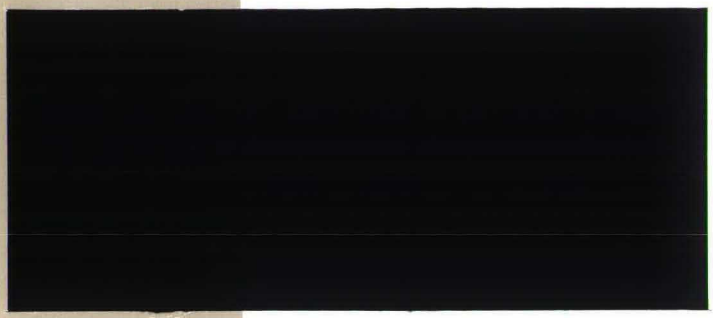
entER for omic Research

Discussion paper



Tilburg University

Center
for
Economic Research

No. 9668

IMPROVED DESIGN OF QUEUEING SIMULATION EXPERIMENTS WITH HIGHLY HETEROSCEDASTIC RESPONSES

By Russell C.H. Cheng and Jack P.C. Kleijnen

July 1996

ISSN 0924-7815

Improved Design of Queueing Simulation Experiments with Highly Heteroscedastic Responses

Russell C. H. Cheng

Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF, United Kingdom.

Fax: +44-1227-827932. E-mail: R.C.H.Cheng@ukc.ac.uk

Jack P. C. Kleijnen

Department of Information Systems and Auditing/CentER, School of Management and Economics, Tilburg University, Tilburg, 5000 LE Tilburg, Netherlands

Fax: +31-13-4663377. E-mail: kleijnen@kub.nl

May 1996

Improved Design of Queueing Simulation Experiments with Highly Heteroscedastic Responses

Russell C.H.Cheng

Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent, CT2 7NF, United Kingdom.

Fax: +44-1227-827932. E-mail: R.C.H.Cheng@ukc.ac.uk

Jack P. C. Kleijnen

CentER/Department of Information Systems and Auditing, School of Management and Economics, Tilburg University, Tilburg, 5000 LE Tilburg, Netherlands

Fax: +31-13-4663377. E-mail: kleijnen@kub.nl

Acknowledgements

The first author was partly financed by a grant from the Center for Economic Research, University of Tilburg. The second author was partly financed by the UK-Dutch Joint Scientific Research Programme of The British Council and the Netherlands Organization for Scientific Research (NWO). The authors wish to thank Dr. W.B. Liu of the University of Kent at Canterbury, for suggestions that simplified the proof of Theorem 1.

Abstract

Simulation experiments for analysing the steady-state behaviour of queueing systems over a range of traffic intensities are considered, and a procedure is presented for improving their design. In such simulations the mean and variance of the response output can increase dramatically with traffic intensity; the design has to be able to cope with this complication. A regression metamodel of the likely mean response is used consisting of two factors, namely a low-degree polynomial and a factor accounting for the exploding mean as the traffic intensity approaches its saturation. The best choice of traffic intensities at which to make simulation runs depends on the variability of the simulation output, and this variability is estimated using analytical heavy traffic results. The numbers of customers simulated at each traffic intensity are built up using a multistage procedure that systematically increases the efficiency of the simulation experiment. The asymptotic properties of the procedure are investigated theoretically.

The procedure is shown to be robust and to have improved efficiency compared with more naive procedures. A result of note is that even when the range of interest includes high traffic intensities, the highest traffic load simulated should remain well away from its upper limit; but the number of customers simulated should be concentrated at the higher traffic intensities used. Empirical results are included for simulations of a single server queue with different priority rules and for a complicated queueing network. These support the theoretical results, demonstrating that the proposed procedure can increase the accuracy of the estimated metamodel significantly compared with more naive methods.

Keywords: response surface, interpolation, congestion, polynomial regression, variance heterogeneity, runlength, sequentialization

1 Introduction

A procedure for improving the design of experiments (DOE) of steady-state simulations of queueing systems is presented. The objective is to better estimate regression metamodels of the system when the traffic intensity is allowed to vary over a range of values including those approaching saturation. Congested queues are known to be hard to simulate accurately, because both the mean and the variance of the steady-state output typically become unbounded as the traffic intensity increases to saturation. Whitt (1989) and Asmussen (1992) discuss such steady-state behaviour; and Whitt in particular considers how the length of a simulation run should be increased in order to maintain accuracy in estimating queue length or waiting time. In this paper we consider a comprehensive approach to determine which traffic rate intensities to simulate and how many customers to simulate at each intensity, particularly when a range of intensities is of interest. Though we consider only steady-state simulations, our method can be applied to terminating simulations as well.

To estimate the mean response of interest, the procedure uses a regression metamodel consisting of two factors: a low degree polynomial and a factor accounting for the unbounded behaviour of the mean response as the traffic intensity approaches saturation. The best choice of traffic intensities at which to simulate depends on the variance of the response; and the proposed method uses analytic heavy traffic results to guide this choice. The best choice of the number of customers to simulate at each intensity then depends on the true variance of the output at the selected traffic intensities. These variances are increasingly more accurately determined using a multistage procedure that progressively improves the variance estimates

whilst simultaneously increasing the run allocations at each of the select traffic intensities in an efficient way. Thus this multistage procedure corrects any errors made in using heavy traffic theory to estimate the response variance. Once the simulation runs are completed, backwards elimination is used to adjust the degree of the polynomial to ensure that the metamodel is a good fit to the observed output. The procedure is studied analytically and shown to have asymptotic optimal properties.

Empirical results are included from simulation experiments of the M/M/1/FIFO queue, two priority variations: M/M/1/SPT (shortest processing time first) and M/M/1/LPT (longest processing time first), and a complicated queueing network (involving a system of terminals with some unusual features). These results show that the procedure leads to significant improvement in accuracy of the fitted metamodel compared with naive designs that, for instance, use evenly spread values of traffic intensity with equal numbers of customers simulated at each intensity.

The theoretical results, corroborated by the empirical results, also show that the method is robust; the bias error, when the assumed form for the unbounded behaviour of the mean is incorrect, being made negligible through fitting of the low-order polynomial factor.

So far the literature has paid relatively little attention to the topic of this paper. In the simulation area, DOE has focussed on classic designs such as 2^{k-p} and central composite designs, possibly combined with variance reduction techniques (see Kleijnen 1987, 1992). Regression models have been considered: by Reiman, Simon and Willie (1992), who show how known theoretical results about light traffic and heavy traffic behaviour can be incorporated into the analysis; by Cheng (1990) and Cheng and Traylor (1993), who show how conditional sampling and use of known theoretical results can be combined; and by Vollebregt (1996), who investigates a problem similar to our problem but not with heavy traffic queueing.

Outside the simulation field, optimal design in regression analysis is discussed in the seminal paper, Kiefer and Wolfowitz (1959) (to whom Reiman et al. (1992) also refer), and also by Fedorov (1972) and more recently by Atkinson and Donev(1992), Ermakov and Melas(1995) and Pukelsheim (1993). However, the focus in this literature is on the case where the response variable has constant variance.

The paper is organised as follows. In Section 2 we introduce the regression metamodel to be considered. In Section 3 we give theoretical results on how to optimally design the simulation experiment and describe the multistage procedure for implementing the runs in practice. In Section 4 we discuss the method for fitting the regression metamodel, once the results have been obtained. In Section 5 we discuss the amount of computing effort needed to make a simulation

run. This effort obviously needs to be taken into account in assessing overall efficiency, and is one that needs clarification if our suggested method is to be applied properly. Section 6 gives a summary of the proposed procedure for ease of reference for the practitioner. In Section 7 we give a number of examples and compare our analytical results with simulation results. These examples show the good agreement between the two, and illustrate the substantial efficiency gains possible with our approach. Section 8 gives conclusions.

2 The Regression Metamodel

We suppose that the simulation experiment is made up of a number of independent runs. We assume that y, the output (response) of a run, is determined by x, an independent input variable, and that this input/output relationship can be represented by the following regression metamodel (response surface):

$$y_{ij} = \eta(x_i, \beta) + \epsilon_{ij} \quad (i = 1, ..., n) \quad (j = 1, ..., m_i) \tag{1}$$

with

$\eta(x, \beta) = (\beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k) f(x)$: the mean of y;

ϵ_i: approximation error of the metamodel, with mean 0 and variance σ_i^2;

$\beta = (\beta_1, ..., \beta_k)'$: vector of k unknown parameters representing input effects;

$f(x)$: a known function (discussed in the next paragraph).

We make the simulation runs at only n distinct input values $x_1, x_2, ..., x_n$, with m_i observations (replications) placed at the ith point, x_i $(i = 1, ..., n)$. We shall be considering how best to choose the x_i and the m_i, and therefore call the x_i, the *design points*. We write $\mathbf{x} = (x_1, x_2, ..., x_n)$, $\mathbf{m} = (m_1, m_2, ..., m_n)$, and denote the total number of runs by N; that is, $\sum_{i=1}^{n} m_i = N$.

The purpose of introducing the factor $f(x)$ is to allow regression models that have unbounded responses; in particular $f(x)$ allows saturated queueing situations. Consider, for example, the M/M/1 queue. Suppose the arrival rate is unity. Then the steady-state expected waiting time has the form

$$E(y) = x/(1 - x) \tag{2}$$

where x is the traffic intensity. If we do not know the correct expression for $E(y)$ but know only that queue saturation occurs as $x \to 1$, then we may assume

$$y_i = (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)/(1 - x_i) + \epsilon_i \tag{3}$$

where the vector of parameters $\beta = (\beta_0, \beta_1, \beta_2)'$ is unknown and is to be estimated. This model is of the form (1).

Another example is the problem discussed by Reiman et al. (1992) where the expected squared waiting time $E(y^2)$ is to be estimated in an M/G/1 queue and the service time has a certain mixture distribution. They show that

$$E(y^2) = (\frac{5}{2} + x + \frac{3}{8}x^2)/(1-x)^2 \tag{4}$$

where x is again the traffic intensity. So in this case $f(x)$ in (1) is $(1-x)^{-2}$.

We must consider the variance of the response as well as its mean. We do not assume that the error variance is constant; it may depend on x_i. For example, Whitt (1989) and Asmussen (1992) show that the variance of ϵ in (1), for the M/G/1 queue, is $O[(1-x)^{-4}]$ as $x \to 1$. We shall assume first that the form of the dependence is known, that is:

$$Var(\epsilon) = [g(x)\sigma]^2 \tag{5}$$

where $g(x)$ is known, but σ is not. In this case, as pointed out by Kiefer and Wolfowitz (1959), homogeneity of variance can be restored simply by dividing (1) by $g(x)$, which gives

$$\begin{aligned} z_{ij} &= y_{ij}/g(x_i) \\ &= (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k) r(x_i) + \delta_{ij} \end{aligned} \tag{6}$$

where

$$r(x_i) = f(x_i)/g(x_i). \tag{7}$$

Then $Var(\delta_{ij}) = \sigma^2$, a constant, independent of x.

We shall suppose that we are interested in the behaviour of the queue over a range of x values: $x \in [x_L, x_U]$, where x_U may be close to unity - the near saturation case, and that the objective of the simulation is to estimate $\eta(x, \beta)$ over this range of x values. We use the obvious estimator:

$$\hat{\eta}(x, \beta) = \eta(x, \hat{\beta}) \tag{8}$$

where $\hat{\beta}$ is the ordinary least squares (OLS) estimator:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}. \tag{9}$$

Here $\mathbf{z} = (z_{11}, z_{12}, ..., z_{nm_n})'$ is the vector of standardized simulation responses (of dimension N), and $\mathbf{X}$ is the $N \times (k+1)$ matrix of independent variables, with the row vector

$$\mathbf{x_i} = (1, x_i, x_i^2, ..., x_i^k) r(x_i)$$

appearing m_i times in $\mathbf{X}$. The matrix $\mathbf{X}'\mathbf{X}$ is persymmetric (that is, all elements in any diagonal perpendicular to the main diagonal are the same; see, for example, Aitken, 1964, p.121), its (i,j)th entry being

$$(\mathbf{X}'\mathbf{X})_{ij} = \sum_{l=1}^{n} n_l \, x_l^{i+j} \quad (i,j = 1,2,...,k) \tag{10}$$

where $n_i = m_i r^2(x_i)$, for $i = 1, 2, ..., n_i$.

The discussion so far assumes that the design points x_i and the number of runs m_i made at each point, are prescribed. We now consider how best to choose the x_i and m_i, where the objective of our DOE is to optimize some measure of the variance of $\eta(x, \hat{\beta})$. A natural measure is the average weighted variance over the range of interest:

$$p = \int_{x_L}^{x_U} w(x) Var[\eta(x, \hat{\beta})] dx \Big/ \int_{x_L}^{x_U} w(x) dx \tag{11}$$

with weight function $w(x)$. The simplest case is $w(x) = 1$ for all $x \in [x_L, x_U]$. An alternative has all the weight concentrated at one point, $x_0 \in [x_L, x_U]$, when p in (11) reduces to

$$p = Var[\eta(x_0, \hat{\beta})]. \tag{12}$$

This latter alternative might not appear appropriate when a range of x values is being considered. It turns out that the choice of $\mathbf{x}$ and $\mathbf{m}$ is not all that sensitive to the precise form of p. Hence, if x_0 is a typical value in the range of interest, then (12) is an adequate measure to use. An interesting case is when we wish to focus on the behaviour of the queue as $x \to 1$. The factor in $\eta(x, \beta)$ that depends explicitly on β is $\beta_0 + \beta_1 x + ... + \beta_k x^k$, and this will be close in value to $\beta_0 + \beta_1 + ... + \beta_k$ when x is close to unity. A simple convenient performance measure is thus

$$Var[\lim_{x \to 1}(\hat{\beta}_0 + \hat{\beta}_1 x + ... + \hat{\beta}_k x^k)] = Var(\hat{\beta}_0 + \hat{\beta}_1 + ... + \hat{\beta}_k). \tag{13}$$

In situations where $Var[\eta(x, \hat{\beta})]$ becomes infinite as $x \to 1$, the average as given by (11) is dominated by values of the integrand close to the upper limit x_U, especially when x_U is close to unity. Optimizing either (11) or (13) then gives similar values for both the x_i and the m_i. We give numerical examples later.

We discuss the calculation of n, x_i and m_i in the next section.

3 Optimal Selection of Simulation Inputs

3.1 Error Variance of Known Form

We consider first the situation where the form of the dependence of the error variance on x is known, i.e. $g(x)$ in (5) is known.

Kiefer and Wolfowitz (1959) address a general version of the optimum design problem in regression analysis. They point out that, in general, a direct approach to the problem leads rapidly to intractable algebra. They give an ingenious approach that transfers the difficulty into a Chebyshev approximation problem, thus allowing use of the extensive literature on this topic. They give examples drawn from polynomial regression. Also see Ermakov and Melas (1995).

It is in principle possible to make use of the Kiefer and Wolfowitz approach in our problem. However, there are two aspects that make an alternative, direct, approach competitive in our case. Firstly, it turns out that the persymmetric form of the matrix $\mathbf{X}'\mathbf{X}$ makes the direct approach much more tractable. Secondly, the form of our regression function, being non-standard, gives rise to a non-standard Chebyshev approximation that is arguably no easier, and is possibly more difficult, to solve than the problem resulting from our direct approach, except in specific, very simple cases, which happen to be analytically tractable.

We start with the optimal choice of n (the number of distinct x values). Kiefer and Wolfowitz (1959, Theorem 2) show that if the functions multiplying the parameters $\beta_0, ..., \beta_k$ are linearly independent, then n should be chosen equal to the number of unknowns, $k+1$ in our case. Thus the optimal choice is $n = k+1$, and we assume that this holds for the remainder of this section.

We now investigate how to choose the m_i (the number of replications at x_i, with $i = 1, 2, ..., n$). Let $\mathbf{V}$ be the so-called Vandermonde matrix:

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ & & \dots & \\ x_1^k & x_2^k & \dots & x_n^k \end{bmatrix},$$

$|\mathbf{V}|$ its determinant, and $|\mathbf{V}_{st}|$ the cofactor of its (s,t)th element. We have the following theorem.

Theorem 1: Let $\mathbf{A} = \mathbf{X}'\mathbf{X}$ be as defined in (10). Then the inverse of $\mathbf{A}$ has the form:

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \mathbf{B}^{(i)}/[m_i r^2(x_i)] \tag{14}$$

where the matrices $\mathbf{B}^{(i)}$ have (s,t)th element

$$(\mathbf{B}^{(i)})_{st} = \frac{|\mathbf{V}_{is}||\mathbf{V}_{it}|}{|\mathbf{V}||\mathbf{V}|}, \quad s,t = 1,2,...,n,$$

which are independent of the m_i and the form of $f(.)$. □

Proof: See Appendix.

Corollary: The average weighted variance of $\hat{\eta}$, defined in (11), reduces to

$$p(\mathbf{x},\mathbf{m}) = \sigma^2 \sum_{i=1}^{n} \left(\frac{a_i}{r_i}\right)^2 m_i^{-1} \tag{15}$$

with

$$a_i = \sqrt{trace(\mathbf{B}^{(i)}\mathbf{C})}, \quad i = 1,2,...,n, \tag{16}$$

where $\mathbf{C}$ is the matrix whose (i,j)th element is

$$(\mathbf{C})_{ij} = \int_{x_L}^{x_U} w(x) f^2(x) x^{i+j-2} dx, \quad i,j = 1,2,...,n. \tag{17}$$

In (15), for simplicity, we have written r_i for $r(x_i)$.

Theorem 1 can be used to derive the optimal m_i that will minimize (15), given x_i. It will be convenient to write

$$m_i = \pi_i N \quad i = 1,2,...,n, \tag{18}$$

so that π_i is the proportion of runs made at x_i. We write $\pi = (\pi_1, \pi_2, ..., \pi_n)'$. We have the following theorem.

Theorem 2: Let the design point $\mathbf{x}$ be fixed, and let the total number of simulations be equal to N, so that $\sum_{i=1}^{n} m_i = N$. Then the variance (15) is minimized if

$$\pi_i = \frac{a_i/r_i}{(\sum_{j=1}^{k+1} a_j/r_j)}. \tag{19}$$

The minimized value is

$$p(\mathbf{x}) = \frac{\sigma^2}{N}[\sum_{j=1}^{k+1}(a_j/f_j)]^2, \tag{20}$$

where $f_j = f(x_j)$.

Proof: Minimization of (15) subject to $m_i > 0$ for all i, and $\sum_{i=1}^{n} m_i = N$, is a convex programming problem. An easy application of the method of Lagrange multipliers gives the result. □

We have ignored the implicit constraint that the m_i have to be integers. Often in practice we can assume that the m_i are so large that rounding to an integer makes a negligible difference. However there is a more subtle reason why rounding is not an issue. Though it has been more natural in the regression context to formulate the problem in terms of number of runs of given length, in fact it is the allocations of total computational effort at each design point, measured (say) in terms of the number of customers simulated, that comprise the more fundamental design variables. An integer constraint on the (large) number of customers simulated at each design point is clearly of no concern. Thus if rounding in terms of given run lengths is a problem, the run lengths themselves can always be adjusted to ensure that the m_i are integers. We shall discuss this more fully in Section 5.

In simple cases, Theorem 2 is sufficient to enable the optimal settings of the x_i to be found explicitly by solving the system of equations

$$\partial p(\mathbf{x})/\partial x_i = 0, \ \ (i = 1, 2, ..., n).$$

However it can happen that the smallest design point value, x_1, is located at the limit of the allowable range of values, so it will not correspond to where $\partial p(\mathbf{x})/dx_1 = 0$. An alternative, which we consider preferable because it applies even to complicated cases, is to numerically minimize (20) using a standard direct search algorithm. We have found that the Nelder-Mead algorithm is perfectly satisfactory (see Nash, 1979, for example). The formulas are in a form particularly suited to numerical optimization. The calculation of p involves a potentially expensive numerical quadrature. This expense is minimized by confining the quadrature to the calculation of the matrix C, which does not involve the design points x_i. Thus, as far as the numerical minimization is concerned, C is a constant that needs be calculated only once at the start of the minimization.

3.2 Error Variance of Unknown Form

In practice $g(x)$ in (5) will not be known, or will only be known approximately. It is possible in principle to develop a procedure which will allow the x_i and m_i to be estimated increasingly more accurately as N, the total number of runs, is increased. Such a procedure will of necessity be elaborate. In our experience, identifying and having the correct design points is of secondary importance compared with having the correct number of runs at each point. This is especially so because the range of x values of interest is often dictated by the practical context of the problem so that the choice of design points is restricted. We shall assume that an approximation,

$g_a(x)$, for $g(x)$ is available from asymptotic theory or other considerations. For example, as has already been pointed out, $Var(\epsilon)$ in (3) is $O[(1-x)^{-4}]$ as $x \to 1$. Moreover $Var(\epsilon) \to 0$, as $x \to 0$. Thus we could take $g_a(x) = x(1-x)^{-4}$ for this case.

We now suggest the following multistage strategy, which fixes the design points approximately, and then adopts a method of allocating runs which converges, as the N increases, to the correct optimal proportions for these given design point settings. Kleijnen and Van Groenendaal (1995) have discussed the method in detail. We give below a full theoretical justification for the technique. The method is as follows.

1. Let p, the performance criterion to be optimized, be as given in (11) or (13). The number of design points is set at $n = k+1$, where k is the highest degree of the polynomial considered necessary to guarantee a good fit for the regression model. Use a variance approximation $g_a(x)$ for $g(x)$ in (5) to calculate the optimal choice for $\mathbf{x}$, at least according to this approximation, by numerically minimizing $p(\mathbf{x})$ in (20). The design points are now fixed at these values.

2. Let N be the total number of runs to be made. Initially make m_0 pilot runs at each x_i, with $nm_0 << N$. This yields initial sample variances

$$s_i^2 = \sum_{l=1}^{m_0} [y_{il} - \bar{y}_i(m_0)]^2/(m_0 - 1),$$

which estimate the true variances $[g(x_i)\sigma]^2$. The number of initial runs is not critical, and in the examples given below it was set at approximately 30% of the total number of anticipated runs.

3. Additional runs are now assigned and made at each x_i in a multistage way. Each design point is considered in turn in a fixed order, $1, 2, ..., n$, say, cycling repeatedly through this fixed sequence. The number of runs, $m_{i,j+1}$, carried out at x_i in the $(j+1)$th cycle is calculated as follows. Suppose that, after j cycles, M_{ij} runs have been made at x_i. Then the updated sample variances

$$s_{ij}^2 = \sum_{l=1}^{M_{ij}} [y_{il} - \bar{y}_i(M_{ij})]^2/(M_{ij} - 1)$$

better estimate the true variances $[g(x_i)\sigma]^2$ than did the initial estimates s_i^2. These updates are used in $\tilde{r}_{ij} = f_i/s_{ij}$ to estimate π_i, the optimal loading at x_i, defined in (19):

$$\hat{\pi}_{ij} = \frac{a_i/\tilde{r}_{ij}}{(\sum_{l=1}^{k+1} a_j/\tilde{r}_{lj})}. \tag{21}$$

Set

$$m_{i,j+1} = \max\{\lfloor (j+1)\hat{\pi}_{ij} - M_{ij} \rfloor,\ 0\} \tag{22}$$

where $\lfloor . \rfloor$ denotes "integer part of", so that the total number of runs made at x_i after $(j+1)$ cycles is

$$M_{i,j+1} = \max_{l \leq j} \{ \lfloor (l+1)\hat{\pi}_{i,l} \rfloor \}.$$

The rationale for this procedure is that, given the current estimates of the optimal allocations in (21), now $(j+1)\hat{\pi}_{ij}$ is the number of runs that should be allocated to point x_i. However, because runs have already been made in previous cycles, this updated desired number is not always achievable, and $M_{i,j+1}$ is the closest we can get to the desired value.

4. The process is continued until the required total number of runs has been made, i.e. $\sum_{j=1}^{n} M_j = N$.

The following Theorem shows that as $N \to \infty$, the process makes an allocation where the proportions converge, almost surely, to the correct optimal values.

Theorem 3 Let π_i denote the (unknown) optimal proportions (19), which is realsies (20), the $p(\mathbf{x})$ value that minimizes $p(\mathbf{x}, \mathbf{m})$ for given $\mathbf{x}$. Let M_i be the number of runs made at x_i by the end of the process, when a total of N runs has been made. Then, if $N \to \infty$,

$$\hat{\pi}_{ij} \to \pi_i \ \ and \ \ M_i / \sum_{l=1}^{n} M_l \to \pi_i \ \ \ a.s. \ \ for \ all \ i. \tag{23}$$

Proof: See Appendix.

As already remarked, our procedure is not fully optimal if the initial variance approximation $g_a(x)$ is different from $g(x)$, since the method does not update the values of the design point. For this reason the design resulting from this procedure is optimal only with regard to the given design points, rather than fully optimal. However, the right choice of design point values is a less serious problem compared with the right choice of run allocations, and this is what the method focuses on. We shall give numerical examples in Section 5.

4 Specifying the Polynomial Regression Metamodel

Apart from selecting the x_i and m_i, the other main practical issue is the fitting of the regression function $\eta_k(x, \beta)$ to the simulation output, where we introduce the subscript k to denote the degree of the polynomial factor in η. The main issue discussed in this section is the choice k.

Our basic assumption is that k can be selected sufficiently large, $k = k_0$ say, for the bias error to be negligible compared with the random error, when the model $\eta_{k_0}(x, \beta)$ is fitted. A question that arises is when this assumption will be valid. The main problem is that we allow $f(x)$ to become unbounded. For example, if it is assumed that $f(x) = (1-x)^{-1}$, and this is incorrect, then the assumption will not be valid as $x \to \infty$. The simplest solution is to restrict x

to a range where the regression function is smooth and bounded. In particular, suppose $\eta_T(x)$ is the unknown true regression function, that it continuous and hence bounded in the interval $[x_L, x_U]$, and that $f(x)$ is any continuous strictly positive function in the same interval $[x_L, x_U]$. Then if

$$\eta(x, \beta) = (\beta_0 + \beta_1 x + ... + \beta_k x^k) f(x),$$

it follows from the Weierstrass approximation theorem (see Johnson and Riess, 1982, for example) that for any $\epsilon > 0$, k and β can be found for which

$$\max_{x \epsilon [x_L, x_U]} |\eta(x, \beta) - \eta_T(x)| < \epsilon.$$

In other words, a polynomial can be found which will allow $\eta(x, \beta)$ to approximate $\eta_T(x)$ to arbitrary accuracy. From now on, we shall assume that the range of x can be restricted in this way.

It is of interest to see if a polynomial of degree less than k_0 can be used. Seber (1977) has reviewed methods for checking this. We adopt Seber's recommendation of using backwards selection, where the models $\eta_k(x, \beta)$ for $k = k_0, k_0 - 1, ...$ are successively fitted. At each stage, after fitting $\eta_k(x, \beta)$, the highest remaining parameter β_k is tested to see if it can be assumed to be zero. The procedure stops at the first k $(= k'$, say) for which β_k is significantly non-zero and the degree of the polynomial is then taken to be k'.

Assuming for the moment $Var(\epsilon) = [\sigma g(x)]^2$ to be known, it is simplest to use the OLS formula (9) for $\hat{\beta}$, based on the normalized observations z_{ij}. We can write the residual sum of squares, when standardized for the variance, as

$$RSS_k = \sum_{i=1}^{k+1} \sum_{j=1}^{m_i} \left(\frac{y_{ij} - \eta_k(x_i, \hat{\beta})}{\sigma g(x_i)} \right)^2. \tag{24}$$

To test if $\beta_k = 0$, we can use (see Seber, Chapter 4)

$$d_k^2 = RSS_k - RSS_{k-1}. \tag{25}$$

When the errors are normally distributed, then d_k^2 has the chi-squared distribution with one degree of freedom: χ_1^2 . The hypothesis that $\beta_k = 0$ can then be rejected at the $(1 - \alpha)100$ percent level if

$$d_k^2 > \chi_1^2(\alpha) \tag{26}$$

where $\chi_1^2(\alpha)$ is the upper $1 - \alpha$ quantile of the χ_1^2 distribution. We set the α level at 5% in the examples given later.

When $Var(\epsilon)$ is not known, then σg_i can be replaced with the estimated standard deviation of the observations at x_i :

$$s_i = \left[\sum_{j=1}^{m_i} [y_{ij} - \bar{y}_i(m_i)]^2/(m_i - 1) \right]^{1/2} . \tag{27}$$

As $s_i \to \sigma g_i$ almost surely as $m_i \to \infty$, we have that d_k^2 is asymptotically χ_1^2 as $m_i \to \infty$ for all i. A similar adjustment applies to the estimation of β. So we still use (9), but now the y_{ij} are rescaled using s_i rather than g_i so that z_{ij} is calculated as $z_{ij} = y_{ij}/s_i$ rather than as in (6). The procedure has been suggested by Kleijnen and van Groenendaal (1995) who call it *estimated weighted least squares*.

Of course the usual caveat applies: the overall level of confidence decreases if the test is applied sequentially to several different β. This is a well-known situation which we do not discuss further here.

A question that arises is whether x_i and m_i, which are selected based on $k = k_0$, will be satisfactory for $k = k' \neq k_0$. We shall give some numerical examples which indicate that there appears to be relatively little loss of efficiency (also see Atkinson and Donev (1994)).

5 Translating Regression Variables into Simulation Variables

Without loss of generality we suppose that it takes one unit of computer time to generate and process one simulated customer. We let b_i be the length of each run at design point x_i, as measured in such units. The cost involved is then directly measured in run length and there is no ambiguity whether we talk of the number of customers processed or of run length.

So far it has been most natural, because we have focused on regression analysis aspects, to treat the m_i as being the decision variables, and the response from a run as the basic unit of observation; the total number of runs $N = \Sigma m_i$ being fixed. However, a more insightful view is to consider as decision variables, not the m_i, but

$$c_i = b_i m_i ,$$

which gives the total computing effort expended at design point x_i. Let the total available computer time be C units. Then it is simplest to regard C as being fixed, with selection of the c_i as being subject to $\sum c_i = C$, which therefore replaces the original condition $\sum m_i = N$. In the case of steady-state simulation it is well known that if the response is some form of sample average, then its variance is $O(b_i^{-1})$ as $b_i \to \infty$. Thus, provided the b_i are sufficiently large,

we can write (5) in the form

$$Var(\epsilon_{ij}) \simeq [g(x_i)\sigma]^2/b_i \tag{28}$$

where $g(x_i)$ does not depend on m_i or b_i. Then (5), (7) and (28) yield $r(x_i) = f(x_i)b_i^{\frac{1}{2}}/g(x_i)$. Defining $t_i = f(x_i)/g(x_i)$, (15) becomes

$$p(\mathbf{x}, \mathbf{m}) = \sigma^2 \sum_{i=1}^{k+1} \left(\frac{a_i}{t_i}\right)^2 (b_i m_i)^{-1}. \tag{29}$$

We see therefore that minimization of (29) subject to $\sum b_i m_i = C$, is precisely the same problem as the original, except that $b_i m_i$ replaces m_i, and t_i replaces r_i. The solution is therefore analogous to (19):

$$b_i m_i = \frac{a_i/t_i}{(\sum_{j=1}^{k+1} a_j/t_j)} C \quad i = 1, 2, ..., k+1. \tag{30}$$

Thus we have the result that the only requirement is that $b_i m_i$, the total run length at the point x_i, should be as given by (30). We are free to choose either the value of b_i (that is, how long to make each run), or to choose the value of m_i (that is, how many runs to make at the point x_i). Whichever value is chosen, the other is given by (30). Consequently the total number of observations N can be regarded as either fixed or variable, as we like. For example, if N is given, we can choose any set of m_i satisfying $\sum m_i = N$, and then fix the run lengths using (30). The only proviso is that m_i should not be so large that it makes b_i too small, because (28), which is an asymptotic result, might then no longer hold. The number of runs N is thus in effect arbitrary, so we could in principle set $N = k + 1$, and make a single run at each point: $m_i = 1$ for all i. However, we would then need to use batching or spectral analysis, say, to assess lack of fit. Conversely, making the m_i too large would mean making very many short runs. In general this is a bad idea, as each run will require a setup time to reach steady state, and it is usually more efficient to make one long run (see Cheng, 1976). In our numerical examples we have chosen b_i proportional to $g_a(x_i)$ so that runs give outputs of roughly equal variance. Additionally the b_i are sufficiently large so that initialization bias is not a serious problem, but at the same time not so large that they make m_i so small that making them integer seriously affects their optimal setting.

6 Intermezzo: Summary of the Method

For convenience we set out the full suggested procedure for fitting the regression metamodel (1):

1. Select: (i) the bounding factor $f(x)$ in (1); (ii) the run lengths b_i in (28); (iii) the variance function $g(x)$ in (5); (iv) the total number of runs N, below (1); (v) the performance index p in (11); (vi) k_0, the maximum degree of the regression polynomial factor; see (1) and Section 4.

2. Find the optimal design points $\mathbf{x}$, by numerically minimizing p in (20). The Nelder-Mead simplex search is suggested.

3. Make the N simulation runs using the sequential procedure described in Steps 2 and 3 of Section 3.2.

4. Progressively fit the regression metamodels $\eta_k(x, \beta)$ to the simulation results, for $k = k_0$, $k_0 - 1, ...,$ using backwards selection based on (26) to select the final fitted model.

7 Examples: Numerical and Simulation Results

We give examples to illustrate the points discussed in previous sections.

7.1 Sensitivity of the Performance Measure to x and m: M/M/1 Queue Example

7.1.1 Exact numerical Results

Firstly we consider the sensitivity of the performance measure p to the choice of $\mathbf{x}$ and $\mathbf{m}$. We give numerical results based on the theory of Section 3, and also simulation results.We take as our example the M/M/1 queue with the objective of estimating the waiting time response curve (3) in the range $[0.5, 0.95]$. In this case, $(1-x)^{-1}$, the natural choice for $f(x)$ in (1), happens to be the correct one. This form for $f(x)$ occurs in many other queues, so in this regard the M/M/1 queue is a good typical example.

Even in this elementary case, $g(x)$ cannot be given explicitly. We take $g(x) \simeq x^{1/2}(1-x)^{-2}$, the form suggested in Section 3.2. It should be emphasised that our choice here of $g(x)$ is made mainly for illustration and other choices are possible. For the performance criterion, we use the average variance (11), with weight $w(x) = 1$.

The optimum design values are obtained by minimizing p given in (20). For the case where the end points are allowed to vary, the chosen form of $g(x)$ is already too complicated to allow a tractable solution except for the case where $n = 2$. (The result for this case is $x_1 = 0$ and $x_2 = 1/3$.) It is generally much simpler to use numerical optimisation directly on (20). We used the well-known Nelder-Mead simplex search modified to satisfy the constraints $0.5 < x_1 < x_2 < ... < x_n < 0.95$. The modification is as follows. At any step, the components of $\mathbf{x}$ are updated one at a time. If the change to any component would result in breaking a constraint

then we reduce the step length so that the distance to the constraint boundary is halved. For example if the jth step value of x_i required by the original algorithm is $x_i^{(j)} = x_i^{(j-1)} - \delta$, and $x_i^{(j)} < x_{i-1}^{(j-1)}$,breaking the constraint, then we simply set $x_i^{(j)} = (x_i^{(j-1)} + x_{i-1}^{(j-1)})/2$. For initial values of the x_i we used values evenly spread in the upper half of the interval, and encountered no difficulties in locating the optimum.

Table 1 gives the optimal design points, $\mathbf{x}$, and design allocations, π, for $n = 2$ through 6, for both the case where the end design points are fixed and where they are allowed to vary. The values Np/σ^2 are the performance index calculated at these optimal design points and loadings, and then scaled by N/σ^2. The values Np_E/σ^2 are where the optimised design points are still used, but the allocations are all set equal (i.e. $\pi_i = n^{-1}$).

There are several points of interest:

(i) The optimal loadings heavily favour the end where the variability is highest.

(ii) There is a large improvement in the performance index as one goes from design points evenly spread with equal allocationss, to design points and allocations that are optimally adjusted. For example in the case $n = 3$, Np/σ^2 decreases from $22,443$ to $1,273$.

(iii) When the endpoints are allowed to vary optimally, the design points are located well away from the upper endpoint. Nevertheless, this is perhaps of theoretical interest only, because, unless there is considerable certainty concerning the precise form for $f(x)$, it is inadvisable to use design points that do not cover the range of interest. Otherwise extrapolation, with its well-known attendant risks, would be needed to estimate the response outside the experimental range.

(iv) Still in the case where the endpoints are allowed to vary, the lower end point is located at zero. This is outside the range over which the performance index (viz. the average variance of the estimator) is obtained, . Moreover, the true average waiting time, the estimated average waiting time and its variance are all identically zero at $x_1 = 0$. The design allocation reflects this by being zero also. This has the interpretation that the known (zero) value of the regression function at $x_1 = 0$ should be used in fitting the model, but there is evidently no need to make any simulation runs at this design point.

(v) In the case where the endpoints are fixed, the minimised variance does not increase monotonically as n increases. However, the minimised value remains reassuringly stable, especially in the case where the other points are optimally chosen.

(vi) In Table 1b the points are equally spaced and the allocations are optimized subject to these design point values. The resulting allocations and optimized Np/σ^2 values are similar to

those of Table 1a. For example when $n = 4$, if we use fixed endpoints, but optimized design points and loadings otherwise, then $Np/\sigma^2 = 9,680$. If the design points are equally spread out, but we use optimised loadings at these values, then Np/σ^2 increases to only 9,929. If, instead, optimised design points are used with evenly spread loadings, then the performance index jumps to $Np_E/\sigma^2 = 23,900$. If evenly spread design points are used as well as equal loadings then $Np_E/\sigma^2 = 24,739$. This supports our view that priority should be attached to optimizing the loadings over optimizing the design point values.

The last column in Table 1a and Table 1c shows the optimal settings using the performance criterion based on the variance (13). However, the corresponding values of Np/σ^2 and Np_E/σ^2 are for the original criterion, (11), where this has been calculated at the settings of that column. For example, in the case of fixed endpoints (Table 1a), the value of (11) using the design settings based on (13) is $p \doteq 11,441$. This is only 10% more than the minimum value of 10,383 when its own settings are used. This indicates that we do not lose much efficiency when using the simpler criterion (13) to obtain the design points and allocations.

Insert Table 1 *about here*

7.1.2 Simulation Results

To test the above (exact) numerical results we carried out some simulations. To simplify the discussion, we consider

$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 + ... + \hat{\beta}_k, \tag{31}$$

which estimates $\eta(x, \beta)$ as $x \to 1$; its use simplifies the simulation experiments and the discussion of the results without altering our broad conclusions. So we shall use the performance criterion (13).

As in the numerical calculations we set $f(x) = (1-x)^{-1}$ and $g(x) = x^{1/2}(1-x)^{-2}$. In this example, we present the 'best case scenario'; that is, (i) n is regarded as fixed; (ii) the selected $f(x)$ is assumed to be correct so that no model fitting is done, and (iii) $g(x)$ is assumed to be correct, so that the sequential procedure for updating the variance estimates at the different design points is not implemented. The optimal value of n is $n = 2$, the optimal input values and allocations (given previously in Table 1c) being $x_1^* = 0$, $x_2^* = 1/3$ and $\pi_1^* = 0$, $\pi_2^* = 1$. The value for Np/σ^2 is then 208. If $n = 3$ is assumed, then the optimal input values and allocations are $x_1^* = 0$, $x_2^* = 0.107$, $x_2^* = .721$ and $\pi_1^* = 0$, $\pi_2^* = 0.16$, $\pi_2^* = 0.84$ (Table 1c) and the corresponding value for Np/σ^2 is 1273. If, instead, the design points are evenly spread over

the interval [0.5, 0.95] and the allocations are all equal, then Np/σ^2 increases to 22,091 when $n = 2$, and to 23,391 when $n = 3$ (Table 1b).

Simulations were carried out using both optimised and equally spread input values and allocations. The results for the case $n = 2$ are given in Table 2a, and those for $n = 3$ in Table 2b. The underlying true expected waiting time being estimated is given by (2), so the correct parameter values are $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = 0$. Thus consistent estimators of the coefficients are obtained by using either $n = 2$ or 3, and the true value of θ, corresponding to (31) is unity. For each n, the total number of customers simulated for each set of design points is 30,000, these being allocated to the design points according to the corresponding allocation under consideration. The service rate was set at unity, so the traffic intensities were therefore equal to the arrival rates in all cases. The average waiting times of customers in each of the runs was recorded and the regression model (3) was then fitted to the observations, yielding the estimated value $\hat{\theta}$ as the response. As already mentioned, the only exceptional case is where the theory calls for a design point of zero with no runs to be made. This can simply be interpreted to mean that the known value of the waiting time at $x_1 = 0$ is used to impose a constraint on the fitting procedure. In our example this corresponds to immediately setting $\beta_0 = 0$, and using the results at the other design points to estimate the other parameters. Each experiment was replicated 100 times, giving 100 independent $\hat{\theta}$ estimates. The sample mean and sample variance of the 100 $\hat{\theta}'s$ are given in Table 2, as well as a 95% confidence interval for θ based on these values. From each experiment, a 95% confidence interval for θ was also calculated, using the $\hat{\theta}$ value from that experiment and the standard estimate of its variance based on the residual sum of squares.The observed coverage of these 100 confidence intervals is also given in the Table 2.

Comparing the sample variances in the optimized and non-optimized versions, we see an over hundred-fold (0.000149 : 0.0198) and over twenty fold (0.00186 : 0.0465) reduction in the variability for the cases $n = 2$ and $n = 3$.. These reductions are as predicted by the theory: 208 : 22091 and 1273 : 23391 respectively; see Tables1b and 1c. Moreover this represents a genuine saving as the computing times were almost identical for the non-optimized and the optimized experiments (the computer time for 100 macro-replications was 260 seconds (plus or minus 3 seconds). The 95% confidence intervals based on the 100 macro-replications include the true value of θ (= 1) in all cases; the observed coverages are similarly satisfactory.

Insert Table 2 about here

7.2 Bias Error when Specifying the Regression Metamodel: Priority Queue Examples

7.2.1 Exact Numerical Results

Next we consider the extent to which bias error can be corrected by the polynomial factor when $f(x)$ is only an approximately known. We again illustrate with the M/M/1 queue, but not FIFO. By adjusting the priority rule, behaviour for $f(x)$ that is markedly different from $f(x) \simeq (1-x)^{-1}$ can be obtained. Again we give both numerical results and simulation results.

The service time of an arriving customer is assumed to be exponentially distributed and to be precisely known at the moment of customer arrival, who takes a position in the queue dependent on this value. We consider the case of shortest processing time (SPT) first, and the case of longest processing time (LPT) first. We consider the nonpremptive case. Kleinrock (1969, Para. 3.6) gives a general formula for $W(x)$, the average steady-state waiting time of a customer with service time x. For the M/M/1 case, with arrival rate x, service rate of unit, and SPT and LPT respectively, the formula reduces after some manipulation to

$$E(W) = xf_S(x) \ \ and \ \ E(W) = xf_L(x)$$

where

$$f_S(x) = \int_0^\infty e^{-u}[1 - x + x(1+u)e^{-u}]^{-2}du$$

and

$$f_L(x) = \int_0^\infty e^{-u}[1 - x(1+u)e^{-u}]^{-2}du.$$

The behaviour of $f_S(x)$ and $f_L(x)$ is substantially different from that for the M/M/1/FIFO queue, when $f(x) = (1-x)^{-1}$. If we write $f_S(x)$ as $f_S(x) = (1-x)^{-a_S(x)}$, and $f_L(x)$ as $f_L(x) = (1-x)^{-a_L(x)}$, then we find that, as x increases from 0.5 to 0.99 say, $a_S(x)$ remains roughly constant, changing from 0.51 to 0.62, whereas $\alpha_L(x)$ remains substantially constant with a value of approximately 1.5. Thus $f(x)$ increases substantially more quickly than $f_S(x)$ as x increases, and so overestimates it. Conversely, it increases substantially more slowly than $f_L(x)$, and so underestimates it.

Suppose now that we use $\eta(x, \beta) = (\beta_0 + ... + \beta_k x^k)/(1-x)$ to estimate $xf_S(x)$ and $xf_L(x)$. Figures 1a and 1b show the fitted polynomial $\beta_0 + ... + \beta_k x^k$ for the cases $k = 1$, 2, and 4, where the $\beta's$ are the values obtained from the least squares fit of $\eta(x, \beta)$ to $xf_S(x)$ and $xf_L(x)$

using the design points from Table 1 with $n = 5$. For both the SPT and the LPT cases, the fit obtained with $k = 1$ is not satisfactory. The maximum relative error

$$\delta = \max_{x\epsilon[0.5,\ 0.95]} \left| \frac{\eta(x,\beta)}{xf(x)} - 1 \right|$$

is $\delta = 0.234$ and 1.86 for the SPT and the LPT cases respectively, when $k = 1$. In fact the fitted curve does not even remain positive for the LPT case. However, the fit obtained with $k = 4$ is satisfactory, with $\delta = 0.009$ for the SPT case and $\delta = 0.069$ for the LPT case. The fit is thus reasonably close over the interpolation range $[0.5, 0.95]$.

The example supports the earlier analysis which indicated that the polynomial factor can correct the bias resulting from an incorrect form for $f(.)$, provided the range of interest is restricted to where $\eta(.)$ remains bounded. Our conclusion is that for smoothly behaved regression metamodels, the model in (1) is sufficiently flexible to provide an adequate fit without recourse to a polynomial of unacceptably high degree.

Insert Fig. 1a, 1b about here.

7.2.2 Simulation Results

The above numerical results were compared with simulations. The simulation used the full method so that the sequential procedure for updating the variance estimates at the different design points is tested. Both SPT and LPT cases were considered. In both, to test that the method does correct for bias, we tried fitting $\eta(x,\beta) = (\beta_0 + \beta_1 x + ... + \beta_k x^k)(1-x)^{-1}$ so that the incorrect $f(x) = (1-x)^{-1}$ is used. Three versions of $g_a(x)$ were used: $g_1(x) = \sqrt{x(1-x)^{-3}}$, $g_2^2(x) = \sqrt{x(1-x)^{-4}}$, $g_3^2(x) = \sqrt{x(1-x)^{-5}}$. Moreover the use of optimal design points and allocations was compared with the naive design where the design points are equally spaced over a given range of x values. We again took this range to be $[0.5,\ 0.95]$. Table 3 gives the results obtained from 100 macro-replications. The highest degree of the polynomial factor was taken to be $k_0 = 4$ (i.e. quartic), so that the number of parameters was $k_0 + 1 = 5$. In each experiment $g_a(x)$ was used to estimate the optimum design point values. The approximation $g_a(x)$ was also used to estimate the number of customers per run at each design point that would make $Var(\epsilon)$ constant (see Section 5.1). An initial set of 12 runs was then made at each design point to estimate the actual variability of ϵ at each x_i. Runs were then allocated and executed at the design points as in Theorem 3, until approximately $30,000$ customers were simulated in total. (The total number of customers was nearly, but not exactly the same in each macro-replication. This was because run lengths were held at the value fixed at the pilot stage,

to enable the variance estimates s_i to be calculated from observations based on unchanging run length throughout the macro-replication.)

We conclude three main points, based on Table 3:

(i) Irrespective of the form assumed for $g(x)$, the sample variance of the $\hat{\theta}$ remains substantially constant (approximately 0.002 for optimum $\mathbf{x}$; and 0.004 for equally spaced $\mathbf{x}$ for SPT), indicating that the sequential adaptive adjustment of the number of runs made at each design point does correct the inaccuracy in the initial form of $g(x)$.

(ii) The sample variance for the optimized case is roughly half that of the sample variance for the evenly spread design point case (0.002 compared with 0.004). This is very much in line with the results for the cases presented in Tables 1a and b.

(iii) For the fit to be adequate we should have

$$\eta(x, \hat{\beta}) = (\hat{\beta}_0 + \hat{\beta}_1 x + ... + \hat{\beta}_k x^k)/(1-x) \simeq x f_S(x) \ \ (and \ \ x f_L(x) \ respectively),$$

so that

$$(\hat{\beta}_0 + \hat{\beta}_1 x + ... + \hat{\beta}_k x^k) \simeq x(1-x) f_S(x) \ \ (and \ \ x(1-x) f_L(x) \ respectively).$$

As we showed in the previous subsection, the assumed form of $f(x) = (1-x)^{-1}$ is very different from the actual forms $f_S(x)$ and $f_L(x)$, so the main source of concern is if the polynomial factor is able to correct for this. Figures 2a and 2b summarise the fit obtained from the 100 macro-replicates. The degree of the fitted polynomial varied with the macro-replicate; however a quadratic was selected the most often, and a linear model almost never. The envelope of the 100 fitted polynomials is compared with the true curve for each of the two priority queues; there is little appreciable bias.

Insert Table 3 about here

Insert Fig. 2a, 2b about here

7.3 Terminal PAD Controller Simulation

The final example has behaviour quite different from that of an M/M/1 queue. It is included to show that the full method can be applied to complex queueing networks, and that it leads to worthwhile improvements in efficiency.

We consider a model of a packet assembly/disassembly device (PAD) as described in Molloy (1989). The PAD receives characters generated from a number of terminals, K say, and assembles them into packets for transmission into a network. We assume that the input of characters into each terminal is Poisson, each with same rate x. The characters from a terminal wait in their own input buffer, one for each terminal. A packet is formed once the buffer is full, or if a special control character arrives, whichever event occurs first. It is then tagged with a fixed number of identification overhead characters and the whole is sent to an output buffer, where it waits in a FIFO queue for transmission. The output queue serves all the terminals, and transmits characters at a constant rate (C char/sec) into the network so long as there are characters to transmit.

We consider the objective of measuring T, the average character delay. The queue is quite complex in its overall operation because of the way packets are formed. The average delay experienced by a character does not change monotonically with x. In fact, the average delay of a character increases to infinity if either $x \to 0$ or if $x \to R$, where R is the saturation input rate for the system. The unbounded delay as x approaches zero is not of particular concern, as it is due simply to slow packet formation because of the low input rate. The behaviour as $x \to R$ is the important case, corresponding to true saturation; this is what we focus on. At first sight, one might expect $R = C/K$. However, R is strictly less than this value because overhead characters are added to packets, making the traffic rate greater than xK. Molloy (1989) gives two approximations for T , but neither seems particularly accurate.

We estimated T by simulation, measuring in appropriate standardised units so that $C/K = 1$. Details of the simulation model are given in Molloy (Para. 8.7.3). We simulated ten input terminals. The size of each terminal buffer was set equal to 32. The probability that a character is the special control character was set equal to 0.02. One overhead character was added to each packet. Some direct calculation shows that $R \simeq 0.959$ in this case; however, we did not make explicit use of this in choosing $f(x)$. Instead, we assumed $f(x) = [x(1-x)]^{-1}$, relying on the polynomial factor to compensate for possible bias. Two simulations, with x in the range [0.5, 0.95] were carried out using a polynomial factor of degree 6. In one simulation the design points and allocations were optimally assigned, whilst the other used the naive design with equally spaced design points and evenly spread allocations. For the same reasons as in the M/M/1 queue examples, it is reasonable to focus on the estimator $\hat{\theta}$ given in (31), and to use the performance measure (13) for the sequential run allocation. The approximation $g_a(x) = [x(1-x)]^{-2}$ was used to set the design point values. This yielded the points and

allocations given in Table 4a. Notice that the condition of using both end points leads to similar design point values in both the optimised and non-optimised cases. However, a heavy allocation is made to the final x design points in the optimized case.

Insert Table 4 about here

Insert Fig. 3 about here

Table 4b summarizes the results of 200 macro-replications for the two designs. As indicated above, because of the way run lengths are constructed, the total number of packets simulated was slightly different in the two cases, with approximately $76,000$ in each macro-replication for the optimized case and approximately $70,000$ for the case with equally distributed design points and run allocations. The estimated ratio of the performance measures is $72595 : 212198$; this is based on the assumption that a polynomial of degree 6 is needed. In the experiments, a polynomial of degree 5 was selected in the majority of the macro-replications in both cases. Using the same design points and allocations when only a polynomial of degree 5 is needed, the ratio of performance measures falls to $31617 : 46497$. However these are only estimates of likely performance, being based on assumed forms of $f(x)$ and $g_a(x)$. The observed ratio of performance measures: $2.86 \times 10^{-5} : 1.23 \times 10^{-4}$ (i.e. a four-fold improvement) is somewhat better than either estimate, even allowing for the 10% difference in the number of packets simulated in the two cases. Figure 3, which shows the upper and lower envelopes of the 200 fitted regression metamodels using the optimised and non-optimised design points, corroborates the observed performance measures. Our main conclusion is that the difference in the allocations between the two designs leads to a substantial improvement in the efficiency of the experiment when our suggested procedure is used.

8 Summary and Conclusion

We have proposed a general regression metamodel for use in queueing simulations where the character of the output of interest is to be examined over a range of traffic intensities. In such a situation the mean and the variance of the output will increase considerably with the traffic intensity, and the metamodel specifically allows for this. We have suggested a straightforward experimental procedure which allows initial information of the likely variability of both the mean and the variance to be used to select design point values. The main steps of the procedure have been discussed and have been justified theoretically.

In particular we have shown that the procedure is robust. Incorporation of a polynomial factor in the regression metamodel enables bias error to be corrected for in a model fitting stage.

As far as efficiency is concerned, the main requirement is to take into account the variabilitiy of the response when allocating runs to the design points. We have shown how the optimal number of runs depends on this variability. Our procedure uses a simple multistage sampling scheme that allows the variance of the response at each design point to be continuously updated, so that the proportion of observations made at each point converges towards the optimum as the number of runs increases.

Results from simulation experiments indicate that the method works very much as suggested by the theory and gives worthwhile improvements in efficiency.

It would be useful to extend our method to allow multiple outputs and factors, and this is the subject of future research.

Appendix

Proof of Theorem 1: The notation is as in Section 2. Recall that $\mathbf{A}$ has (s,t)th element

$$a_{st} = \sum_{i=1}^{k+1} n_i x_i^{s+t-2}$$

where $n_i = m_i r^2(x_i)$. Using identities such as

$$\begin{vmatrix} a_{10} & \dots & a_{1i}+b_{1i} & \dots & a_{1k} \\ . & \dots & . & \dots & . \\ . & \dots & . & \dots & . \end{vmatrix} = \begin{vmatrix} a_{10} & \dots & a_{1i} & \dots & a_{1k} \\ . & \dots & . & \dots & . \\ . & \dots & . & \dots & . \end{vmatrix} + \begin{vmatrix} a_{10} & \dots & b_{1i} & \dots & a_{1k} \\ . & \dots & . & \dots & . \\ . & \dots & . & \dots & . \end{vmatrix},$$

we can write the determinant of $\mathbf{A}$ as

$$|\mathbf{A}| = \sum_{\substack{i_l \neq i_m \\ 1 \le i_1, \dots, i_{k+1} \le k+1}} n_{i_1} n_{i_2} \dots n_{i_{k+1}} \begin{vmatrix} 1 & x_{i_2} & . & x_{i_{k+1}}^k \\ x_{i_1} & x_{i_2}^2 & . & . \\ . & . & . & . \\ x_{i_1}^k & x_{i_2}^{k+1} & . & x_{i_{k+1}}^{2k} \end{vmatrix}.$$

Thus

$$|\mathbf{A}| = \sum_{\substack{i_l \neq i_m \\ 1 \le i_1, \dots, i_{k+1} \le k+1}} n_{i_1} n_{i_2} \dots n_{i_{k+1}} x_{i_1}^0 x_{i_2}^1 \dots x_{i_{k+1}}^k \begin{vmatrix} 1 & 1 & . & 1 \\ x_{i_1} & x_{i_2} & . & x_{i_{k+1}} \\ . & . & . & . \\ x_{i_1}^k & x_{i_2}^k & . & x_{i_{k+1}}^k \end{vmatrix}$$

$$= n_1 n_2 ... n_{k+1} |\mathbf{V}| \sum_{\substack{i_l \neq i_m \\ 1 \leq i_1, ..., i_{k+1} \leq k+1}} \pm x_{i_1}^0 x_{i_2}^1 ... x_{i_{k+1}}^k = n_1 n_2 ... n_{k+1} |\mathbf{V}|^2.$$

A similar calculation shows that $\mathbf{A}_{ij}$, the cofactor of the (i,j)th element of $|\mathbf{A}|$, has the form

$$| \mathbf{A}_{ij} | = \sum_{s=1}^{k+1} |\mathbf{V}_{si}| |\mathbf{V}_{sj}| n_s^{-1} \prod_{t=1}^{k+1} n_t,$$

and substituting these expressions into

$$(\mathbf{A}^{-1})_{ij} = | \mathbf{A}_{ji} | / | \mathbf{A} |$$

gives the required form (14) for $\mathbf{A}^{-1}$. □

Proof of Theorem 3

It is simplest to study the build up of runs, cycle by cycle (rather than let $N \to \infty$ directly). The adjustment from cycles to N does not affect the convergence limits in (23).

By the strong law of large numbers, for given i, if $M_{ij} \to \infty$, then $s_{ij}^2 \to [g(x_i)\sigma]^2$ *a.s.*, and so $a_i/\tilde{r}_{ij} \to a_i/r_i$ *a.s.* Suppose that there is some set of indices $I \subseteq \{1, 2, ..., n\}$ for which $M_{ij} \not\to \infty$, for $i \in I$. This implies that, after some cycle, j_0, all additional runs are made only at x_i for which $i \notin I$. Thus $a_i/\tilde{r}_{ij}$ is constant for all $j \geq j_0$ for $i \in I$ and $a_i/\tilde{r}_{ij} \to a_i/r_i$ *a.s.* as $j \to \infty$ for $i \notin I$. It follows that $\hat{\pi}_i \to c_i$ *a.s.* for all i, where the c_i are all constants greater than zero. But this means that for $i \in I$, $\lfloor (j+1)\hat{\pi}_{ij} \rfloor \to \infty$ as $j \to \infty$, implying that $M_{ij} \to \infty$. This contradicts the definition of I so that $M_{ij} \to \infty$ as $j \to \infty$ for all i. Thus $a_i/\tilde{r}_{ij} \to a_i/r_i$ *a.s.* for all i, and so $\hat{\pi}_{ij} \to \pi_i$ *a.s.* for all i.

Given $\epsilon > 0$, there exists $j_0(\epsilon)$ such that $|\hat{\pi}_{ij} - \pi_i| < \epsilon$ for all $j > j_0$. Let $M' = \max_{j < j_0} \{\lfloor (j+1)\hat{\pi}_{ij} \rfloor\}$. Then there is a $j_1 > j_0$ such that

$$\lfloor (j+1)\hat{\pi}_{ij} \rfloor > j(\pi_i - \epsilon) > j_1(\pi_i - \epsilon) > M' \quad \textit{for all} \quad j > j_1.$$

Therefore for $j > j_1$

$$\lfloor (j+1)(\pi_i - \epsilon) \rfloor < M_{i,j+1} = \max_{l \leq j} \{\lfloor (l+1)\hat{\pi}_{i,l} \rfloor\} < \lfloor (j+1)(\pi_i + \epsilon) \rfloor.$$

Now

$$\pi_i - (j+1)^{-1} - \epsilon < \lfloor (j+1)(\pi_i - \epsilon) \rfloor / (j+1)$$

and

$$\lfloor (j+1)(\pi_i + \epsilon) \rfloor / (j+1) < \pi_i + \epsilon.$$

Therefore

$$\pi_i - (j+1)^{-1} - \epsilon < M_{i,j+1}/(j+1) < \pi_i + \epsilon \quad \textit{for} \; j > j_1,$$

and if we take j_1 sufficiently large we have

$$\pi_i - 2\epsilon < M_{i,j+1}/(j+1) < \pi_i + \epsilon \quad for\ j > j_1.$$

As ϵ is arbitrary we have therefore that $M_{i,j+1}/(j+1) \rightarrow \pi_i \ \ a.s.$ Finally, as $\sum \pi_i = 1$, we can replace $(j+1)$ by $\sum_{l=1}^{n} M_{l,j+1}$ without affecting the limit, i.e.,

$$M_{i,j+1} / \sum_{l=1}^{n} M_{l,j+1} \rightarrow \pi_i \ \ a.s. \tag{32}$$

as $j \rightarrow \infty$.

The above limit is obtained by counting in cycles and letting $j \rightarrow \infty$. However it is readily verified that the adjustment to count directly in terms of N and to replace $M_{i,j+1}$ by M_i in (32), does not alter the result. This proves the second part of the theorem. □

References

Aitken, A.C. (1964), Determinants and Matrices. Oliver and Boyd, Edinburgh.

Asmussen, S. (1992) "Queueing Simulation in Heavy Traffic." Math. of Operations Research, 17, pp. 84-111.

Atkinson, A.C. and Donev, A.N. (1992) Optimal Experimental Designs. Clarendon Press, Oxford.

Cheng, R.C.H. (1976), "A note on the effect of initial conditions on a simulation ruco" Operational Research Quarterly, 27, 2, ii, pp. 467-470.

Cheng, R.C.H. (1990), Fitting parametric models by conditional simulation. In Proceedings of the 1990 Winter Simulation Conference (eds. O. Balci, R.P. Sadowski and R.E. Nance), IEEE, Piscataway, New Jersey, pp. 333-336.

Cheng, R.C.H. and Traylor, L. (1993), "Conditional simulation techniques for estimating the behaviour of nearly saturated queues.", UKSS Conference Proceedings 1993, pp. 105-109.

Ermakov, S.M. and Melas, V.B. (1995), Design and Analysis of Simulation Experiments. Kluwer Academic Publishers, Dordrecht, Netherlands.

Fedorov, V.V. (1972), Theory of Optimal Experiments. Academic Press, New York.

Johnson, L.W. and Riess, R.D. (1982), Numerical Analysis, 2nd Edition. Addison-Wesley, Reading, Massachusetts.

Kiefer, J. and Wolfowitz, J. (1959), " Optimum designs in regression problems." Annals Mathematical Statistics, 30, pp. 271-294.

Kleijnen, J.P.C. (1987), Statistical tools for simulation practitioners. Marcel Dekker, New York.

Kleijnen, J.P.C. (1992), "Regression metamodels for simulation with common random numbers: comparison of validation tests and confidence intervals." Management Science, 38, no. 8, August, pp. 1164-1185.

Kleijnen, J.P.C. and Van Groenendaal, W.J.H. (1995), "Two-stage versus sequential sample-size determination in regression analysis of simulation experiments." Amer. J. Math. and Mgmnt Sciences, 15, pp. 83-114.

Kleinrock, L. (1976), Queueing Systems Vol. II: Computer Applications. John Wiley: New York.

Molloy, M.K. (1989), Fundamentals of Performance Modeling. Macmillan: New York.

Nash, J.C. (1979), Compact Numerical Methods for Computers: linear algebra and function minimisation. Adam Hilger Ltd: Bristol.

Pukelsheim, F. (1993), Optimal Design of Experiments. Wiley, New York.

Reiman M.I., Simon B., and Willie J.S. (1992), "Simterpolation: a simulation based interpolation approximation for queueing systems." Operations Research, 40, no. 4, July-August, pp. 706-723.

Seber, G.A.F. (1977), Linear Regression Analysis. John Wiley, New York.

Vollebregt, T.A.J. (1996), "Experimental design for simulation", PhD Dissertation, Dept. of Management, Univ. of Canterbury, New Zealand.

Whitt, W. (1989), "Planning queueing simulations." Management Science, 35, no. 11, November, pp. 1341-1366.

Table 1

Performance Measure for Selected Combinations of $\mathbf{x}$ *and* π

1a : *Optimized* $\mathbf{x}$, *Fixed Endpoints* 0.5 *and* 0.95

			Average Variance Criterion								*Single Point Var Criterion*	
n	2		3		4		5		6		5	
	x	π	x	π	x	π	x	π	x	π	x	π
	.50	.022	.500	.013	.500	.009	.500	.007	.500	.006	.500	.027
	.95	.978	.756	.119	.617	.039	.564	.024	.542	.017	.545	.059
			.950	.868	.819	.191	.706	.064	.641	.035	.668	.085
					.950	.761	.859	.258	.769	.090	.827	.178
							.950	.647	.885	.309	.950	.651
									.950	.542		
Np/σ^2	11,552		9,746		9,680		10,383		11,571		11,441	
Np_E/σ^2	22,091		22,443		23,900		25,422		27,714		27,636	

1b : *Evenly Spaced Design Points in the Interval* (0.5, 0.95)

		Average Variance Criterion						
n	2	3		4		5		6	
	x	π	x	π	x	π	x	π	x	π
	.50	.022	.500	.013	.50	.009	.5000	.007	.50	.005
	.95	.978	.725	.101	.65	.048	.6125	.031	.59	.022
			.950	.886	.80	.179	.7250	.084	.68	.055
					.95	.764	.8375	.242	.77	.124
							.9500	.636	.88	.280
									.95	.514
Np/σ^2	11,552		9,808		9,929		11,342		14,281	
Np_E/σ^2	22,091		23,391		24,739		26,739		30,966	

1c : *Optimized* $\mathbf{x}$

			Average Variance Criterion								*Single Point Var Criterion*	
n	2		3		4		5		6		5	
	x	π	x	π	x	π	x	π	x	π	x	π
	.00	.0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.33	1.0	.107	.160	.054	..057	.032	.025	.022	.012	.031	.052
			.721	.840	.423	.114	.270	.036	.187	.016	.258	.069
					.846	.829	.629	.117	.471	.036	.605	.135
							.896	.821	.744	.146	.911	.743
									.921	.790		
Np/σ^2	208		1,273		3,246		5,500		7,496		6,088	
Np_E/σ^2	415		2,793		9,131		18,989		29,084		20,718	

Table 2

Estimating E(W) in the M/M/1Queue

2a. *Using* $\theta = \beta_0 + \beta_1$. *Correct value is* $\theta = 1$

Non − Optimized		*Optimized*	
x	*# of customers/per run*	x	*# of customers/per run*
0.50	15,000	0.000	0
0.95	15,000	0.333	30,000
# of runs at each x : 5			
Results from 100 *simulations* :			
Sample Mean of $\hat{\theta}'s$ 0.974		*Sample Mean of* $\hat{\theta}'s$ 0.999	
Sample Var of $\hat{\theta}'s$ 0.0198		*Sample Var of* $\hat{\theta}'s$ 0.000149	
95% *Confidence Interval for* θ :			
(.946, 1.002)		(.997, 1.001)	
Observed Coverage of Computed 95% $CI's$:			
95%		97%	

2b. *Using* $\theta = \beta_0 + \beta_1 + \beta_2$. *Correct value is* $\theta = 1$

Non − Optimized		*Optimized*	
x	*# of customers/per run*	x	*# of customers/per run*
0.500	10,000	0.000	0
0.725	10,000	0.107	4,800
0.950	10,000	0.721	25,200
# of runs at each x : 5			
Results from 100 *simulations* :			
Sample Mean of $\hat{\theta}'s$ 0.974		*Sample Mean of* $\hat{\theta}'s$ 0.994	
Sample Var of $\hat{\theta}'s$ 0.0465		*Sample Var of* $\hat{\theta}'s$ 0.00186	
95% *Confidence Interval for* θ :			
(.931, 1.016)		(.986, 1.002)	
Observed Coverage of Computed 95% $CI's$:			
96%		98%	

Table 3

Sample Mean and Variance of $\hat{\theta}$ From 100

Macro – Replications of the M/M/1/SPT and M/M/1/LPT

		$M/M/1/SPT$		$M/M/1/LPT$	
		$\bar{\hat{\theta}}$	$\widehat{Var}\,(\hat{\theta})$	$\bar{\hat{\theta}}$	$\widehat{Var}\,(\hat{\theta})$
g_1	*Optimized*	.203	.0019	5.66	1.81
	Even Spread	.201	.0031	5.28	4.40
g_2	*Optimized*	.210	.0020	5.64	1.95
	Even Spread	.210	.0022	4.90	4.06
g_3	*Optimized*	.214	.0013	5.63	1.70
	Even Spread	.216	.0034	5.05	4.27

Table 4

Results for PAD Controller Example

4a : *Design Points and Allocations*

Optimized	x	0.5	0.527	0.601	0.703	0.810	0.903	0.95
(*Fixed Endpoints* 0.5 *and* 0.95)	$\hat{\pi}$	0.025	0.052	0.060	0.080	0.129	0.282	0.373
Even Spread	x	0.5	0.575	0.650	0.725	0.800	0.875	0.95
	π	0.143	0.143	0.143	0.143	0.143	0.143	0.143

4b : *Estimated Mean and Variance of* 200 $\hat{\theta}$ *Estimates*

	$\bar{\hat{\theta}}$	$\widehat{Var}(\hat{\theta})$	95% *C.I. for* θ	*Total Simulation Time*
Optimized	.0157	2.86×10^{-5}	(0.0150, 0.0164)	1.46 (*hours*)
Even Spread	.0168	1.23×10^{-4}	(0.0153, 0.0184)	1.28 (*hours*)

Figure 1a: Best Fit Polynomials: M/M/1/SPT

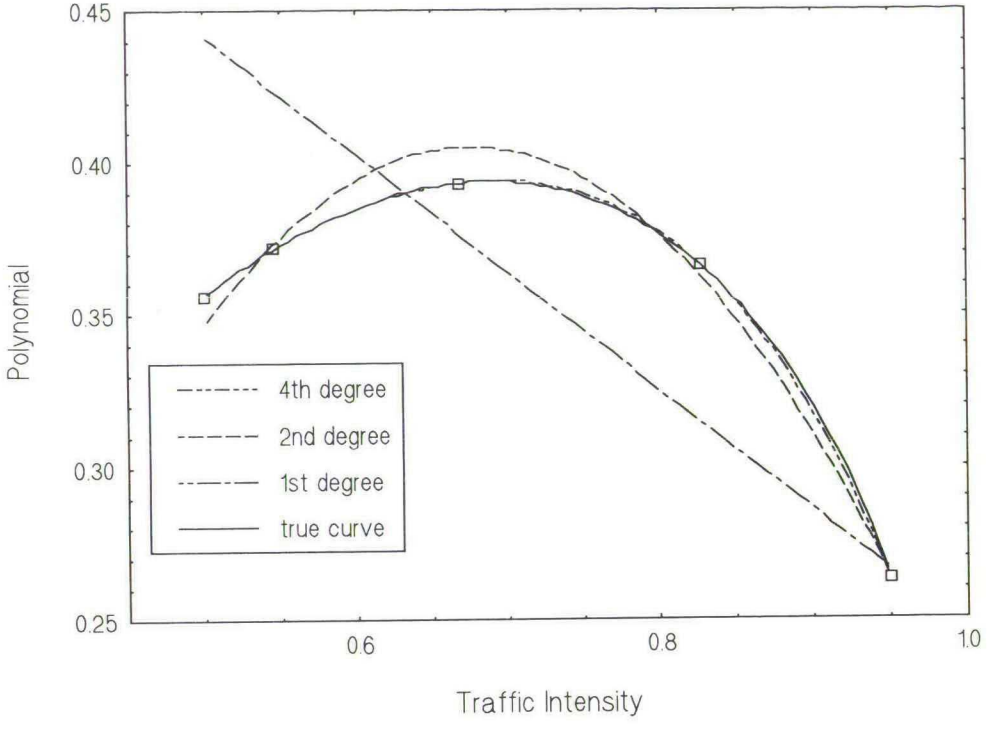


Figure 2a: Envelope of 100 fitted Polynomials: M/M/1/SPT

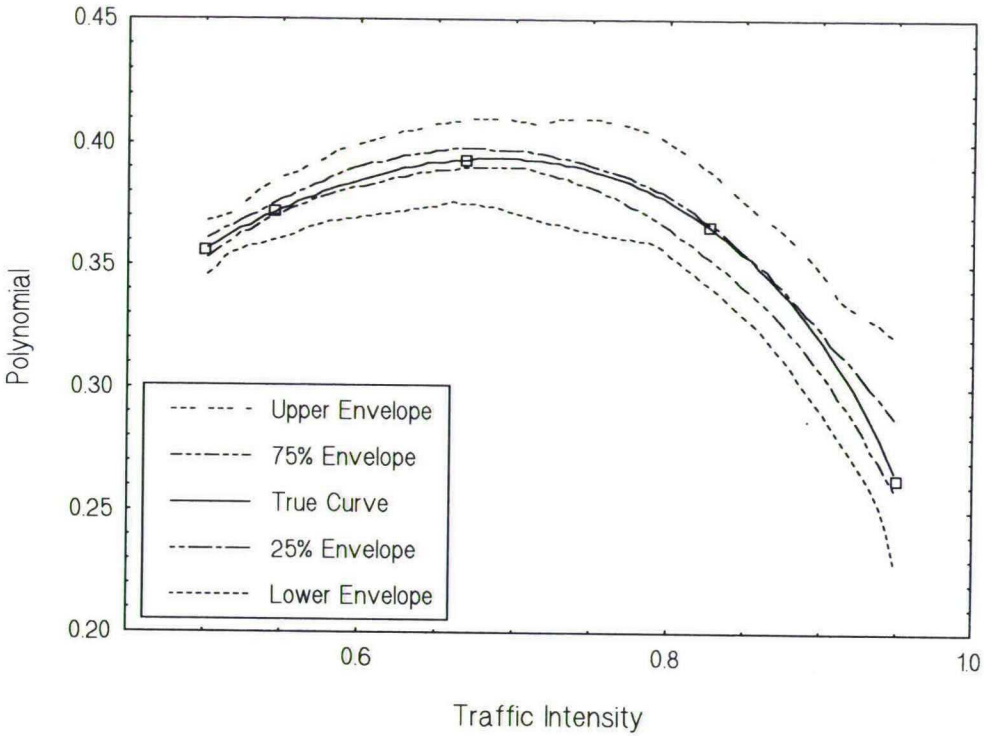


Figure 1b: Best Polynomial Fits: M/M/1/LPT

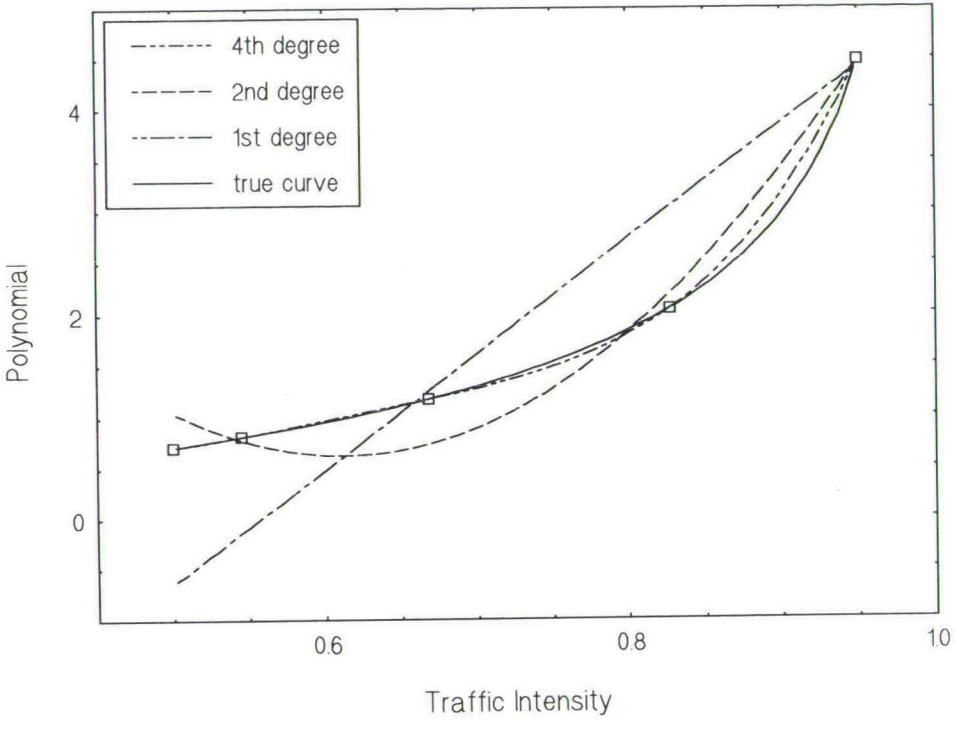


Figure 2b: Envelope of 100 Fitted Polynomials: M/M/1/LPT

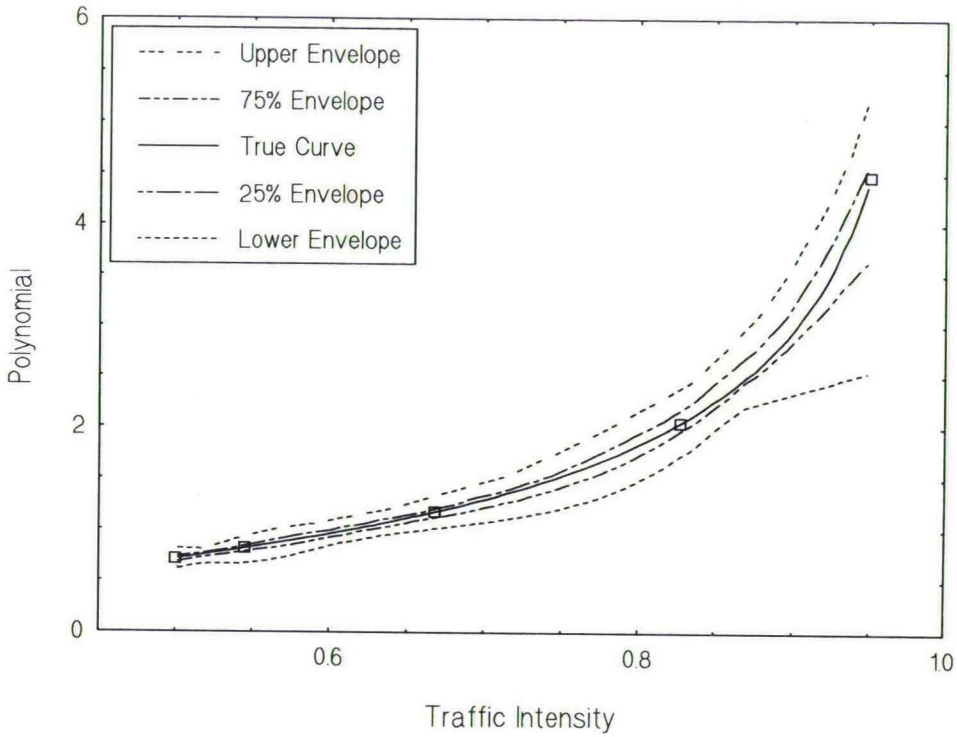
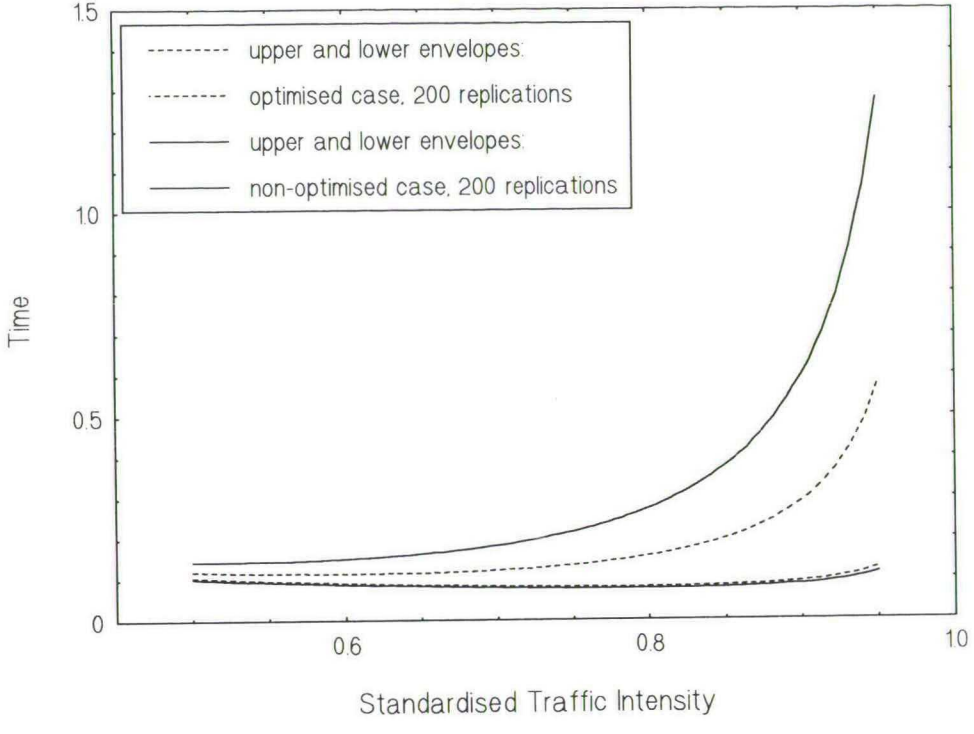


Figure 3: Average Character Processing Time in PAD Queue



No.	Author(s)	Title
95116	F. Kleibergen and H. Hoek	Bayesian Analysis of ARMA models using Noninformative Priors
95117	J. Lemmen and S. Eijffinger	The Fundamental Determiniants of Financial Integration in the European Union
95118	B. Meijboom and J. Rongen	Clustering, Logistics, and Spatial Economics
95119	A. de Jong, F. de Roon and C. Veld	An Empirical Analysis of the Hedging Effectiveness of Currency Futures
95120	J. Miller	The Effects of Labour Market Policies When There is a Loss of Skill During Unemployment
95121	S. Eijffinger, M. Hoeberichts and E. Schaling	Optimal Conservativeness in the Rogoff (1985) Model: A Graphical and Closed-Form Solution
95122	W. Ploberger and H. Bierens	Asymptotic Power of the Integrated Conditional Moment Test Against Global and Large Local Alternatives
95123	H. Bierens	Nonparametric Cointegration Analysis
95124	H. Bierens and W. Ploberger	Asymptotic Theory of Integrated Conditional Moment Tests
95125	E. van Damme	Equilibrium Selection in Team Games
95126	J. Potters and F. van Winden	Comparative Statics of a Signaling Game: An Experimental Study
9601	U. Gneezy	Probability Judgements in Multi-Stage Problems: Experimental Evidence of Systematic Biases
9602	C. Fernández and M. Steel	On Bayesian Inference under Sampling from Scale Mixtures of Normals
9603	J. Osiewalski and M. Steel	Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models
9604	J. Potters and J. Wit	Bets and Bids: Favorite-Longshot Bias and Winner's Curse
9605	H. Gremmen and J. Potters	Assessing the Efficacy of Gaming in Economics Educating
9606	J. Potters and F. van Winden	The Performance of Professionals and Students in an Experimantal Study of Lobbying
9607	J.Kleijnen, B. Bettonvil and W. van Groenendaal	Validation of Simulation Models: Regression Analysis Revisited
9608	C. Fershtman and N. Gandal	The Effect of the Arab Boycott on Israel: The Automobile Market
9609	H. Uhlig	Bayesian Vector Autoregressions with Stochastic Volatility

No.	Author(s)	Title
9610	G. Hendrikse	Organizational Change and Vested Interests
9611	F. Janssen, R. Heuts and T. de Kok	On the (R,s,Q) Inventory Model when Demand is Modelled as a compound Bernoulli Process
9612	G. Fiestras-Janeiro P. Borm and F. van Megen	Protective Behaviour in Games
9613	F. van Megen, G. Facchini, P. Borm and S. Tijs	Strong Nash Equilibria and the Potential Maximizer
9614	J. Miller	Do Labour Market Programmes Necessarily Crowd out Regular Employment? - A Matching Model Analysis
9615	H. Huizinga	Unemployment Benefits and Redistributive Taxation in the Presence of Labor Quality Externalities
9616	G. Asheim and M. Dufwenberg	Admissibility and Common Knowlegde
9617	C. Fernández, J. Osiewalski and M. Steel	On the Use of Panel Data in Bayesian Stochastic Frontier Models
9618	H. Huizinga	Intrafirm Information Transfers and Wages
9619	B. Melenberg and B. Werker	On the Pricing of Options in Incomplete Markets
9620	F. Kleibergen	Reduced Rank Regression using Generalized Method of Moments Estimators
9621	F. Janssen, R. Heuts and T. de Kok	The Value of Information in an (R,s,Q) Inventory Model
9622	F. Groot, C. Withagen and A. de Zeeuw	Strong Time-Consistency in the Cartel-Versus-Fringe Model
9623	R. Euwals and A. van Soest	Desired and Actual Labour Supply of Unmarried Men and Women in the Netherlands
9624	M. Khanman, M. Perry and P.J. Reny	An Ex-Post Envy-Free and Efficient Allocation Mechanism: Imperfect Information Without Common Priors
9625	C. Eaves, G. van der Laan D. Talman and Z. Yang	Balanced Simplices on Polytopes
9626	H. Bloemen and A. Kalwij	Female Employment and Timing of Births Decisions: a Multiple State Transition Model
9627	P. Bolton	Strategic Experimentation: a Revision
9628	C. Fershtman and Y. Weiss	Social Rewards, Externalities and Stable Preferences

No.	Author(s)	Title
9629	P.M. Kort, G. Feichtinger R.F. Hartl and J.L. Haunschmied	Optimal Enforcement Policies (Crackdowns) on a Drug Market
9630	C. Fershtman and A. de Zeeuw	Tradeable Emission Permits in Oligopoly
9631	A. Cukierman	The Economics of Central Banking
9632	A. Cukierman	Targeting Monetary Aggregates and Inflation in Europe
9633	F. de Roon, C. Veld and J. Wei	A Study on the Efficiency of the Market for Dutch Long Term Call Options
9634	B. van Aarle, L. Bovenberg M. Raith	Money, Fiscal Deficits and Government Debt in a Monetary Union
9635	G. van der Laan, D. Talman and Z. Yang	Existence of and Equilibrium in a Competitive Economy with Indivisibilities and Money
9636	I. Brouwer, J. van der Put C. Veld	Contrarian Investment Strategies in a European Context
9637	M. Berg, F. van der Duyn Schouten and J. Jansen	Optimal Service Policies to Remote Customers with Delay-Limits
9638	F.C. Drost and C.A.J. Klaassen	Efficient Estimation in Semiparametric GARCH Models
9639	A.G. de Kok and F.B.S.L.P. Janssen	Demand Management in Multi-Stage Distribution Chain
9640	D.A. Freund, T.J. Kniesner A.T. LoSasso	How Managed Care Affects Medicaid Utilization A Synthetic Difference-in-Differences Zero-Inflated Count Model
9641	M.J. Lee	Instrumental Variable Estimation For Linear Panel Data Models
9642	W. Härdle, E. Mammen and M. Müller	Testing Parametric versus Semiparametric Modelling in Generalized Linear Models
9643	C. Dustmann, N. Rajah and A. van Soest	Part-Time Work, School Success and School Leaving
9644	S.C.W. Eijffinger and M. Hoeberichts	The Trade Off Between Central Bank Independence and Conservativeness
9645	R. Sarin and P. Wakker	A Single-Stage Approach to Anscombe and Aumann's Expected Utility

No.	Author(s)	Title
9646	J.P. Ziliak and T.J. Kniesner	The Importance of Sample Attrition in Life Cycle Labor Supply Estimation
9647	P.M. Kort	Optimal R&D Investments of the Firm
9648	M.P. Berg	Performance Comparisons for Maintained Items
9649	H. Uhlig and Y. Xu	Effort and the Cycle: Cyclical Implications of Efficiency Wages
9650	M. Slikker and A. van den Nouweland	Communication Situations with a Hierarchical Player Partition
9651	H.L.F. de Groot	The Struggle for Rents in a Schumpeterian Economy
9652	R.M. de Jong and J. Davidson	Consistency of Kernel Estimators of heteroscedastic and Autocorrelated Covariance Matrices
9653	J. Suijs, A. De Waegenaere and P. Borm	Stochastic Cooperative Games in Insurance and Reinsurance
9654	A.N. Banerjee and J.R. Magnus	Testing the Sensitivity of OLS when the Variance Matrix is (Partially) Unknown
9655	A. Kalwij	Estimating the Economic Return to Schooling on the basis of Panel Data
9656	M. Lind and F. van Megen	Order Based Cost Allocation Rules
9657	A. van Soest, P. Fontein and Rob Euwals	Earnings Capacity and Labour Market Participation
9658	C. Fernández and M.F.J. Steel	On Bayesian Modelling of Fat Tails and Skewness
9659	R. Sarin and P. Wakker	Revealed Likelihood and Knightian Uncertainty
9660	J.R. Magnus and J. Durbin	A Classical Problem in Linear Regression or How to Estimate the Mean of a Univariate Normal Distribution with Known Variance
9661	U. Gneezy and J. Potters	An Experiment on Risk Taking and Evaluation Periods
9662	H.J. Bierens	Nonparametric Nonlinear Co-Trending Analysis, with an Application to Interest and Inflation in the U.S.
9663	J.P.C. Blanc	Optimization of Periodic Polling Systems with Non-Preemptive, Time-Limited Service
9664	M.J. Lee	A Root-N Consistent Semiparametric Estimator for Fixed Effect Binary Response Panel Data
9665	C. Fernández, J. Osiewalski and M.F.J. Steel	Robust Bayesian Inference on Scale Parameters

No.	Author(s)	Title
9666	X. Han and H. Webers	A Comment on Shaked and Sutton's Model of Vertical Product Differentiation
9667	R. Kollmann	The Exchange Rate in a Dynamic-Optimizing Current Account Model with Nominal Rigidities: A Quantitative Investigation
9668	R.C.H. Cheng and J.P.C. Kleijnen	Improved Design of Queueing Simulation Experiments with Highly Heteroscedastic Responses