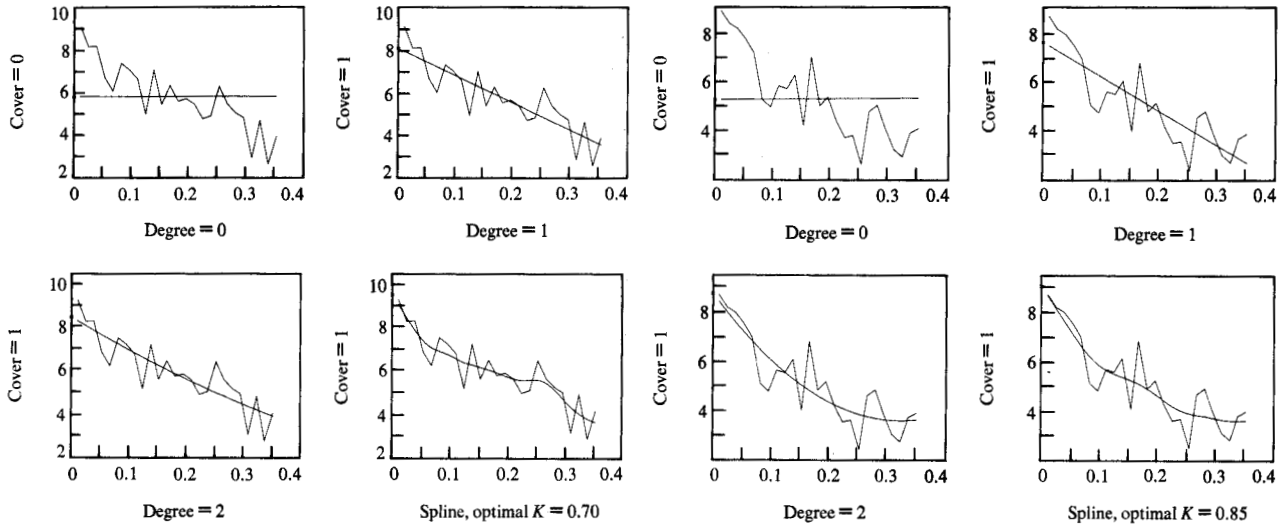
Replication 4: cross validation selects degree = 2; sequential regression selects degree = 1.

Figure 2 Model B waiting time with run length = 1125: replications 1–4.

pronounced, and the confidence interval width variances for the adaptive procedures decrease below those of the quadratic.

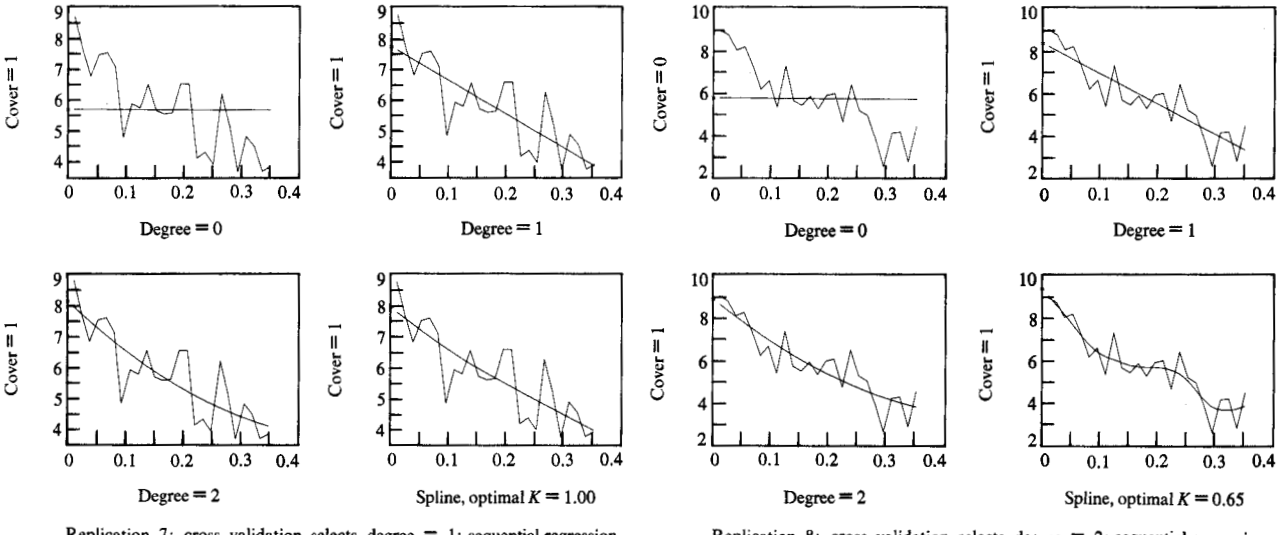
The smoothing splines generally exhibit good fixed sample coverages throughout the entire range of run lengths. The small sample confidence interval width variances are, however, much larger than those of any of the other procedures, and their large sample variances are approximately equal to those of the quadratic.

To estimate the large sample potential benefit of these methods we tested them against a set of independent observations with a flat spectrum. Specifically we ran 200 replications of 5000 independent and identically distributed exponential random variables with mean 1. These observations were batched as described in Section 2, and the methods of Section 3 were applied. The variances of the confidence interval widths for $d = 0$, 1, and 2 were 1.38×10^{-5}, 5.84×10^{-5}, and 14.8×10^{-5}, respectively, while for sequential regression, regression with cross



Replication 5: cross validation selects degree = 1; sequential regression selects degree = 1.

Replication 6: cross validation selects degree = 2; sequential regression selects degree = 2.



Replication 7: cross validation selects degree = 1; sequential regression selects degree = 1.

Replication 8: cross validation selects degree = 2; sequential regression selects degree = 1.

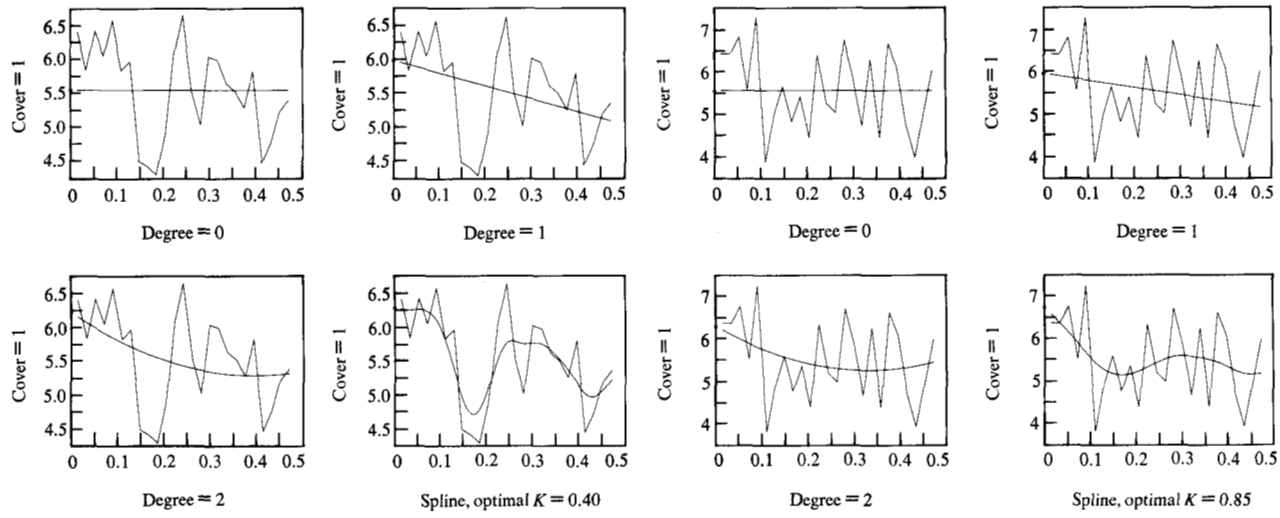
Figure 3 Model B waiting time with run length = 1125: replications 5–8.

validation, and the smoothing splines they were 7.32×10^{-5}, 7.01×10^{-5}, and 18.2×10^{-5}, respectively. Thus even with this "tailor-made data" (see [17]) neither the $d = 0$ nor the $d = 1$ variances were achieved asymptotically by the adaptive procedures, and the splines did not even beat the variance of the $d = 2$ procedure.

In this experiment with independent samples the cross validation selected $d = 0$, 1, and 2 80%, 16%, and 4% of the time, respectively. The sequential regression selected $d = 0$, 1, and 2 88%, 8%, and 4% of the time, respectively. The variances for the adaptive regression procedures are greater than if $d = 0$, 1, and 2 were independently selected with the above corresponding probabilities. This variance increase is explained by the fact that the large error sum of squares resulting in the selection of $d = 1$ or 2 tends to yield extreme values in the intercept estimates, $\hat{g}_1(0)$ and $\hat{g}_2(0)$.

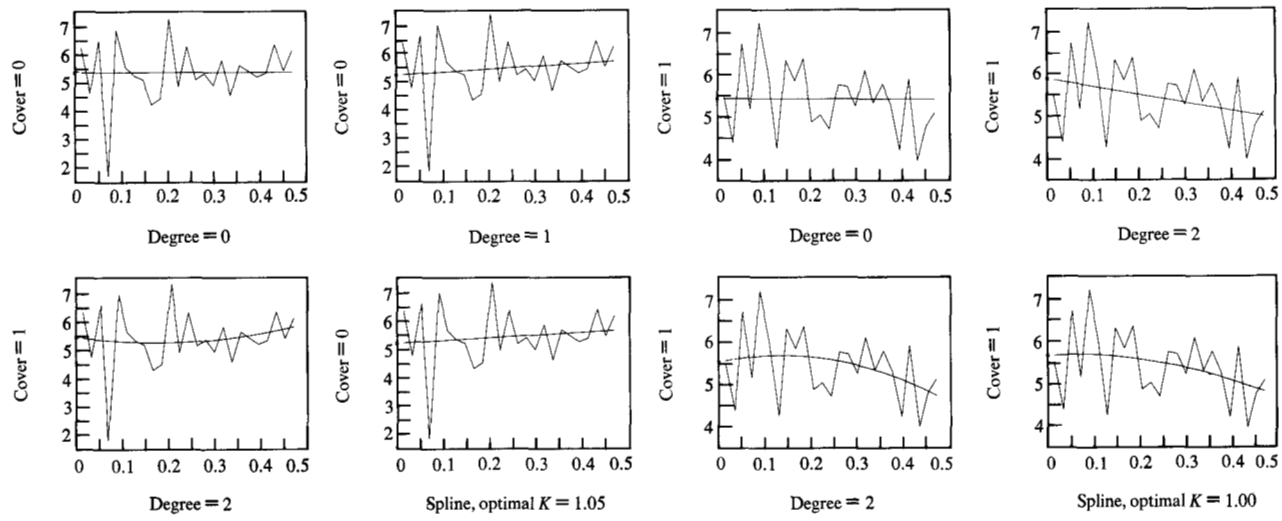
Sequential behavior

In this subsection we examine the behavior of the confidence interval methods when operating within the run



Replication 1: cross validation selects degree = 1; sequential regression selects degree = 0.

Replication 2: cross validation selects degree = 0; sequential regression selects degree = 0.



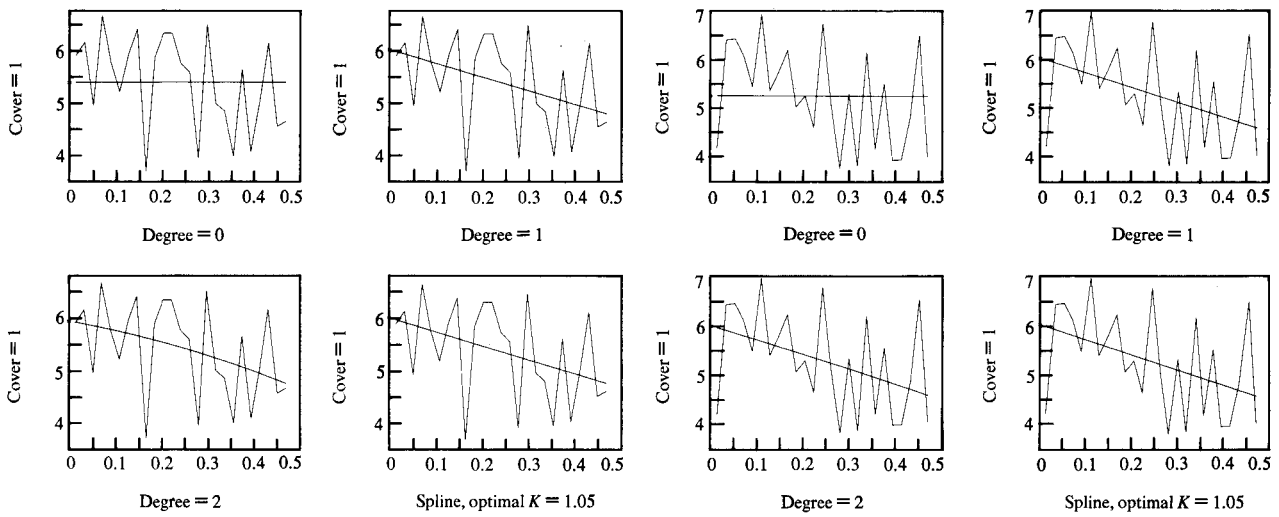
Replication 3: cross validation selects degree = 0; sequential regression selects degree = 0.

Replication 4: cross validation selects degree = 1; sequential regression selects degree = 0.

Figure 4 Model B waiting time with run length = 13500: replications 1–4.

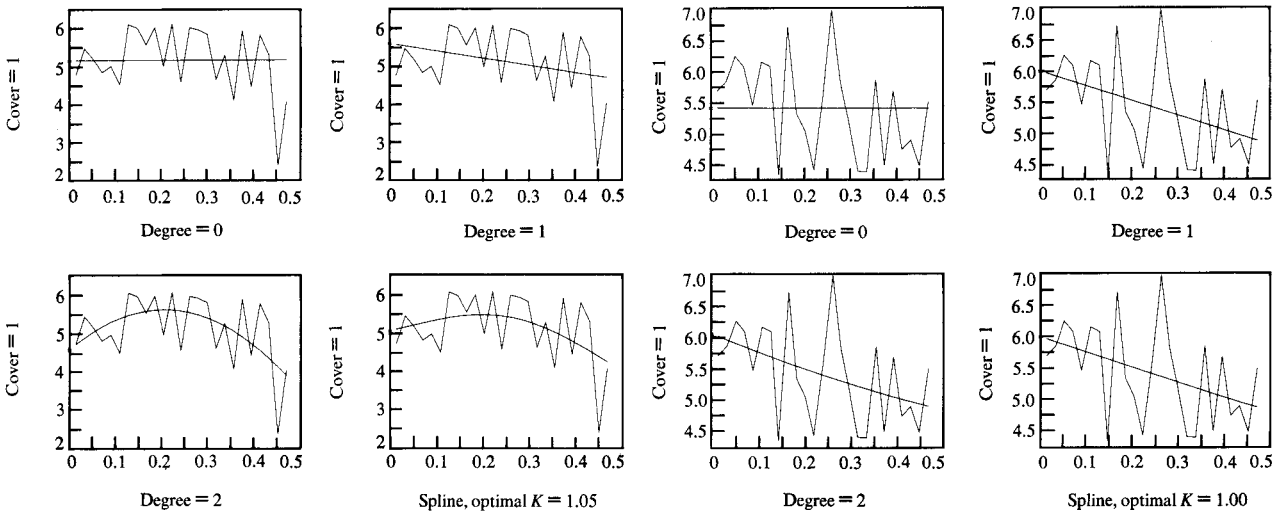
length control procedure. This behavior is summarized in Tables 5–8. Notice that the fixed degree $d = 2$ method has proper coverages for all accuracy requirements ε. However, for large ε the coverages are quite low for $d = 0$ and $d = 1$. This is a consequence of the small sample bias in $\hat{p}_B(0)$ and the resulting poor small sample coverage using $d = 0$ or 1. For small ε the coverages corresponding to $d = 1$ are generally adequate since the accuracy requirement forces the simulation to run long enough so that $d = 1$ produces essentially unbiased estimates of $p_B(0)$. The underestimation of $p_B(0)$ with $d = 0$ and 1 leads to shorter run lengths than with $d = 2$.

The adaptive methods generally do not produce acceptable coverages when operating within the run length control procedure. In the case of the polynomial methods this is related to the poor small-sample coverages of the $d = 0$ and $d = 1$ polynomials. When $d = 0$ or 1 is prematurely selected by the adaptive procedure, $p_B(0)/N_B$ is underestimated, resulting in a small relative half-width and an increase in the probability of passing the relative half-width criterion with a confidence interval which fails to cover the true value. In the case of the smoothing splines it is unexpected since their fixed sample coverages are good. We conjecture that it is related to the large



Replication 5: cross validation selects degree = 1; sequential regression selects degree = 1.

Replication 6: cross validation selects degree = 1; sequential regression selects degree = 1.



Replication 7: cross validation selects degree = 2; sequential regression selects degree = 2.

Replication 8: cross validation selects degree = 1; sequential regression selects degree = 1.

Figure 5 Model B waiting time with run length = 13500: replications 5–8.

small-sample variance which tends again to generate underestimates of $p_B(0)/N_B$ and confidence intervals which pass the relative half-width criterion and fail to cover the true value.

For values of ε in which the maximum run length is not a constraint (all ε's in Tables 5 and 7 and large ε's in Tables 6 and 8), the run lengths for these adaptive procedures are less than those with $d = 2$. This is primarily due to the somewhat shorter confidence intervals produced by the adaptive procedure.

- *Modified adaptive procedures*

We tried making a number of modifications to the adaptive polynomial procedures in an attempt to improve their small-sample and sequential behavior. They were designed to decrease and delay the possibility of selecting $d = 0$ and $d = 1$. These modifications generally operated only within the context of a run length control procedure with a sequence of checkpoints. Among the modifications we tried were

1. Excluding $d = 0$ from consideration,

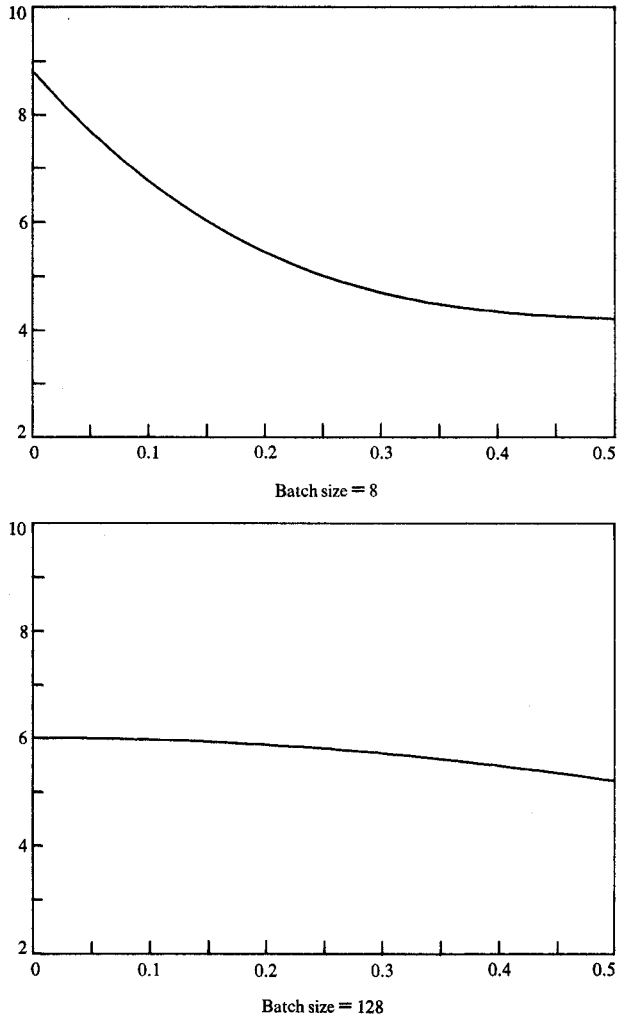


Figure 6 Logarithm of Model B waiting time spectrum with batch sizes of 8 and 128.

2. Allowing the polynomial to move down by at most one degree per checkpoint,
3. Requiring that a test be passed on two successive checkpoints before lowering the degree of the polynomial, and
4. Introducing two significance levels into the sequential regression procedure, one for moving down a degree and one for moving up a degree.

These modifications did result in marginal improvements in both small-sample and sequential coverages, but they seemed *ad hoc* and their performance overly dependent upon our particular experimental data. Furthermore, these modifications tended to increase the large-sample variances over those of the straightforward procedures described in Section 3, thus reducing the potential benefit. The only way we found to achieve the large-sample variances corresponding to $d = 0$ was to make the decisions on moving down a degree irrevocable (*i.e.*, degrees d and higher were excluded from consideration at all checkpoints after the polynomial first dropped from degree d to $d - 1$). However, this resulted in lower small-sample coverages and seemed like an inflexible and potentially dangerous approach.

Similar rules limiting the range of the smoothing parameter of the spline could probably have been devised to improve its behavior, but to do so also seemed arbitrary and contrary to the simplicity and elegance of that method. We did try weighting functions with the splines which gave more weight to the points close to zero. However, this did not result in any significant improvement.

An approach we did not investigate was to adaptively vary K, the number of points to which the polynomial is fit, rather than the degree of the polynomial. Such an investigation would appear prone to the same difficulties encountered in the present one.

- *Experiments with unbatched data*

As we mentioned earlier the performance of the methods we have discussed is not sensitive to whether or not the data are batched or, if batched, to the manner in which the batching is done. Hence, the results and conclusions of the paper do not depend upon the particular batching protocol we have used. It was selected not because of performance considerations but because it made most economical use of storage. The reason for this insensitivity is that, regardless of the batching protocol, the method operates on an increasingly narrow low frequency region of $p(f)$ as the run length N gets large. This takes place directly through the interval $(0, \varepsilon_N)$ in the unbatched case and, in the batched case, by a combination of this direct shrinkage and the filtering, stretching, and aliasing of $p(f)$ caused by the batching. For additional discussion see [2].

To illustrate this insensitivity we now describe the application of the methods to unbatched data for the system response time sequences of Model A. Table 9 gives the results of the fixed run length experiments, and Table 10 gives the results of the sequential experiments. Comparing these tables with the results of the experiments on batched data (Tables 1 and 5) reveals no significant differences. This was our general experience with all the models.

5. Summary

In an earlier paper [2], we described a spectral method for generating confidence intervals from simulation output sequences and evaluated that method within the context

Table 9 Fixed sample size simulation results for response time process in Model A, unbatched case, $\mu = 41.2$.

Run length		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
500	Coverage	0.64	0.76	0.88	0.80	0.82	0.84
	Width	7.62	10.68	15.13	13.49	13.77	14.87
	Var (width)	1.40	3.67	16.30	22.77	23.57	28.71
750	Coverage	0.60	0.82	0.94	0.86	0.86	0.86
	Width	6.70	9.93	13.71	12.37	12.81	12.64
	Var (width)	0.73	3.48	12.64	15.97	16.98	13.77
1125	Coverage	0.64	0.80	0.96	0.90	0.92	0.90
	Width	6.04	9.40	12.45	11.20	11.55	10.86
	Var (width)	0.36	1.59	8.19	10.48	10.57	12.63
1687	Coverage	0.66	0.88	0.92	0.86	0.90	0.86
	Width	5.55	8.82	10.16	9.11	9.43	9.04
	Var (width)	0.26	1.77	6.14	4.17	4.78	5.15
2530	Coverage	0.72	0.86	0.86	0.84	0.86	0.84
	Width	5.18	7.30	7.68	7.11	7.19	7.27
	Var (width)	0.21	2.04	4.25	4.37	4.27	4.93
3795	Coverage	0.84	0.92	0.94	0.90	0.90	0.88
	Width	4.74	6.11	6.56	5.81	6.03	6.15
	Var (width)	0.19	0.89	2.08	2.10	2.19	2.99
5692	Coverage	0.82	0.90	0.88	0.86	0.86	0.88
	Width	4.15	4.93	5.33	4.64	4.88	5.11
	Var (width)	0.15	0.45	1.22	0.89	1.05	0.95
8538	Coverage	0.84	0.88	0.88	0.86	0.88	0.86
	Width	3.51	3.87	4.21	3.80	3.82	3.93
	Var (width)	0.08	0.30	1.04	0.72	0.70	1.25
12 807	Coverage	0.90	0.94	0.94	0.92	0.92	0.92
	Width	2.85	3.06	3.25	2.88	2.99	2.96
	Var (width)	0.04	0.23	0.51	0.11	0.13	0.66
13 500	Coverage	0.92	0.92	0.92	0.94	0.94	0.88
	Width	2.76	2.94	3.23	2.87	2.81	2.82
	Var (width)	0.03	0.20	0.46	0.18	0.17	0.52

Table 10 Sequential simulation results for response time process in Model A, unbatched case, $\mu = 41.2$.

Relative half-width		$d = 0$	$d = 1$	$d = 2$	*Sequential regression*	*Regression with cross validation*	*Splines with cross validation*
0.20	Coverage	0.64	0.76	0.88	0.78	0.78	0.80
	Run length	500	500	649	624	624	633
	Half-width	0.09	0.13	0.16	0.14	0.14	0.14
0.15	Coverage	0.64	0.76	0.88	0.74	0.74	0.78
	Run length	500	540	1124	781	870	896
	Half-width	0.09	0.12	0.13	0.12	0.12	0.12
0.10	Coverage	0.58	0.88	0.82	0.80	0.82	0.80
	Run length	553	2122	2691	2002	2231	2408
	Half-width	0.09	0.09	0.08	0.08	0.08	0.08
0.05	Coverage	0.80	0.86	0.88	0.84	0.84	0.80
	Run length	7011	9335	10 033	8156	8470	8866
	Half-width	0.05	0.04	0.04	0.04	0.04	0.04

of a run length control procedure. This method estimated the variance of the sample mean by estimating the spectral density at zero frequency, $p_B(0)$, of a sequence of batch means. This was accomplished by fitting a quadratic to the logarithm of the averaged periodogram.

This method worked well and was recommended as a solid practical procedure. However, there were two reasons to believe it could be improved upon by applying more flexible, adaptive curve fitting techniques. First, limiting the approximating function to a quadratic appeared somewhat restrictive for small batch sizes. Second, the spectrum $p_B(f)$ becomes smoother and is asymptotically flat as the sample and batch sizes increase. Thus, for large samples, a linear fit would yield an unbiased estimate of $p_B(0)$ with a much smaller variance than the one obtained by fitting a quadratic. This is analogous to an increase in the degrees of freedom in a t-confidence interval.

The present paper is an examination of this approach. Adaptive procedures of three basic types were evaluated: polynomial fits with the degree selected by sequential regression, polynomial fits with the degree selected by cross validation, and smoothing splines with the amount of smoothing determined by cross validation. In no case were we able to realize enough benefit to be able to recommend an adaptive procedure. In each case the process of adaptation created negative effects which either generated poorer performance than the quadratic method or reduced the potential payoff to a marginal point. The performance of the smoothing splines with cross validation was particularly disappointing since this method has flexibility, simplicity, and elegance.

Hence we still recommend the specific fixed quadratic method of [2]. More so than ever it appears as an effective, simple, and practical technique for simulation confidence interval generation and run length control. This method has been incorporated into the internal IBM system simulation analysis tools FIVE and SNAP/SHOT and is planned for installation in the internal IBM simulation tool RESQ. These simulators are described in [18], [19], and [20] respectively.

Acknowledgment

We would like to thank J. Boericke, R. Jensen, C. Sauer, and W. Skwish for their encouragement of this research and for their cooperation in seeing it to a practical fruition.

References

1. P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, Inc., New York, 1968.
2. P. Heidelberger and P. D. Welch, "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations," *Commun. ACM* **24,** 233–245 (1981).
3. D. R. Brillinger, *Time Series, Data Analysis and Theory*, Holt, Rinehart and Winston, Inc., New York, 1975.
4. M. S. Bartlett and D. G. Kendall, "The Statistical Analysis of Variance Hetereogeniety and the Logarithmic Transformation," *J. Roy. Statist. Soc. (Suppl.)* **8,** 128–138 (1946).
5. N. R. Draper and H. Smith, *Applied Regression Analysis, Second Edition*, John Wiley & Sons, Inc., New York, 1981.
6. T. W. Anderson, *The Statistical Analysis of Time Series*, John Wiley & Sons, Inc., New York, 1971.
7. D. M. Allen, "The Relationship Between Variable Selection and Data Augmentation and a Method of Prediction," *Technometrics* **16,** 125–127 (1974).
8. M. Stone, "Cross-validatory Choice and Assessment of Statistical Predictions," *J. Roy. Statist. Soc. Ser. B.* **36,** 111–147 (1974).
9. C. H. Reinsch, "Smoothing by Spline Functions," *Numer. Math.* **10,** 177–183 (1967).
10. G. Wahba and S. Wold, "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation," *Comm. Statist.* **4,** 1–17 (1975).
11. P. Craven and G. Wahba, "Smoothing Noisy Data with Spline Functions, Estimating the Correct Degree of Smoothing by the Method of Generalized Cross Validation," *Numer. Math.* **31,** 377–403 (1979).
12. G. Wahba, "Smoothing Noisy Data with Spline Functions," *Numer. Math.* **24,** 383–393 (1975).
13. G. Wahba, "A Survey of Some Smoothing Problems and the Method of Generalized Cross Validation for Solving Them," *Applications of Statistics*, P. R. Krishnaiah, Ed., North-Holland Publishing Company, New York, 1977, pp. 507–523.
14. G. Wahba and S. Wold, "Periodic Splines for Spectral Density Estimation: The Use of Cross Validation for Determining the Degree of Smoothing," *Comm. Statist.* **4,** 125–141 (1975).
15. C. de Boor, *A Practical Guide to Splines, Applied Mathematical Sciences*, Vol. 27, Springer-Verlag, New York, 1978.
16. L. W. Schruben, "A Coverage Function for Interval Estimators of Simulation Response," *Manage. Sci.* **26,** 18–27 (1980).
17. T. J. Schriber and R. W. Andrews, "A Conceptual Framework for Research in the Analysis of Simulation Output," *Commun. ACM.* **24,** 218–232 (1981).
18. H. C. Nguyen, A. Ockene, R. Revell, and W. J. Skwish, "The Role of Detailed Simulation in Capacity Planning," *IBM Syst. J.* **19,** 81–101 (1980).
19. H. M. Stewart, "Performance Analysis of Complex Communications Systems," *IBM Syst. J.* **18,** 356–373 (1979).
20. Charles H. Sauer, Edward A. MacNair, and Silvio Salza, "A Language for Extended Queuing Network Models," *IBM J. Res. Develop.* **24,** 747–755 (1980).

Received April 28, 1981; revised June 16, 1981

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.