Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

DETECTING BIAS DUE TO INPUT MODELLING IN COMPUTER SIMULATION

Lucy E. Morgan
Barry L. Nelson
Andrew C. Titman
David J. Worthington

Statistics and Operational Research Centre for Doctoral Training in Partnership with Industry
Lancaster University
Lancaster, LA1 4YR, UK

ABSTRACT

Bias due to input modelling is almost always assumed negligible and ignored. It is known that increasing the amount of real-world data available for modelling input processes causes this form of bias to decrease faster than the variance due to input uncertainty. However, this does not mean bias is irrelevant when considering the error in a simulation performance measure caused by input modelling. In this paper we present a response surface approach to bias estimation in simulation models along with a diagnostic test for identifying, with controlled power, bias due to input modelling of a size that would be concerning to a practitioner.

1 INTRODUCTION

Simulation models are driven by input distributions or processes. These input models are often built using real-world observations of the system of interest and are therefore always approximate due to the finite amount of available observations. This causes error in the output performance measures due to the non-linear form of simulation models. The mean squared error (MSE) about the simulation output due to input modelling can be broken down into variance, or "input uncertainty" (IU), and bias due to input modelling.

Methods for quantifying input uncertainty in simulation models exist. Song and Nelson (2015) and Cheng and Holland (1997) present methods for quantifying IU in simulation models with time homogeneous input distributions; this was extended by Morgan et al. (2016) to piecewise-constant non-stationary Poisson arrival processes. However, bias due to input modelling has, to date, been ignored, due to the fact that bias reduces faster than input uncertainty as the quantity of observations increase. To get a full picture of the total error in the simulation output due to input modelling it is not enough to consider IU alone. Bias can be small and therefore hard to accurately estimate without large simulation effort. But this does not mean bias is irrelevant. In context it may have a considerable impact on the error of the simulation response. For this reason we present a new diagnostic test that not only gives an estimate of bias, but has controlled power, $1-\alpha_2$, of rejecting the null hypothesis if bias is more than a relevant threshold, γ. This may be a subjective measure, a threshold of bias the decision makers would be concerned with, or it could be a proportion of IU thought to make a considerable impact on the MSE. Using a hypothesis test to check for the relevance of bias has not previously been considered; this approach could be adapted to many situations in which bias results by taking a function of a random variable.

Our approach is focused on the design of the following two-sided hypothesis test

$$\mathrm{H}_0 : \mathrm{Bias} = 0 \qquad \textit{vs.} \qquad \mathrm{H}_1 : \mathrm{Bias} \neq 0$$

with type I error rate, or size, α_1. The test will be set up to ensure controlled power, $1-\alpha_2$, of rejecting the null hypothesis when the absolute value of the true bias is greater than or equal to $\gamma > 0$. This method ensures that if, on conclusion of the test, the bias is found not to be relevant then the simulation practitioner can be confident that they do not need to consider bias further. Conversely, if a relevant bias is detected then further steps should be taken to include it in the analysis of total model risk.

We begin this paper with a discussion of the current literature in §2. In §3 the formulation of the diagnostic test is presented and an algorithm is given in §3.4. In §4.1 and §4.2 we evaluate the diagnostic test by considering simulation models with quadratic and non-quadratic response surfaces, respectively, and in §4.3 a realistic application of the method in a call centre setting is given. We conclude in §5.

2 BACKGROUND

As described in §1 we not only present a method for bias estimation, but also a diagnostic test with controlled power of identifying whether a relevant bias is present in the simulation output. This idea is common in the field of medical statistics for finding a relevant treatment effect with controlled power (Liu 1997). Although designing an experiment to satisfy certain power constraints is not a new idea, using it to create a simulation diagnostic for assessing whether bias is relevant has never before been tried.

Within our diagnostic test we make use of the delta method, giving an estimate of bias based upon a second-order Taylor Series approximation of the response function. Withers (1987) discussed this method and applied it for the purpose of bias reduction. We apply the delta method to quantify the bias in the output of a simulation model caused by input modelling; compared to other situations where a bias estimate may be required this has the additional complication of simulation noise. Alternative methods for bias quantification are the jackknife and the bootstrap. In the general case, without simulation noise, these methods were found inferior to the delta method in terms of computational efficiency in all but a few special cases where it could be said the jackknife method was comparable (Withers and Nadarajah 2014). In the context of simulation there may be a strict simulation budget, so computational efficiency can be very important.

To be able to use the delta approximation of bias we require evaluation of the Hessian matrix of second-order partial derivatives of the mean simulation response; this is simple if a closed form response function exists. But simulation models are usually unknown functions of their input parameters. Therefore to estimate these partial derivatives we propose using a central composite experimental design (CCD) to fit a response surface model as described by Montgomery (2013). This enables us to investigate the behaviour of the simulation response close to the true input parameter values.

We now present the diagnostic test for assessing whether bias due to input modelling is a relevant error about the simulation output.

3 DETECTING BIAS OF A RELEVANT SIZE

Bias due to input modelling is often small, requiring a large amount of simulation effort to accurately estimate. Although small, in context this bias may make a considerable contribution to the error about the simulation output. A practitioner may have in mind a value that they believe to be a worrying level of bias, γ; this value is informed by the output of the nominal experiment. In this section we present a hypothesis test that will detect, with controlled power, whether the absolute value of bias is greater than or equal to γ. The method still requires an estimate of bias, but it is easier to test for bias than to estimate it accurately. We therefore require less simulation effort.

Before we can present the details of the diagnostic test we first illustrate how to estimate bias due to input modelling in the presence of simulation noise and give an estimate of the variance of this bias estimator; these values come together within the diagnostic test forming the test statistic. To estimate bias we start by considering the delta method approximation. This requires evaluation of the second-order partial derivatives of the response function (Nelson 2013). In reality, the simulation response is likely to be

an unknown function of the input parameters with unknown partial derivatives. We therefore estimate the simulation response function by building a response surface model. Under the assumption that our response surface is locally quadratic, we use a CCD experimental design to fit this model, allowing estimation of the partial derivatives and therefore an estimate of the delta method approximation to bias, henceforth denoted $\widehat{b}$.

Given this bias estimator, $\widehat{b}$, the key to our approach is in controlling the power of the hypothesis test so we have a high probability of rejecting the null hypothesis when $|\widehat{b}| \geq \gamma$ is satisfied. This power is directly controlled by the variance of the bias estimator which can be reduced in two ways: by increasing the number of replications at each design point or by increasing the width of the experimental design. Note that there are limitations to how far we can spread our experimental design before our quadratic approximation fails, whereas, in theory, we could increase our simulation effort at each design point endlessly. We therefore allow for a large number of replications at each design point and use the width of the experimental design to ensure the power holds. We next present the components of our bias detection approach.

3.1 The Delta Method

Let there be L parametric input distributions to the simulation with true input parameters $\boldsymbol{\theta}^c = \{\theta_1^c, \theta_2^c, \ldots, \theta_k^c\}$; note that $k \geq L$ as some distributions may have multiple parameters. For some set of parameters $\boldsymbol{\theta}$, the output of the j^{th} replication of the simulation can be represented by

$$Y_j(\boldsymbol{\theta}) = \eta(\boldsymbol{\theta}) + \varepsilon_j,$$

where $\eta(\boldsymbol{\theta})$ is the expected value of the simulation response and we assume $\varepsilon_j \sim (0, \sigma^2(\boldsymbol{\theta}))$, for $j = 1, 2, \ldots, r$, representing the stochastic estimation error from replication to replication of the simulation.

For each of the L input distributions, $l = 1, 2, \ldots, L$, we have m_l real-world observations from which we can find the maximum likelihood estimators (MLEs), $\boldsymbol{\theta}^{mle} = \{\theta_1^{mle}, \theta_2^{mle} \ldots, \theta_k^{mle}\}$, of the input parameters. Given these estimators, bias due to input modelling, b, is defined as

$$b = \mathrm{E}[\eta(\boldsymbol{\theta}^{mle})] - \eta(\boldsymbol{\theta}^c). \tag{1}$$

Since $\boldsymbol{\theta}^c$ is unknown, we approximate bias using the delta method approach as follows. Assuming the expected simulation response, $\eta(\cdot)$, is twice continuously differentiable about $\boldsymbol{\theta}^c$, it can be expanded as a Taylor Series to second-order

$$\eta(\boldsymbol{\theta}^{mle}) \approx \eta(\boldsymbol{\theta}^c) + d(\boldsymbol{\theta}^{mle})^T \nabla\eta(\boldsymbol{\theta}^c) + \frac{1}{2!} d(\boldsymbol{\theta}^{mle})^T \, \mathrm{H}(\boldsymbol{\theta}^c) d(\boldsymbol{\theta}^{mle}),$$

where $d(\boldsymbol{\theta}^{mle}) = (\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)$ is the difference between the MLEs and true parameters, $\nabla\eta(\boldsymbol{\theta}^c)$ is the $(k \times 1)$ gradient vector of the response function and $\mathrm{H}(\boldsymbol{\theta}^c)$ represents the $(k \times k)$ Hessian matrix of second derivatives with respect to the k input parameters. Using this Taylor series expansion an estimate of bias is given by

$$\begin{aligned} b &= \mathrm{E}[\eta(\boldsymbol{\theta}^{mle})] - \eta(\boldsymbol{\theta}^c) \\ &\approx \frac{1}{2} \mathrm{E}[d(\boldsymbol{\theta}^{mle})^T \, \mathrm{H}(\boldsymbol{\theta}^c) d(\boldsymbol{\theta}^{mle})], \end{aligned}$$

as $d(\boldsymbol{\theta}^{mle}) = (\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c) \to 0$ in probability, due to the consistency of the MLEs. This gives, after some matrix manipulation,

$$b^{approx} = \frac{1}{2} \mathrm{tr}(\Omega \, \mathrm{H}(\boldsymbol{\theta}^c)), \tag{2}$$

the delta method approximation of bias, where tr() denotes the trace of a matrix and $\Omega = \text{Var}(\boldsymbol{\theta}^{mle})$. When the simulation response is an unknown function and the true input parameters $\boldsymbol{\theta}^c$ are unknown, we estimate b^{approx} to give our estimate of bias

$$\widehat{b} = \frac{1}{2}\text{tr}(\widehat{\Omega}\,\widehat{\text{H}}(\boldsymbol{\theta}^{mle})). \tag{3}$$

For this we require estimates of both the covariance matrix of the input parameters and the Hessian matrix of second-order partial derivatives. The estimated covariance matrix, $\widehat{\Omega} = \widehat{\text{Var}}(\boldsymbol{\theta}^{mle})$, will be obtained from the real-world data and is from here forward assumed known. Estimating the Hessian is a little more tricky. We chose a response surface modelling approach, quantifying the curvature of the response surface by investigating the behaviour of the function close to $\boldsymbol{\theta}^{mle}$.

3.2 Fitting a Response Surface Model

Central to our method is the further assumption that, locally to $\boldsymbol{\theta}^c$, our response surface is quadratic and can be approximated by

$$\eta(\boldsymbol{\theta}) \approx \beta_0 + \boldsymbol{\theta}^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\theta}^T\mathbf{B}\boldsymbol{\theta},$$

where $\boldsymbol{\beta}$ is the vector of coefficients belonging to the linear terms and $\mathbf{B}$ is the $(k \times k)$ matrix of coefficients belonging to the interaction and quadratic terms. We will use a CCD centred at $\boldsymbol{\theta}^{mle}$ to fit the response surface; see Figure 1 for an example of a 2-dimensional, $k = 2$, design. We chose to use a CCD because they are well known and allow the estimation of higher-order regression coefficients which could be used to check the fit of the response surface. Within the experimental design, let n_F denote the number of factorial

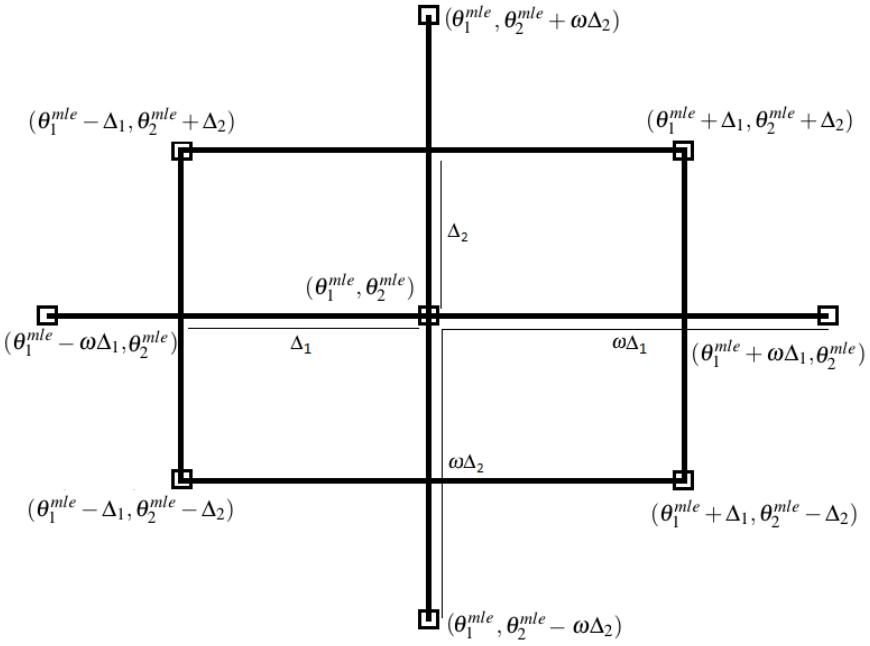

Figure 1: A CCD design with dimension $k = 2$.

points, n_A the number of axial points and n_C the number replications of the centre point. The total number of design points n is therefore $n = n_F + n_A + n_C = 2^k + 2k + n_C$, which depends on the number of input parameters, k. As suggested by Montgomery (2013) we let there be multiple design points at the centre, $n_C > 1$, allowing more information collection at the most important design point $\boldsymbol{\theta}^{mle}$. At each design

point, $i = 1,2,\ldots,n$, we complete r replications of the simulation model. The total number of replications is therefore $n \times r$.

As seen in Figure 1, we position the factorial and axial points relative to the centre point, $\boldsymbol{\theta}^{mle}$. Let Δ_i be the distance to a factorial point from the centre point in the i^{th} direction, $i = 1,2,\ldots,k$, and similarly let τ_i be the distance to the axial points. We set

$$\Delta_i = a\sqrt{\mathrm{Var}(\theta_i^{mle})} \qquad \text{and} \qquad \tau_i = \omega\Delta_i = a\omega\sqrt{\mathrm{Var}(\theta_i^{mle})},$$

where a is the number of standard deviations the factorial points are from the centre point in the i^{th} direction. Here ω is the scaled distance from the centre to the axial points; we set $\omega = \sqrt{(\sqrt{n_F n} - n_F)/2}$ as suggested by Dean and Voss (1999) for creating orthogonal designs, although we note here that due to the assumed quadratic nature of the response surface, orthogonality does not hold.

Given the averaged output of the simulation, $\bar{Y}(\hat{\boldsymbol{\theta}}_i)$, at each design point $i = 1,2,\ldots,n$, we use regression analysis to evaluate the estimates $\widehat{\mathbf{B}}$, which in turn allows us to estimate the Hessian, where

$$\hat{\mathrm{H}}(\boldsymbol{\theta}^{mle}) = \begin{bmatrix} 2\widehat{\mathrm{B}}_{11} & \widehat{\mathrm{B}}_{12} & \ldots & \widehat{\mathrm{B}}_{1k} \\ \widehat{\mathrm{B}}_{21} & 2\widehat{\mathrm{B}}_{22} & & \\ \vdots & & \ddots & \\ \widehat{\mathrm{B}}_{k1} & & & 2\widehat{\mathrm{B}}_{kk} \end{bmatrix},$$

and therefore the bias, using $\widehat{b}$, as in Equation (3). The variance of this bias estimator, conditional on the value of $\widehat{\Omega}$, can be expanded as follows

$$\begin{aligned} \mathrm{Var}(\widehat{b}) =& \mathrm{Var}\left[\frac{1}{2}\mathrm{tr}(\widehat{\Omega}\widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle}))\right] \\ =& \frac{1}{4}\mathrm{Var}\left[2\sum_{i=1}^{k}\widehat{\mathrm{B}}_{ii}\widehat{\Omega}_{ii} + \sum_{j=1}^{k}\sum_{i=1,i\neq j}^{k}\widehat{\mathrm{B}}_{ij}\widehat{\Omega}_{ij}\right] \\ =& \sum_{i=1}^{k}\sum_{j\geq i}^{k}\mathrm{Var}(\widehat{\mathrm{B}}_{ij})\widehat{\Omega}_{ij}^2 + 2\sum_{i\leq j}^{k}\sum_{l\geq m, ij<lm}^{k}\mathrm{Cov}(\widehat{\mathrm{B}}_{ij},\widehat{\mathrm{B}}_{lm})\widehat{\Omega}_{ij}\widehat{\Omega}_{lm}, \end{aligned}$$

requiring the calculation of $\mathrm{Var}(\widehat{\mathbf{B}})$, the variance-covariance matrix of regression coefficients belonging to the interaction and quadratic terms. Given that we can estimate the stochastic estimation error, $\widehat{\sigma}^2$, from the nominal experiment, this matrix has special form

$$\mathrm{Var}(\widehat{\mathrm{B}}_{ii}) = \frac{\widehat{\sigma}^2 s}{ra^4\widehat{\Omega}_{ii}^2}, \qquad \mathrm{Var}(\widehat{\mathrm{B}}_{ij}) = \frac{\widehat{\sigma}^2 f}{ra^4\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}} \quad \text{and} \quad \mathrm{Cov}(\widehat{\mathrm{B}}_{ii},\widehat{\mathrm{B}}_{jj}) = \frac{\widehat{\sigma}^2 g}{ra^4\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}},$$

which we will exploit later when it comes to setting the width of the CCD in our hypothesis test. Here, s, f and g are constants independent of a and $\widehat{\Omega}$. Note that $\mathrm{Cov}(\widehat{\mathrm{B}}_{ij},\widehat{\mathrm{B}}_{lm}) = 0$ when $i \neq j$ and $l \neq m$ if we use a CCD, therefore $\mathrm{Var}(\widehat{b})$ has the form

$$\mathrm{Var}(\widehat{b}) = \frac{\widehat{\sigma}^2}{ra^4}\left[sk + f\sum_{i=1}^{k}\sum_{j>i}^{k}\frac{\widehat{\Omega}_{ij}^2}{\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}} + gk(k-1)\right]. \tag{4}$$

This variance estimator only accounts for the variability of the Hessian as $\widehat{\Omega}$, the covariance matrix, and $\widehat{\sigma}^2$, the common variance about the simulation output, can both be estimated from the nominal experiment.

At this point we have presented a method for estimating the bias about the simulation response caused by input modelling and have also provided a variance estimate associated with it. We could stop here but as was argued at the start of this section, testing for bias is easier than finding an accurate point estimate of it and requires less simulation effort. We will now present our key idea, a diagnostic test for detecting bias of relevant size γ with controlled power.

3.3 How to Detect a Relevant Bias

We begin by considering the following hypothesis test

$$\mathrm{H}_0 : b = 0 \qquad \text{vs.} \qquad \mathrm{H}_1 : b \neq 0$$

with test statistic

$$\mathrm{T} = \frac{\hat{b}}{\sqrt{\mathrm{Var}(\hat{b})}}.$$

This hypothesis test asks the question: is bias significantly different from 0? We shall assume that

$$\frac{\hat{b} - b}{\sqrt{\mathrm{Var}(\hat{b})}} \sim \mathrm{N}(0,1) = \mathrm{Z}. \tag{5}$$

We want an experimental design where the following significance and power hold

$$\mathrm{P}[\mathrm{T} < \mathrm{Z}_{\alpha_1/2}, \mathrm{T} > \mathrm{Z}_{1-\alpha_1/2} \big| b = 0] = \alpha_1 \tag{6}$$

$$\mathrm{P}[\mathrm{T} < \mathrm{Z}_{\alpha_1/2}, \mathrm{T} > \mathrm{Z}_{1-\alpha_1/2} \big| \ |b| \geq \gamma] \geq 1 - \alpha_2 \tag{7}$$

given a relevant bias, γ, set by the practitioner. We know that Equation (6) is guaranteed by (5). Constraint (7) says that if bias is relevant we want controlled power, probability $1-\alpha_2$ of rejecting the null hypothesis. This holds when

$$\sqrt{\mathrm{Var}(\hat{b})} \leq \frac{\gamma}{Z_{1-\alpha_2} - Z_{\alpha_1/2}}. \tag{8}$$

We can therefore control the power of our experiment using the variance about our bias estimator, $\mathrm{Var}(\hat{b})$. Recall from Equation (4) that $\mathrm{Var}(\hat{b})$ is a function of the width of the CCD, set using a, and r, the number of replications at each design point, along with $\widehat{\Omega}$ and $\widehat{\sigma}^2$, constants estimated in the nominal experiment. As previously mentioned, due to the limitations on how far we can spread our design until our quadratic assumption breaks down, we choose to fix r at some appropriately large number and find the value of a where our power holds.

Returning to Equation (4) we see that a, the parameter controlling the width of the design, can be factored out of $\mathrm{Var}(\hat{b})$. Thus our problem simplifies to finding a such that Constraint (7) holds which gives

$$a \geq \left[\frac{\widehat{\sigma}^2 t^2}{r\gamma^2} \left(sk + f \sum_{i=1}^{k} \sum_{j>i}^{k} \frac{\widehat{\Omega}_{ij}^2}{\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}} + gk(k-1) \right) \right]^{\frac{1}{4}}. \tag{9}$$

Given a that satisfies Constraint (9) we can set up a CCD for use within the hypothesis test which will detect with controlled power, $1-\alpha_2$, a relevant bias, if the bias due to input modelling is truly at least γ. We now present an algorithm for the diagnostic test above.

3.4 Algorithm

Preliminary Step. Estimate $\boldsymbol{\theta}^c$ and Ω by $\boldsymbol{\theta}^{mle}$ and $\widehat{\Omega}$. Run the nominal experiment to estimate σ^2 by $\widehat{\sigma}^2$. Set γ, a bias we wish to detect, α_1 the size of the test and $1-\alpha_2$ the power.

1. Set r, the number of replications at each design point. To find a such that the power holds we must first evaluate s, f and g. Initially let $a=1$, noting that any positive value would suffice; create the $\left(n \times \left(1+2k+\frac{k(k-1)}{2}\right)\right)$ design matrix X, centred at $(0,0,\ldots,0)$ for convenience with $\Delta_i = a\sqrt{\mathrm{Var}(\theta_i^{mle})}$ and $\tau_i = \omega\Delta_i$, for $i=1,2,\ldots,k$. Given X, evaluate s, f and g as follows

$$s = (X^{\mathrm{T}}X)^{-1}_{[\frac{(k+1)(k+2)}{2},\frac{(k+1)(k+2)}{2}]}\Delta_k^4, \qquad f = (X^{\mathrm{T}}X)^{-1}_{[k+2,k+2]}\Delta_1^2\,\Delta_2^2,$$
$$g = (X^{\mathrm{T}}X)^{-1}_{[\frac{(k+1)(k+2)}{2}-1,\frac{(k+1)(k+2)}{2}]}\Delta_{k-1}^2\,\Delta_k^2$$

   and thus evaluate the value of a for which the power holds using Equation (9).
2. Re-build the design matrix X given a for which the power holds.
3. For i in 1 to n: Run r replications of the simulation at design point, $\boldsymbol{\theta}_i$, corresponding to row i of the design matrix; average over the r replications to find $\bar{Y}(\boldsymbol{\theta}_i)$.
4. Using the simulation output from the n design points, $\bar{Y}(\boldsymbol{\theta}_i)$ for $i=1,2,\ldots,n$, extract $\widehat{\mathbf{B}}$ from $(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\bar{\boldsymbol{Y}}(\boldsymbol{\theta})$, giving $\widehat{\mathrm{B}}_{11},\widehat{\mathrm{B}}_{12},\ldots,\widehat{\mathrm{B}}_{(k-1)k},\widehat{\mathrm{B}}_{kk}$.
5. Evaluate $\widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle})$; thus evaluate $\widehat{b}$ and $\mathrm{Var}(\widehat{b})$.
6. Calculate the test statistic, $\mathrm{T} = \frac{\widehat{b}}{\sqrt{\mathrm{Var}(\widehat{b})}}$. If $|\mathrm{T}| \geq Z_{1-\alpha_1/2}$ is satisfied reject the null hypothesis.

4 EMPIRICAL EVALUATION

In this section we will evaluate the diagnostic test presented in this paper by considering how well the power holds: firstly in a system where the simulation response surface is truly quadratic, and then for a tractable $M/M/1/C$ queueing model. Finally we illustrate the use of the diagnostic test in a realistic call centre setting to show how in practice the diagnostic test could be used, and suggest follow-up actions for when a relevant bias is found.

4.1 A Truly Quadratic Model

Consider a quadratic response function. As an example, when $k=2$ let the response function be given by

$$\eta(\boldsymbol{\theta}) = 2+3\theta_1+\theta_2+4\theta_1\theta_2+\theta_1^2+2\theta_2^2. \tag{10}$$

Here we let θ_1^c and θ_2^c be the true mean parameters from the following bivariate normal distribution

$$X_1, X_2 \sim \mathcal{N}\left((\theta_1^c,\theta_2^c)^T, \begin{pmatrix} \xi_1^2 & 0 \\ 0 & \xi_2^2 \end{pmatrix}\right)$$

with $\mathrm{Cov}(\theta_1^{mle},\theta_2^{mle}) = 0$ and $\mathrm{Var}(\theta_i^{mle}) = \xi_i^2/m$. Given this response function we know the Hessian matrix exactly, therefore the delta approximation gives $b^{approx} = \xi_1^2/m + 2\xi_2^2/m$ which is exact since (10) is quadratic.

Let us now assume that the response function, $\eta(\boldsymbol{\theta})$, is unknown to us. We wish to evaluate the performance of the diagnostic test when the underlying response surface is truly quadratic. To do this we investigate how well the power holds when the relevant bias, γ, is set equal to b^{approx}, the true bias in this quadratic case. For this experiment let the power be set to $1-\alpha_2 = 0.8$. We therefore wish to illustrate our diagnostic test having probability 0.8 of rejecting the null hypothesis when $\gamma = b^{approx}$.

Table 1: How power holds when $\gamma = b^{approx}$ given a truly quadratic response function.

θ_1^c	θ_2^c	r	m	b^{approx} $(=\gamma)$	$\widehat{\mathrm{E}}[\widehat{b}]$	$\widehat{\mathrm{Var}}[\widehat{b}]$	$\widehat{p}$
5	2	1000	40	0.2125	0.2111	4.52×10^{-3}	0.79

To show our diagnostic test has this desired power we run a macro-experiment, repeating the diagnostic test $G = 1000$ times. An estimate of power will be given by the proportion of times the null hypothesis is rejected; we denote this estimate $\widehat{p}$. In Table 1 $\widehat{p}$ is recorded along with $\widehat{\mathrm{E}}[\widehat{b}]$ and $\widehat{\mathrm{Var}}[\widehat{b}]$, the sample mean and variance of the bias estimates recorded over the $G = 1000$ macro-replications. Also reported is b^{approx}, the true bias in this quadratic example, which we set equal to γ, the relevant bias.

To complete the diagnostic test we use the methods presented in §3.1 to §3.3. Given true input parameters $\theta_1^c = 5$ and $\theta_2^c = 2$ with $\xi_1^2 = 2$ and $\xi_2^2 = 1.5$, $m = 40$ observations of X_1 and X_2 were generated from the bivariate normal distribution and used to estimate the MLEs, $\boldsymbol{\theta}^{mle}$, and $\widehat{\Omega}$. We set the number of replications to $r = 1000$ then built a response surface model using a CCD centred at $\boldsymbol{\theta}^{mle}$ with width set using $a = 0.283$ selected to ensure a power of $1-\alpha_2 = 0.8$. In each replication we ran the simulation by adding $\mathcal{N}(0, 0.01)$ noise to (10). Given the response surface model the bias estimator, $\widehat{b}$, and its variance, $\mathrm{Var}(\widehat{b})$, could be evaluated enabling the calculation of the test statistic, T, and the conclusion of the diagnostic test. This process was repeated $G = 1000$ times to gain the results shown in Table 1.

In Table 1 we see that when the response function is truly quadratic, the diagnostic test holds power very close to $1-\alpha_2 = 0.8$ as desired. We also see that the average of the bias estimates, $\widehat{\mathrm{E}}[\widehat{b}]$, is very close to the true bias.

We will now investigate how well the diagnostic test performs when the quadratic assumption does not hold, by studying a tractable $M/M/1/C$ queueing model.

4.2 $M/M/1/C$ Queueing Model

Consider an $M/M/1/C$ queueing model with true arrival rate θ_1^c, service rate θ_2^c and finite capacity C. Here inter-arrival times of customers, A_i, follow an exponential distribution $A_i \sim \mathrm{Exp}(\theta_1^c)$, as do the service times, $S_i \sim \mathrm{Exp}(\theta_2^c)$, for $i = 1, 2, \ldots, m$ observations. For this queueing model the expected number of customers in the system, $\mathrm{E}[Y|\boldsymbol{\theta}]$, can be expressed in closed form

$$\eta(\boldsymbol{\theta}) = \mathrm{E}[Y|\boldsymbol{\theta}] = \frac{\theta_1}{\theta_2 - \theta_1} - \frac{(C+1)\theta_1^{C+1}}{\theta_2^{C+1} - \theta_1^{C+1}}. \tag{11}$$

It is therefore possible to derive the second-order partial derivatives yielding $\mathrm{H}(\boldsymbol{\theta}^c)$; this allows the evaluation of b^{approx}, the delta method approximation of bias.

We shall now, for the purpose of the experiment, assume that the true response function, Equation (11), is unknown. We want to evaluate the quality of our diagnostic test for detecting a relevant bias when the response function is not truly quadratic. To do this we will look at both the $M/M/1/10$ and $M/M/1/100$ queueing models over a number of parameter settings to see how well the power, set at $1-\alpha_2 = 0.8$, holds when relevant bias, γ, is set equal to the delta approximation of bias b^{approx}. As before, to measure the power we record the proportion of times the null hypothesis was rejected over $G = 1000$ macro-replications of the diagnostic test, $\widehat{p}$. The results of the experiments are given in Table 2.

The diagnostic test was completed as follows. Instead of running a nominal experiment we used the true input distributions to generate m observations from the arrival and service distributions, A_i, S_i for $i = 1, 2, \ldots, m$, then estimated the MLEs, $\boldsymbol{\theta}^{mle}$, and the covariance matrix, $\widehat{\Omega}$; we know that $\mathrm{Cov}(\theta_1^{mle}, \theta_2^{mle}) = 0$. Also, rather than directly simulating the $M/M/1/C$ queue we add $\mathcal{N}(0, 0.05)$ noise to (11) for each replication. The number of replications to be run at each design point was set to $r = 500$ allowing the identification of the value of a required for the power to hold at $1-\alpha_2 = 0.8$. A CCD design, centred at $\boldsymbol{\theta}^{mle}$, was then created using a to set the distance to the design points. Replications of the simulation

Table 2: How power holds when $\gamma = b^{approx}$ given an $M/M/1/C$ queueing model.

			$M/M/1/10$			$M/M/1/100$		
Exp	$\frac{\theta_1^c}{\theta_2^c}$	m	b^{approx}	$\widehat{\mathrm{E}}[\widehat{b}]$	$\widehat{p}$	b^{approx}	$\widehat{\mathrm{E}}[\widehat{b}]$	$\widehat{p}$
1	0.25	40	0.019	0.024 (4.98×10^{-4})	0.766	0.019	0.025 (6.26×10^{-4})	0.787
2	0.25	100	0.007	0.008 (1.25×10^{-4})	0.79	0.007	0.009 (1.30×10^{-4})	0.789
3	0.50	40	0.134	0.174 (3.86×10^{-3})	0.704	0.150	0.855 (1.73×10^{-1})	0.659
4	0.50	100	0.053	0.063 (1.07×10^{-3})	0.775	0.060	0.085 (2.98×10^{-3})	0.741
5	0.50	1000	0.005	0.006 (6.77×10^{-5})	0.818	0.006	0.007 (7.80×10^{-5})	0.822
6	0.83	100	0.164	0.114 (3.09×10^{-3})	0.611	3.300	6.623 (1.24×10)	0.611
7	0.83	1000	0.016	0.015 (1.98×10^{-4})	0.712	0.330	0.570 (2.64×10^{-2})	0.713
8	0.83	5000	0.003	0.003 (3.82×10^{-5})	0.777	0.066	0.071 (1.08×10^{-3})	0.765

were run at each design point and the response surface fitted allowing evaluation of $\widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle})$, the estimated Hessian matrix. We were therefore able to estimate the delta approximation of bias, $\widehat{b}$, and its variance, $\mathrm{Var}[\widehat{b}]$, allowing us to calculate the test statistic and conclude the hypothesis test. This process was repeated over $G = 1000$ macro-replications giving $\widehat{p}$ and $\widehat{\mathrm{E}}[\widehat{b}]$, the average of the bias estimates, both are recorded in Table 2.

In Table 2, we see that across all experiments, whether $C = 10$ or 100, as the amount of input data is increased $\widehat{p}$ gets closer to the desired power $1-\alpha_2 = 0.8$ and the average bias estimate $\widehat{\mathrm{E}}[\widehat{b}]$ gets closer to the delta approximation b^{approx}. Both parameter estimates improve due to the the increase in information which sees $\boldsymbol{\theta}^{mle}$ get closer to $\boldsymbol{\theta}^c$, the true input parameters. This is important in our method as, ideally, we would centre our CCD at $\boldsymbol{\theta}^c$ to find the curvature of the response function at that point, $\mathrm{H}(\boldsymbol{\theta}^c)$.

Experiments 6, 7 and 8 look at the system under high traffic intensity, $\rho = \theta_1^c/\theta_2^c = 0.833$. In Experiment 6, where $m = 40$, we saw a reasonably high proportion of instances ($\approx 10\%$) where the estimated traffic intensity exceeded 1, i.e. $\rho = \theta_1^{mle}/\theta_2^{mle} > 1$. When this occurs the number of people in the queue will increase up to capacity and remain around that level. The behaviour of the response surface in these cases is not quadratic and therefore the delta method does not perform well which is reflected in the average bias estimate, $\widehat{\mathrm{E}}[\widehat{b}]$, and power, $\hat{p}$. One way to fix this problem is to collect more data, m, until $\theta_1^{mle}/\theta_2^{mle} < 1$ consistently, as we did in Experiments 7 and 8 where the bias estimate $\widehat{\mathrm{E}}[\widehat{b}]$ gets closer to the delta approximation.

This problem is not unique to bias estimation: it will occur in any simulation model with finite capacity and traffic intensity close to 1. If the amount of data available is small and we cannot accurately estimate the input parameters it is easy to conclude that a system will become saturated when in reality it might not.

In experiments 6, 7 and 8, where a high traffic intensity was investigated, we see the effect of the shape of the true response surface on how well the power holds. The shape of the response surface is driven by the capacity, C. This directly links to how closely $\boldsymbol{\theta}^c$ can be estimated by $\boldsymbol{\theta}^{mle}$. In Figure 3 we see that for the $M/M/1/100$ queue, with higher capacity, there is a more dramatic change in the response surface for small changes of θ_1 and θ_2 than there is for the lower capacity, $M/M/1/10$, queue seen in Figure 2. Close to $\rho = 1$, where the response surface changes more dramatically, more observations, m, are needed to ensure we are estimating the Hessian, $\mathrm{H}(\boldsymbol{\theta}^c)$, close enough to $\boldsymbol{\theta}^c$ to capture the true curvature at that point. This could also be affected by the variability of the MLEs; when the variance is large even if we have $\boldsymbol{\theta}^{mle}$ close to $\boldsymbol{\theta}^c$ on average, we could see large variability in the response from replication to replication. In the higher capacity system small changes in the inputs have a larger effect on the simulation output which is used to fit the response surface and therefore estimate the Hessian. For the lower capacity queueing model the distance between $\boldsymbol{\theta}^{mle}$ and $\boldsymbol{\theta}^c$ has a less pronounced effect on the simulation response as the response surface changes.

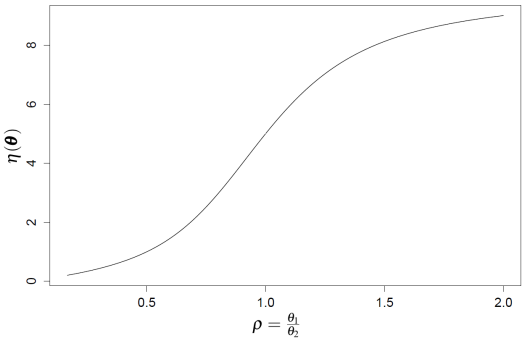

Figure 2: $M/M/1/10$

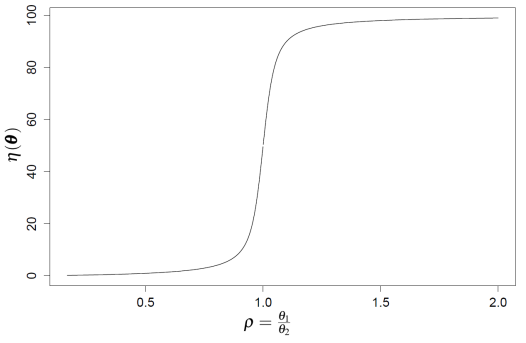

Figure 3: $M/M/1/100$

We also note that for the $M/M/1/100$ queueing model, in Experiments 3, 6 and 7 $\widehat{p}$ is lower than the desired power of 0.8 but the average of the bias estimates in these cases, $\widehat{\mathrm{E}}[\widehat{b}]$, is higher than b^{approx}. Intuitively, this seems contradictory as we would expect to reject the null hypothesis more often if bias is much more extreme than $\gamma = b^{approx}$. In these cases we also see $\widehat{\mathrm{E}}[\widehat{b}]$ has large standard error. Investigating the test statistics over the $G = 1000$ macro replications, using Q-Q plots, illustrated that these were the cases where the distribution of the test statistics was far from the assumed normal distribution. In Experiments 4, 5 and 8 given more input data the normality assumption held much better. For the $M/M/1/10$ queueing model the normality assumption held well in all cases. This again illustrates the importance of centring the CCD close to $\boldsymbol{\theta}^c$, especially when there is a sharp change in the shape of the response surface.

As an aside we also considered the trade off between the variables a and r, used to set the width of the experimental design. To improve the quadratic assumption it is tempting to shrink a and increase the number of replications at each design point to ensure the power still holds. This is very expensive computationally; to halve a, and thus the width of the design, in the experiments above we would have had to increase the number of replications at each design point to $r = 8000$. Looking at the experiments above we saw little improvement on the estimated power $\widehat{p}$ from halving a. This is because shrinking the width of the design would only be helpful if the CCD was centred very close to $\boldsymbol{\theta}^c$; no amount of computational effort will improve our estimate of $\widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle})$ if the design is centred at $\boldsymbol{\theta}^{mle}$ far from $\boldsymbol{\theta}^c$.

4.3 A Realistic Example - NHS 111 Healthcare Call Centre

We will now illustrate our bias detection diagnostic on the simulation of a real-world system with a non-stationary input process. The nominal experiment is based on observations of an NHS111 healthcare call centre simulated using an $M(t)/G/S(t)$ queueing model with a piecewise-constant Poisson arrival process. Using the methods discussed by Morgan et al. (2016) we were able to quantify the total IU about the expected waiting time of callers, E(WTime). The system discussed is staffed to meet the NHS target level of service, P(Wait $>$ 1 min) < 0.05; in the nominal experiment an estimate of the expected waiting time of customers was found to be E(WTime) $= 0.0674$ minutes. In the following tests we use the value of IU, defined to be $\mathrm{Var}[\eta(\boldsymbol{\theta}^{mle})]$, to guide our choice of γ. Let $\gamma = \sqrt{\upsilon \times \mathrm{IU}}$ where $0 < \upsilon < 1$. This gives us the threshold bias thought to have an important effect on the MSE. Estimates of $\boldsymbol{\theta}^c, \Omega$ and σ^2, given by $\boldsymbol{\theta}^{mle}, \widehat{\Omega}$ and $\widehat{\sigma}^2$, were collected within the nominal experiment. In practice the estimate of the simulation estimation error, $\widehat{\sigma}^2$, could have been used here to aid the choice of r, the number of replications at each design point. For example, if we had a noisy simulation a large value of r would be required. In the following experiments $r = 500$ replications were performed at each design point.

Given observations of the NHS111 healthcare call centre system we conducted two experiments with different levels of input data. We denote by m_1 the number of days of observations of the arrival process and m_2 the number of service time observations. The desired power was set equal to $1 - \alpha_2 = 0.8$ and the

Table 3: Results of the bias detection test for an NHS 111 healthcare call centre when considering expected waiting time of callers, E(WTime).

Exp	m_1	m_2	IU $(\text{Var}[\eta(\boldsymbol{\theta}^c)])$	γ $(\upsilon = 0.3)$	$\widehat{b}$	$\text{Var}[\widehat{b}]$	p-value
1	10	20068	4.336×10^{-5}	3.61×10^{-3}	6.03×10^{-3}	1.638^{-6}	1.23×10^{-6}
2	26	52711	1.907×10^{-5}	2.39×10^{-3}	4.04×10^{-3}	7.215×10^{-7}	9.97×10^{-7}

size to $\alpha_1 = 0.05$. For these experiments the relevant bias, γ, was set using $\upsilon = 0.3$, meaning we consider bias squared higher than 0.3 times the value of IU to be concerning.

From Table 3 we see that in Experiment 1, where we considered a smaller number of observations of the arrival process, m_1 =10 days, we have sufficient evidence to reject the null hypothesis as the p-value is less than $\alpha_1 = 0.05$. We can therefore conclude that the amount of bias due to input modelling about the expected waiting time of callers, E(WTime), is significant. In addition, using the output of the diagnostic test we can calculate a confidence interval about our estimate of bias as $\widehat{b} \pm Z_{1-\alpha_1}\sqrt{\text{Var}(\widehat{b})}$. This allows us to make a statement about how $\widehat{b}$ compares to the relevant bias γ. In Experiment 1 a 90% confidence interval about $\widehat{b}$ is given by $(3.92\times10^{-3}, 8.14\times10^{-3})$. This does not contain γ, the relevant bias, and we can therefore conclude at the 90% significance level that bias due to input modelling is more than γ in this system. In Experiment 2 we observed $m_1 = 26$ days of observations of the arrival process which saw a reduction in IU and therefore γ, but again we were able to reject the null. A 90% confidence interval about $\widehat{b}$ is given by $(2.64\times10^{-3}, 5.44\times10^{-3})$ which again does not contain γ. In both cases the practitioner can be confident that the level of bias due to input modelling is high enough for them to be concerned.

When the bias is relevant, as is the case here, it should be taken into account in assessing the total uncertainty about the simulation output due to input modelling. This will allow more informed decisions to be made. At this point it could be considered sensible to spend more computational effort running the delta approximation alone to obtain a more accurate estimate of the bias due to input modelling.

The practitioner may, alternatively, wish to reduce bias to a level that does not concern them by collecting more input data. Changing the number of intervals describing the piecewise-constant Poisson arrival process may also have an affect on the bias due to input modelling. Morgan et al. (2016) used change point analysis as a pre-processing step in their IU quantification method. This aided the choice of arrival intervals but did not guarantee an arrival function that represented the true arrival process well or that had minimal error due to input modelling. Our method now provides the bias estimate needed to be able to compare two arrival functions in terms of the MSE due to input modelling.

5 CONCLUSION

This paper presents a diagnostic test with controlled power of detecting bias due to input modelling of a relevant size in simulation models.

Within the diagnostic test the experimental design is centred at $\boldsymbol{\theta}^{mle}$ our best estimate of the true input parameters, $\boldsymbol{\theta}^c$. When the response surface, at the point $\boldsymbol{\theta}^c$, is sensitive to the input parameters we found more input data was required to estimate the delta approximation of bias well and retain the desired power. Although the problem here seems to be the discrepancy between $\boldsymbol{\theta}^{mle}$ and $\boldsymbol{\theta}^c$ it could be that the quadratic assumption is not satisfactory; this assumption is always approximate with small samples, in which case a higher-order approximation should be used.

Although we used the CCD within our method we acknowledge that we are restricted to moderate dimensionality at this point and that fractional factorial designs could be exploited here to improve scalability and computational efficiency of the test. We also note that many open questions still remain in this area, for example, how large should we set r the number of replications at each design point to control the variability of our bias estimator? Or how do we optimally set the experimental design parameters ω and n_C? Even given these follow-up questions we have shown that for sensible, but arbitrary, choice of design

parameters our method is a good step forward to being able to detect when bias due to input modelling is of a concerning size.

REFERENCES

Cheng, R. C., and W. Holland. 1997. "Sensitivity of Computer Simulation Experiments to Errors in Input Data". *Journal of Statistical Computation and Simulation* 57 (1-4): 219–241.

Dean, A., and D. Voss. 1999. *Response Surface Methodology*. Springer-Verlag, New York.

Liu, W. 1997. "On Some Sample Size Formulae for Controlling both Size and Sower in Clinical Trials". *Journal of the Royal Statistical Society: Series D (The Statistician)* 46 (2): 238–251.

Montgomery, D. C. 2013. *Design and Analysis of Experiments*. John Wiley & Sons.

Morgan, L. E., A. C. Titman, D. J. Worthington, and B. L. Nelson. 2016. "Input Uncertainty Quantification for Simulation Models with Piecewise-Constant Non-Stationary Poisson Arrival Processes". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 370–381. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Nelson, B. 2013. *Foundations and Methods of Stochastic Simulation: A First Course*. Springer Science & Business Media.

Song, E., and B. L. Nelson. 2015. "Quickly Assessing Contributions to Input Uncertainty". *IIE Transactions* 47:893–909.

Withers, C. S. 1987. "Bias Reduction by Taylor Series". *Communications in Statistics-Theory and Methods* 16 (8): 2369–2383.

Withers, C. S., and S. Nadarajah. 2014. "Bias Reduction: The Delta Method versus the Jackknife and the Bootstrap.". *Pakistan Journal of Statistics* 30 (1): 143–151.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the EPSRC funded EP/L015692/1 STOR-i Centre for Doctoral Training, NSF Grant CMMI-1068473 and GOALI sponsor Simio LLC. We also thank Bruce Ankenman for insightful discussion and suggestions.

AUTHOR BIOGRAPHIES

LUCY E. MORGAN is a Ph.D. student of the Statistics and Operational Research Centre for Doctoral Training in Partnership with Industry at Lancaster University. Her research interests are input uncertainty in simulation models and arrival process modelling. Her email address is l.e.morgan@lancaster.ac.uk.

BARRY L. NELSON is the Walter P. Murphy Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University and a Distinguished Visiting Scholar in the Lancaster University Management School. He is a Fellow of INFORMS and IIE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. His e-mail address is nelsonb@northwestern.edu.

ANDREW C. TITMAN is currently a Lecturer in Statistics in the Department of Mathematics and Statistics at Lancaster University. His research interests include survival and event history analysis and latent variable modelling. His email address is a.titman@lancaster.ac.uk.

DAVID J. WORTHINGTON is a senior lecturer in Operational Research in the Department of Management Science in Lancaster University Management School. He researches the modelling and management of time-dependent queueing systems. His email address is d.worthington@lancaster.ac.uk.