European Journal of Operational Research 279 (2019) 869–881

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

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 (STOR-i) Lancaster University Lancaster, LA1 4YR, UK

ARTICLE INFO

Article history:
Received 31 May 2018
Accepted 3 June 2019
Available online 8 June 2019

Keywords:
Simulation
Bias
Uncertainty
Input modelling

ABSTRACT

This is the first paper to approach the problem of bias in the output of a stochastic simulation due to using input distributions whose parameters were estimated from real-world data. We consider, in particular, the bias in simulation-based estimators of the expected value (long-run average) of the real-world system performance; this bias will be present even if one employs unbiased estimators of the input distribution parameters due to the (typically) nonlinear relationship between these parameters and the output response. To date this bias has been assumed to be negligible because it decreases rapidly as the quantity of real-world input data increases. While true asymptotically, this property does not imply that the bias is actually small when, as is always the case, data are finite. We present a delta-method approach to bias estimation that evaluates the nonlinearity of the expected-value performance surface as a function of the input-model parameters. Since this response surface is unknown, we propose an innovative experimental design to fit a response-surface model that facilitates a test for detecting a bias of a relevant size with specified power. We evaluate the method using controlled experiments, and demonstrate it through a realistic case study concerning a healthcare call centre.

1. Introduction

In stochastic simulation the "stochastic" element of the simulation comes from the input models that drive it. In this paper we focus on parametric input models, probability distributions or stochastic processes that are estimated from observations of the real-world system of interest. Since we can only ever collect a finite number of observations, error, with respect to what the simulation says about the real-world system performance, is inevitable.

In this paper 'response' means the expected value of a simulated output performance measure. Error caused by input modelling can be broken down as MSE = Variance + Bias^2; that is, the mean squared error (MSE) due to input modelling is made up of the variability of the simulation response caused by input modelling, known in the literature as input uncertainty variance (IU variance), and the squared bias due to input modelling. Barton (2012) explains that, even in very reasonable simulation scenarios, analysis of the response of interest can be very different when error due to input modelling is included. Barton (2012) was referring to the IU variance, but the same idea holds for the bias due to input modelling. In simulation models where a large number of replications of the simulation are completed, effectively driving out the inherent simulation noise caused by random-variate generation, ignoring the input modelling uncertainty can lead to overconfidence in the simulation results. Underestimating the error of the simulation response is dangerous, especially when this output may be used to guide important decisions about a real-world system.

To date the main focus of research in this area has been on IU variance quantification, while the bias caused by input modelling has been virtually ignored. This was partially justified by the knowledge that, as the number of real-world observations of the input models increases, the bias due to input modelling decreases faster than the input uncertainty: given m observations of an input model, it is known that the IU variance is $O(1/m)$, whereas the bias squared due to input modelling is typically $O(1/m^2)$ (Nelson, 2013). Despite this, the bias can still be substantial for finite m. Since in reality we can never collect an infinite number of observations, bias should not be ignored.

To facilitate understanding, we consider the simulation of a healthcare call centre. More specifically, we look at the UK National Health Service (NHS) 111 system. The NHS 111 system was designed to take some of the strain from other healthcare systems in the UK, for example, emergency departments and doctors' surgeries. Ringing NHS 111 allows a caller to talk to a healthcare professional who can advise them on what care they need. The NHS 111 call centre can be represented as a stochastic queueing model

* Corresponding author.
E-mail addresses: l.e.morgan@lancaster.ac.uk (L.E. Morgan), nelsonb@northwestern.edu (B.L. Nelson), a.titman@lancaster.ac.uk (A.C. Titman), d.worthington@lancaster.ac.uk (D.J. Worthington).

https://doi.org/10.1016/j.ejor.2019.06.003
0377-2217/

with a non-stationary arrival process and a stationary service distribution. Since we have only a finite number of observations from which to estimate these input models, they are not correct; this error propagates through the NHS 111 simulation model to the performance measures of interest.

This paper presents a delta-method approach to estimating the bias caused by input modelling in stochastic simulation. The delta-method is based on a second-order Taylor series approximation and therefore requires the quantification of the second-order partial derivatives of the response surface. In simulation, the response of interest is most often an unknown function of its input models which means we cannot directly evaluate its derivatives. We therefore propose the use of an experimental design to fit a response surface model from which the second-order partial derivatives can be estimated.

As a key feature of this paper, we also present a bias detection test with controlled power for detecting a bias due to input modelling greater than a pre-chosen value, γ, considered to be a bias of a relevant size. In this way when the bias is small, and therefore not of concern to us, we require less computational effort to conclude that the bias is not significantly different from zero than to accurately estimate it. Also, when the bias is large, i.e., greater than γ, we have a high probability of detecting it. In Section 3.1 we describe a novel way in which we construct the experimental design used to estimate the response surface, which allows a practitioner to easily control the power of the bias detection test.

The bias detection test also hinges on our choice of a "bias of a relevant size". When there is no clear choice for γ from the problem context, we propose using the estimated value of the IU variance as a benchmark: if the bias is a small fraction of the IU variance, then it contributes little to the overall MSE, while if it is a large fraction of the IU variance then it should not be ignored. In Section 4.2 the IU variance is used to guide the choice of the relevant bias, γ, for the NHS 111 system.

We begin this paper with a discussion of the current literature in Section 2. In Section 3 we present our delta-method approach to bias estimation and the diagnostic test along with an algorithm to aid implementation. In Section 4.1 we complete a controlled experiment to evaluate the diagnostic test for response functions with different forms, under varying numbers of observations and replications; and in Section 4.2 a realistic application of the method in the NHS 111 system is given. We conclude in Section 5. All proofs are left to the appendix.

A preliminary proposal of the ideas presented here appeared in Morgan, Titman, Worthington, and Nelson (2017), but it did not contain the key supporting theory: the proof that asymptotically the delta approximation of bias, scaled by the number of observations, converges to the scaled true bias; the proof that, under certain assumptions, the scaled estimate of the delta approximation of bias converges to the scaled delta approximation of bias; or the proof that, without significant simulation effort, the variability of the jackknife estimator of bias can easily be obscured by simulation noise.

2. Background

To date, estimating the IU variance has been the main focus of research in quantifying error caused by input modelling. See Song, Nelson, and Pegden (2014) for a careful definition and discussion of IU variance quantification techniques. A number of methods for quantifying the IU variance in simulation models exist covering both frequentist and Bayesian methodologies (Barton, 2012). Of these, Cheng and Holland (1997) present a delta-method approach for simulation models with time-homogeneous parametric input distributions; this was extended by Morgan, Titman, Worthington, and Nelson (2016) for simulation models with piecewise-constant non-stationary Poisson arrival processes. In Section 4 these two methods will be used to estimate the IU variance and thus guide our choice of a relevant bias.

When one refers to quantifying the 'bias' it is typically the bias of an estimator of a population parameter given a sample of data, averaged over the distribution of possible samples. In our computer-simulation context this bias is also averaged over the natural noise due to generating samples of the stochastic inputs. Stated differently, our estimator is a function of both real-world and simulated sampling. Standard methods for bias quantification are the jackknife and the bootstrap (Efron, 1982), with the jackknife often considered the go-to choice. However, for bias estimation without simulation noise, Withers and Nadarajah (2014) found both the jackknife and the bootstrap are 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. When there is also simulation noise, the number of simulation replications required to mitigate it for the jackknife grows as $O(m^2)$, meaning that the simulation effort could become prohibitive or an estimate of the bias could be obscured by the simulation noise when m is large; for a proof of this result see Appendix A. For a review of the conditions under which the delta-method approximation is accurate see Oehlert (1992).

The delta-method requires the second-order partial derivatives of the expected value of the simulation response. Since the expected value of the simulation response is not known, we propose using an experimental design to fit a response surface model of it. To allow estimation of the derivatives of the response surface, we assume a simple type of meta-model, namely, a second-order polynomial. To estimate its second-order terms, we use a central composite design (CCD), which includes a Resolution V, or higher, experimental design; see Montgomery (2013).

The CCD is easy to understand and meets the design resolution requirement, but does suffer in terms of scalability, requiring an exponentially increasing number of design points as the number of input parameters increases. Fractional factorial designs are one way of reducing the number of design points required to fit a response surface. However, few efficient generators exist for creating Resolution V fractional factorial designs with a large number of inputs. An exception is the method of Sanchez and Sanchez (2005) which we use to reduce the number of design points needed to support the quadratic response surface. This method can generate designs with over 120 inputs. Methods for generating Resolution V fractional designs are also discussed by Montgomery (2013) and Box, Hunter, and Hunter (1978) but the allowable number of inputs within these design generators is limited.

Neither quantification nor detection of the bias due to input modelling have previously been considered. In the following section we present the methodology behind our delta-method estimate of the bias due to input modelling and our bias detection test.

3. Detecting bias of a relevant size

Let there be L parametric input distributions that drive the simulation with, $k \geq L$, true input parameters, $\boldsymbol{\theta}^c = \{\theta_1^c, \theta_2^c, \ldots, \theta_k^c\}$. For any set of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$, we denote the output of the jth replication of the simulation as $Y_j(\boldsymbol{\theta}) = \eta(\boldsymbol{\theta}) + \epsilon_j$, where $\eta(\boldsymbol{\theta})$ is the expected value of the simulation output of interest; this could be, for example, the expected fraction of callers that have to wait more than 1 minute to be served.

Let r denote the total number of replications; here we assume ϵ_j, for $j = 1, 2, \ldots, r$, are i.i.d. random variables, with mean zero and variance σ^2, that represent the stochastic estimation error arising within each replication of the simulation, $\epsilon_j \sim$ i.i.d. $(0, \sigma^2)$.

In this contribution we assume that ϵ_j is unaffected by the choice of $\boldsymbol{\theta}$, i.e. the simulation noise is homogeneous. In reality this is likely to be false, but note that the Hessian estimator introduced later to estimate and detect the bias remains valid even given a failure of this assumption. A lower variance estimator could be achieved by taking into account the heterogeneous simulation noise in the experiment design, but this would add significant additional complexity to the method. Also note that the experiment design is used to construct a local (not global) response surface model. We therefore expect the mean response surface and the output variance around it to vary less as the experimental design gets tighter around $\boldsymbol{\theta}^{mle}$.

For each of the $l = 1, 2, \ldots, L$ input distributions we have m_l real-world observations from which we find the maximum likelihood estimators (MLEs) of the input parameters, $\boldsymbol{\theta}^{mle} = \{\theta_1^{mle}, \theta_2^{mle} \ldots, \theta_k^{mle}\}$. By averaging over the r replications of the simulation, driven by $\boldsymbol{\theta}^{mle}$, we gain an estimate of the output performance measure of interest. We call this the *nominal experiment*. We can reduce the stochastic estimation error about our response of interest through further replications of the simulation, but this has no effect on the error due to input modelling which is only affected by $m_1, m_2, \ldots, m_L$.

For the NHS 111 system let $\boldsymbol{\theta}^c$ be the unknown parameters describing the true arrival process and service distribution, and $\boldsymbol{\theta}^{mle}$ be the MLEs of these parameters. The MLEs are estimated from service time and arrival count observations. In total there are m arrivals, and assuming a service time is recorded for each arrival, m service time observations. For any set of parameters $\boldsymbol{\theta}$, the performance measure of interest in the NHS 111 system, $\eta(\boldsymbol{\theta})$, is the expected waiting time of callers. For each replication $Y_j(\boldsymbol{\theta})$ is the average of the waiting times observed in that replication.

The bias due to input modelling arises because we only have a finite number of observations of the real-world system from which to estimate $\boldsymbol{\theta}^c$. This type of bias describes how far, on average, our simulation response is from the real-world performance given the error that arises from estimating the input models. Specifically

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

where the expectation is with respect to the sampling distribution of $\boldsymbol{\theta}^{mle}$. When the simulation response is non-linear in $\boldsymbol{\theta}$, as is usually the case, this bias will always arise; we now approximate it using the delta-method in an innovative way.

Assuming the expected simulation response, $\eta(\cdot)$, is at least twice continuously differentiable about $\boldsymbol{\theta}^c$ it can be expanded as a Taylor series to second-order

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

where $d(\boldsymbol{\theta}^{mle}) = (\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)$ is the difference between the MLEs and the true parameters, $\nabla\eta(\boldsymbol{\theta}^c)$ is the $(k \times 1)$ gradient vector and $\mathrm{H}(\boldsymbol{\theta}^c)$ is the $(k \times k)$ Hessian matrix of the response function. Note that the Hessian matrix, $\mathrm{H}(\boldsymbol{\theta}^c)$, is composed of the second-order partial derivatives with respect to the k input parameters, and approximates the curvature of the response surface. To ease explanation, let m be the common number of observations collected from each of the L input models. The following results hold in slightly modified form for $m_1 \neq m_2 \neq \cdots \neq m_L$, provided $m_i / \sum_{j=1}^{L} m_j \to c_i > 0$ for some fixed value c_i as $m \to \infty$. Taking the expectation of (2), whilst noting that, under mild conditions, $\mathbb{E}[d(\boldsymbol{\theta}^{mle})] = \mathbb{E}[(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)] \to 0$ as $m \to \infty$, we get the delta-method approximation of bias,

$$b \approx \frac{1}{2} \mathbb{E}[d(\boldsymbol{\theta}^{mle})^T \, \mathrm{H}(\boldsymbol{\theta}^c) d(\boldsymbol{\theta}^{mle})] = b^{approx}.$$

After some matrix manipulation, this simplifies to

$$b^{approx} = \frac{1}{2} \operatorname{tr}(\Omega \, \mathrm{H}(\boldsymbol{\theta}^c)) \tag{3}$$

where tr() denotes the trace of a matrix and $\Omega = \mathrm{Var}(\boldsymbol{\theta}^{mle})$ denotes the variance-covariance matrix of the MLEs. For a proof of the asymptotic equivalence of b and b^{approx} as $m \to \infty$ see Appendix B.

As previously noted $\boldsymbol{\theta}^c$ is unknown; if it were known then there would be no error due to input modelling. In simulation studies it is also most often the case that the systems we simulate are complex, and thus no tractable form of our response of interest exists; we will therefore also treat the response function, $\eta(\cdot)$, as unknown. This means the delta approximation of bias, b^{approx}, cannot be evaluated directly; we therefore estimate it by

$$\widehat{b} = \frac{1}{2} \operatorname{tr}(\widehat{\Omega} \, \widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle})). \tag{4}$$

Evaluation of $\widehat{b}$ requires estimates of both the variance-covariance matrix of the input parameters and the Hessian matrix of second-order partial derivatives. In practice we estimate Ω using $\widehat{\Omega} = \mathrm{I}_0(\boldsymbol{\theta}^{mle})^{-1}/m$ the inverse Fisher information evaluated at $\boldsymbol{\theta}^{mle}$. From this point on, $\widehat{\Omega}$ will refer to this plug-in estimate for $\mathrm{Var}(\boldsymbol{\theta}^{mle})$. Notice that using $\widehat{\Omega}$ rather that Ω introduces additional error into $\widehat{b}$, but this error was insignificant in the experiment reported in Morgan et al. (2017).

In brief, Morgan et al. (2017) found that in controlled experiments with a truly quadratic $\eta(\cdot)$ and homogeneous variance, the relative error of $\hat{b}$ to b using $\hat{\Omega}$ was shown to be less than 1%.

Estimating the Hessian is more difficult. For this we choose a response surface modelling approach, quantifying the non-linearity of the response surface by investigating the behaviour of $\eta(\cdot)$ close to $\boldsymbol{\theta}^{mle}$, our estimate of $\boldsymbol{\theta}^c$, see Section 3.1.

Based on our estimate of the bias, we present a bias detection test with high power when $|b| \geq \gamma$. In Section 3.2 we illustrate the use of an experimental design for estimating the Hessian, and therefore the bias. We also present a novel way to construct this experimental design that allows a practitioner to control the power of the bias detection test.

3.1. Estimating the Hessian

To estimate the Hessian we make the further assumption that our response surface is locally quadratic; that is, if we are near enough to $\boldsymbol{\theta}^c$

$$\eta(\boldsymbol{\theta}) = \beta_0 + \boldsymbol{\theta}^T \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{B} \boldsymbol{\theta}, \tag{5}$$

where $\boldsymbol{\beta}$ is the vector of coefficients belonging to the linear terms, $\mathbf{B}$ is the $(k \times k)$ matrix of coefficients belonging to the interaction and quadratic terms and $\boldsymbol{\theta}$ is any vector of input parameter values near $\boldsymbol{\theta}^c$. Note that, if $\eta(\cdot)$ is twice continuously differentiable at $\boldsymbol{\theta}^c$, as assumed in (2), then this is approximately true using Taylor series. In Section 3.3 we suggest a test for lack-of-fit of the quadratic response surface; then in Section 4.1 we evaluate this assumption by considering responses with different functional forms. For now we will assume (5) holds.

By fitting model (5) we can estimate the Hessian matrix of second-order partial derivatives, allowing the evaluation of $\widehat{b}$. It is clear that taking the second-order partial derivatives of (5) with respect to $\boldsymbol{\theta}$ is equivalent to estimating $\mathbf{B}$. As $\boldsymbol{\theta}^c$ is unknown, we will use a central composite design (CCD), centred at $\boldsymbol{\theta}^{mle}$, to fit this model. The CCD is well known and has at least Resolution V, allowing the estimation of quadratic and interaction effects without confounding. Fig. 1 illustrates a CCD design in $k = 2$ dimensions; factorial (purple) and axial (yellow) design points are positioned relative to $\boldsymbol{\theta}^{mle}$, the central (red) design point.

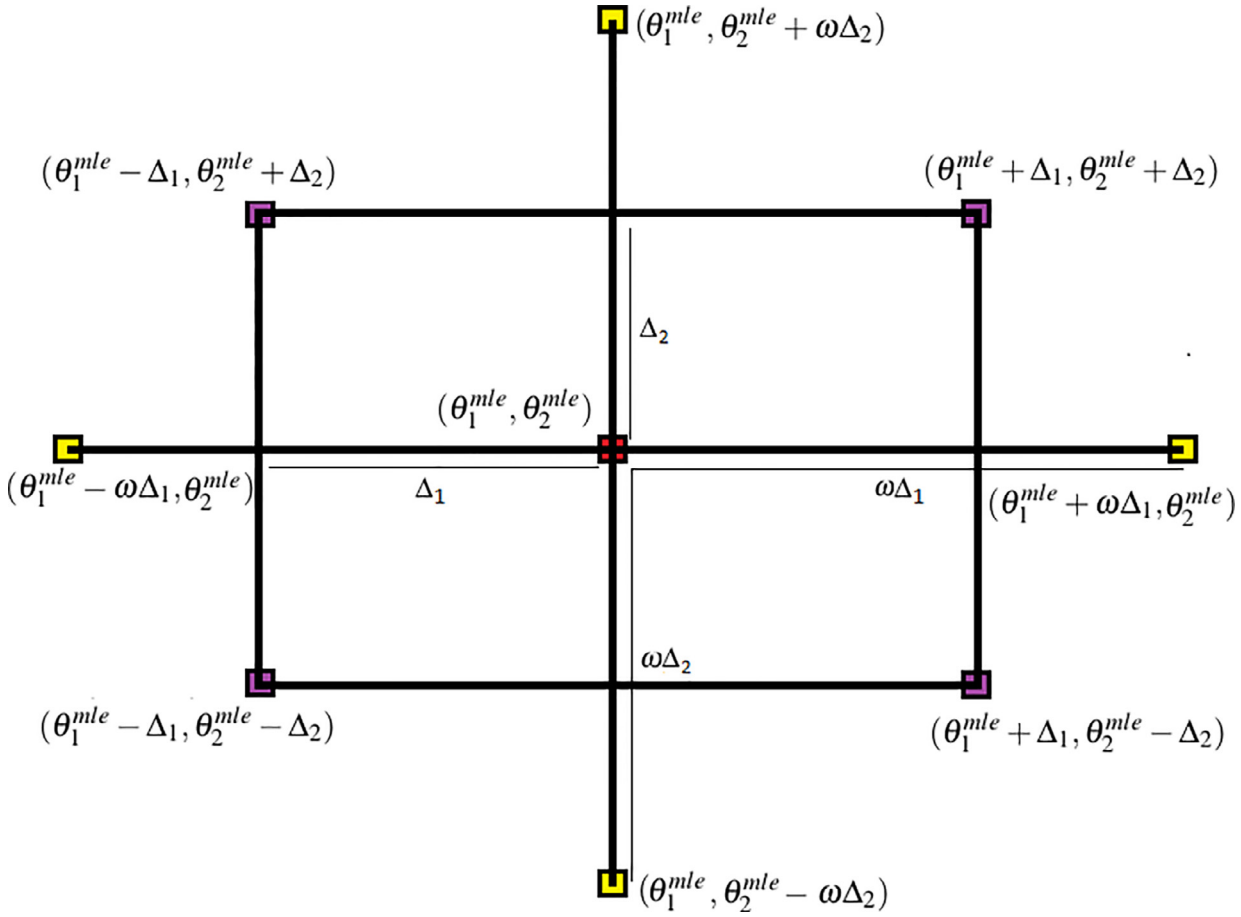


Fig. 1. A CCD design with dimension $k = 2$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To fit model (5), we complete r replications of the simulation model at each design point. Let n_F denote the number of factorial design points and n_A the number of axial design points. As suggested by Montgomery (2013), we will carry out more replications of the experiment at the centre point allowing more information collection at $\boldsymbol{\theta}^{mle}$, the point at which we wish to estimate the Hessian. We let this number be a multiple of r, which allows us to treat the multiple replications at centre point as multiple design points, $n_C > 1$. The total number of design points n is therefore $n = n_F + n_A + n_C = 2^k + 2k + n_C$, and depends on the number of input parameters, k. The total number of simulation replications is $n \times r$.

Clearly, the total number of design points, n, grows exponentially with the number of input parameters, k. For $k = 10$, the number of factorial design points is $n_F = 2^{10} = 1024$, even without considering the axial and centre points of the design. To reduce the size of the design, we therefore propose the use of fractional factorial designs, with the addition of axial and centre points. The key to this is to select a Resolution V, or higher, fractional factorial design to ensure no main effects or two-factor interactions are confounded (Montgomery, 2013).

Sanchez and Sanchez (2005) provide an efficient algorithm for generating Resolution V CCDs with a greatly reduced number of design points using discrete-valued Hadamard–Walsh functions to describe and generate the design. Their method focuses on specifying highly-fractionated Resolution V fractional factorial designs. After the fractional-factorial design has been generated the centre and axial points can then be added just as in the full CCD. When $k = 10$, Sanchez and Sanchez (2005) recommend $n_F = 128$ factorial design points, resulting in $n = 148 + n_C$ design points in total without specifying n_C. This is computationally much cheaper than the $n_F = 1024$ factorial design points, in total $n = 1044 + n_C$ points, needed in the full CCD experiment. In Section 4.2 we implement these reduced designs alongside the full-factorial CCDs in the NHS 111 setting for comparison.

In Fig. 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 ith direction, $i = 1, 2, \ldots, k$, and similarly let τ_i be the distance to the axial points. Experimental designs are often used to investigate the operational range of systems. It is therefore common to work with standardised variables, transforming the original quantitative factors to the values +1 and −1, representing the high and low levels of each factor at the edge of the operational space. We use experimental design quite differently. We are not interested in looking at the behaviour of $\eta(\cdot)$ over the entire range of each input variable. Instead, we are interested in assessing the Hessian of the response surface at the unknown $\boldsymbol{\theta}^c$. By using the standard deviation of the MLEs, $\sqrt{\text{Var}(\theta_i^{mle})}$ for $i = 1, 2, \ldots, k$, to scale the experimental design in each direction, we have a reasonable chance of covering $\boldsymbol{\theta}^c$ without having to spread our design points so wide that we risk violating the quadratic assumption over our design space. Note that, based on similar reasoning we might have chosen to use the variance-covariance matrix of the MLEs, $\text{Var}(\boldsymbol{\theta}^{mle})$, to scale the design. This would take into account dependencies among the input parameters, but would have introduced substantial additional complexity to the method. Given that we cannot prove that either method leads to the optimal design scaling we opt for the simpler option. That is, we set $\Delta_i = a\sqrt{\text{Var}(\theta_i^{mle})}$ and $\tau_i = \omega\Delta_i = a\omega\sqrt{\text{Var}(\theta_i^{mle})}$ 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.

At the ith design point we run r replications of the simulation returning the averaged output of the simulation, $\bar{Y}(\boldsymbol{\theta}_i)$ for $i = 1, 2, \ldots, n$. Given the averaged outputs we use least-squares regression to fit the response surface model and therefore evaluate the Hessian,

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

Given $\widehat{\Omega}$ and $\hat{\mathrm{H}}(\boldsymbol{\theta}^{mle})$, we can now estimate the bias, using $\widehat{b}$, as in Eq. (4).

We can also estimate $\mathrm{Var}(\widehat{b})$. Conditional on the value of $\widehat{\Omega}$, the plug-in estimate of $\mathrm{Var}(\boldsymbol{\theta}^{mle})$, $\mathrm{Var}(\widehat{b})$ is

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

This requires the calculation of $\mathrm{Var}(\widehat{\mathbf{B}})$, the variance-covariance matrix of regression coefficients belonging to the interaction and quadratic terms.

Given we estimated $\widehat{\mathbf{B}}$ by least-squares regression an estimator of $\mathrm{Var}(\widehat{\mathbf{B}})$ is easily obtained under the assumption of normally distributed residuals with homogeneous variance, using standard regression analysis. Note that the assumption of normally distributed residuals is reasonable here since the output at each design point is the average of a large number of replications r. We derived that $\mathrm{Var}(\widehat{\mathbf{B}})$ has special form

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

$$\mathrm{Cov}(\hat{\mathrm{B}}_{ii},\hat{\mathrm{B}}_{jj}) = \frac{\sigma^2 g}{ra^4\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}},$$

where, s, f and g are constants independent of the scaling factor a and $\widehat{\Omega}$. We exploit the common ra^4 scaling in Section 3.2 when it comes to manipulating the CCD width to control the power of our hypothesis test.

Application of our method will always follow a nominal experiment run at $\boldsymbol{\theta}^{mle}$; we therefore have a natural estimator of the simulation noise σ^2; we denote this by $\widehat{\sigma}^2$. In practice we use $\widehat{\sigma}^2$ as a plug-in estimator in the expressions for $\mathrm{Var}(\hat{\mathrm{B}}_{ii})$, $\mathrm{Var}(\hat{\mathrm{B}}_{ij})$ and $\mathrm{Cov}(\hat{\mathrm{B}}_{ii},\hat{\mathrm{B}}_{jj})$.

We derived that when using a CCD, $\mathrm{Cov}(\hat{\mathrm{B}}_{ij},\hat{\mathrm{B}}_{lm}) = 0$ when $i\neq j$ or $l\neq m$, therefore after some simplification our estimate of $\mathrm{Var}(\widehat{b})$ has the form

$$\widehat{\mathrm{Var}}(\widehat{b}) = \frac{\widehat{\sigma}^2}{ra^4}\left[sk + f\sum_{i=1}^{k}\sum_{j>i}^{k}\frac{\widehat{\Omega}_{ij}^2}{\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}} + gk(k-1)\right]. \tag{6}$$

At this point we have presented a method for estimating the bias of the simulation response caused by input modelling and have also provided a variance estimate associated with it. However, in some cases the bias will be small and therefore hard to accurately estimate. When the bias is small, we are not interested in getting a precise estimate of $\widehat{b}$. A bias detection test could therefore save us computational effort since we do not require as much precision to be able to reject a hypothesis as we would perhaps want to use $\widehat{b}$ as a point estimate of the error about our performance measure. Let γ denote the size of the smallest bias due to input modelling that would concern us. We will now present our key idea, a diagnostic test for detecting the bias with controlled power of rejecting the null when $|b| \geq \gamma$.

3.2. A bias detection test

We begin by considering the hypothesis test $\mathrm{H}_0 : b = 0$ vs. $\mathrm{H}_1 : b \neq 0$ with test statistic $\mathrm{T} = \widehat{b}/\sqrt{\widehat{\mathrm{Var}}(\widehat{b})}$. Let the size of the test be denoted by α_1 and the power by $1-\alpha_2$. We shall assume that

$$\frac{\hat{b} - b}{\sqrt{\widehat{\mathrm{Var}}(\hat{b})}} \sim \mathrm{N}(0,1) = \mathrm{Z}, \tag{7}$$

which is a reasonable approximation since $\widehat{b}$ is a linear combination of asymptotically normally distributed least-squares regression estimators, and we expect $\widehat{\mathrm{Var}}(\hat{b})$ to be a good estimate of $\mathrm{Var}(\widehat{b})$ since we have many observations. The key to this test is in controlling the power at a pre-specified level $1-\alpha_2$ so that, when the absolute bias is truly greater than or equal to γ, we have a high probability of rejecting the null hypothesis. We therefore require an experimental design where the following significance and power constraints hold given γ,

$$\mathrm{P}[\ |\mathrm{T}| > \mathrm{Z}_{1-\alpha_1/2} \mid b = 0\ ] = \alpha_1 \tag{8}$$

$$\mathrm{P}[\ |\mathrm{T}| > \mathrm{Z}_{1-\alpha_1/2} \mid\ |b|\ \geq \gamma\ ] \geq 1-\alpha_2. \tag{9}$$

Let the true IU variance of the response of interest be denoted by $\kappa = \mathrm{Var}(\eta(\boldsymbol{\theta}^{mle}))$. Using IU variance quantification techniques we can estimate κ by $\widehat{\kappa}$. We propose that, when the practitioner does not have an obvious value in mind for γ, $\widehat{\kappa}$ can be used to guide this choice. This is a natural suggestion as it looks at the bias within the context of the total MSE due to input modelling. If the bias is very small compared to $\widehat{\kappa}$ it may not be worth taking into account. On the other hand if the bias is large compared to $\widehat{\kappa}$ it would be important, and using $\widehat{\kappa}$ to guide our choice of γ will give us high power of rejecting the null.

We know that Eq. (8) is guaranteed by (7). Constraint (9) holds when

$$\sqrt{\widehat{\mathrm{Var}}(\hat{b})} \leq \frac{\gamma}{\mathrm{Z}_{1-\alpha_2} - \mathrm{Z}_{\alpha_1/2}}. \tag{10}$$

This says that the estimate of the variance of our bias estimator, $\widehat{\mathrm{Var}}(\hat{b})$, can be used to control the power of our test. From Eq. (6) it can be seen that, of the components that make up $\widehat{\mathrm{Var}}(\hat{b})$, only the width of the CCD, controlled via a, and the number of replications at each design point, controlled via r, can be influenced by the practitioner. In many simulation scenarios we are constrained by some fixed simulation budget. When this is the case, and we have a set total simulation budget $n \times r$ that we are willing to spend, we can set a, the scaling parameter of the experimental design, to be the smallest value such that

$$a \geq \left[\frac{\widehat{\sigma}^2 t^2}{r\gamma^2}\left(sk + f\sum_{i=1}^{k}\sum_{j>i}^{k}\frac{\widehat{\Omega}_{ij}^2}{\widehat{\Omega}_{ii}\widehat{\Omega}_{jj}} + gk(k-1)\right)\right]^{\frac{1}{4}}, \tag{11}$$

where $t = \mathrm{Z}_{1-\alpha_2} - \mathrm{Z}_{\alpha_1/2}$, the difference of the critical values given our size and power requirements. Alternatively, we may wish to choose a just large enough so that we can be confident that $\boldsymbol{\theta}^c$ has been covered within the CCD design space and set r appropriate

to it; recall that a was defined in units of the standard deviation of the MLEs. Notice that we can easily rewrite (11) to yield the number of replications as a function of a. Some caution is advised as $r = O(1/a^4)$, which means that a small decrease in the width of the design leads to a great increase in the number of replications required at each design point to estimate the change in the response surface in the smaller region.

Due to the limitations on how far we can spread our design before the quadratic assumption breaks down, we propose fixing an appropriately large r and letting (11) guide our choice of a. In Section 3.3 we describe a lack-of-fit test that can be used to test the quadratic assumption.

Given a and r that satisfy (11), we are able to set up the CCD to ensure that the power holds at the pre-set level, $1-\alpha_2$ within the hypothesis test. We can now carry out the bias detection test knowing that if the bias is truly greater than or equal to γ we have a high probability of rejecting the null hypothesis, H_0.

On completion of the test, even if we reject H_0, we cannot say anything about the size of the bias. We have sufficient evidence to suggest that the bias is non-zero at the α_1% level, and therefore is worth considering, but we cannot be sure that it is greater than or equal to our relevant value of bias, γ. At this point the practitioner may wish to collect further observations of the real system to reduce the error due to input modelling. Another option might be to spend further simulation effort on improving the precision of the estimate $\widehat{b}$ so it can be included in a summary of the total error of the response. Whichever choice is made we have presented a novel method for detecting the bias due to input modelling, a source of error that, before this contribution, had been virtually ignored.

An algorithm for the bias test is summarised below.

1. Preliminary Step. From the real-world input data estimate $\boldsymbol{\theta}^c$ and Ω by $\boldsymbol{\theta}^{mle}$ and $\widehat{\Omega}$. From the nominal experiment estimate σ^2 by $\widehat{\sigma}^2$. Set γ, a bias we wish to detect, α_1 the size, and $1-\alpha_2$ the power, of the test.
2. To ensure the power holds: initially let $a = 1$, noting that any positive value will suffice; create the $\left(n \times \left(1 + 2k + \frac{k(k-1)}{2}\right)\right)$ design matrix X, centred at $(0, 0, \ldots, 0)$ with $\Delta_i = a\sqrt{\text{Var}(\theta_i^{mle})}$ and $\tau_i = \omega\Delta_i$, for $i = 1, 2, \ldots, k$. Given X, evaluate s, f and g as follows
$$s = (X^{\mathrm{T}}X)^{-1}_{[\frac{(k+1)(k+2)}{2}, \frac{(k+1)(k+2)}{2}]}\Delta_k^4, \qquad f = (X^{\mathrm{T}}X)^{-1}_{[k+2,k+2]}\Delta_1^2\,\Delta_2^2,$$
$$g = (X^{\mathrm{T}}X)^{-1}_{[\frac{(k+1)(k+2)}{2}-1, \frac{(k+1)(k+2)}{2}]}\Delta_{k-1}^2\,\Delta_k^2$$
where the subscript $[i, j]$ denotes the element in the ith row and jth column of a matrix. Now use (11) to set a and r, to ensure the power holds.
3. Re-build the design matrix X, centred at $(\theta_1^{mle}, \theta_2^{mle}, \ldots, \theta_k^{mle})$, given a.
4. For each design point $i = 1, 2, \ldots, n$, run r replications of the simulation at $\boldsymbol{\theta}_i$, corresponding to row i of the design matrix; average over the r replications to find $\bar{Y}(\boldsymbol{\theta}_i)$.
5. Using the simulation output from each design point $\bar{Y}(\boldsymbol{\theta}_i)$, for $i = 1, 2, \ldots, n$, estimate the regression coefficients $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k, \widehat{\mathrm{B}}_{11}, \widehat{\mathrm{B}}_{12}, \ldots, \widehat{\mathrm{B}}_{(k-1)k}, \widehat{\mathrm{B}}_{kk})^T = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\bar{\boldsymbol{Y}}(\boldsymbol{\theta})$, giving $\widehat{\mathrm{B}}_{11}, \widehat{\mathrm{B}}_{12}, \ldots, \widehat{\mathrm{B}}_{(k-1)k}, \widehat{\mathrm{B}}_{kk}$.
6. Evaluate $\widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle})$; thus, estimate b and $\text{Var}(\widehat{b})$ by $\widehat{b}$ and $\widehat{\text{Var}}(\widehat{b})$.
7. Calculate the test statistic, $\mathrm{T} = \widehat{b}/\sqrt{\widehat{\text{Var}}(\widehat{b})}$. If $|\mathrm{T}| \geq Z_{1-\alpha_1/2}$ reject the null hypothesis.

3.3. Validating the bias test

Up to this point we have made the assumption that our response surface, $\eta(\cdot)$, is truly quadratic near $\boldsymbol{\theta}^c$. In reality we know that this does not hold in all cases. Take for example a single-server Markovian queue with capacity, C. For this system the expected number of customers in the system in steady state, $\eta(\cdot)$, can be expressed in closed form

$$\eta(\boldsymbol{\theta}) = \frac{\theta_1}{\theta_2 - \theta_1} - \frac{(C+1)\theta_1^{C+1}}{\theta_2^{C+1} - \theta_1^{C+1}},$$

where θ_1 is the arrival rate and θ_2 the service rate. This global (not local) response function is clearly not quadratic. Detection of the bias due to input modelling in this system was empirically explored by Morgan et al. (2017). They found that when the traffic intensity θ_1/θ_2 was close to 1, centring the CCD close to $\boldsymbol{\theta}^c$ was of great importance to ensure power held at $1-\alpha_2$ when γ was set to equal the true bias, b. This was particularly evident in models with high capacity, C, where $\eta(\cdot)$ was sensitive to changes in θ_1 and θ_2.

Although the expected response surface is unlikely to be truly quadratic, as long as the quadratic assumption holds locally within our CCD, we will get a good approximation of the non-linearity of the response surface at $\boldsymbol{\theta}^{mle}$. We therefore propose using a lack-of-fit test to check the quadratic assumption on the response. The "classical" lack-of-fit test, as described by Myers, Montgomery, and Anderson-Cook (1995), compares the error caused by lack-of-fit to the pure error estimated from replications made at the centre of the experiment design. This test assumes homogeneous variance across the design space; Kleijnen (1983) provides a lack-of-fit test based on cross validation if this assumption does not hold.

The "classical" lack-of-fit test comes with certain advantages. Firstly, no additional simulation effort is required to incorporate the lack-of-fit test within the bias detection method since we replicate the centre point in the experimental design, $n_C > 1$; this allows the calculation of the pure error. Also, we do not have to assume any functional form for our response surface; we could have compared the quadratic model to a cubic model for example but there is no guarantee that the cubic part of the model would be the problem in all cases.

Running the lack-of-fit test prior to our bias detection test enables us to examine the quadratic assumption. Of course, a hypothesis is just an assessment of evidence: accepting the null hypothesis does not prove that the approximation of a quadratic surface near $\boldsymbol{\theta}^{mle}$ is good enough to provide a trustworthy estimate of bias. However, rejecting the quadratic fit is a useful warning that the resulting bias estimate might not be trustworthy. By the nature of Taylor series approximation, a smaller-width CCD will tend to imply better conformance to a quadratic approximation. Therefore, one way to react to a significant lack of fit, as long as there is additional computer budget, is to increase r, the number of replications at each design point; this leads to a smaller value of a, the width scaling parameter of the design, whilst preserving the power of the bias test at $1-\alpha_2$ (see Section 3.2 and in particular Eq. (11)).

That said, repeated application of the lack-of-fit test with different sample sizes, the unknown effect of the experiment design used to fit the quadratic model, and the power of the lack-of-fit test all muddy the overall inference. Thus, while we recommend the lack-of-fit test its conclusions are at best advisory, and standard regression diagnostics applied to the quadratic model will also be helpful.

4. Empirical evaluation

In this section we evaluate the bias detection test presented in Section 3. In Section 4.1 we complete a controlled study considering four tractable response surfaces with different functional forms whilst controlling the number of input observations, m, and the number of simulation replications at each design point, r. We then

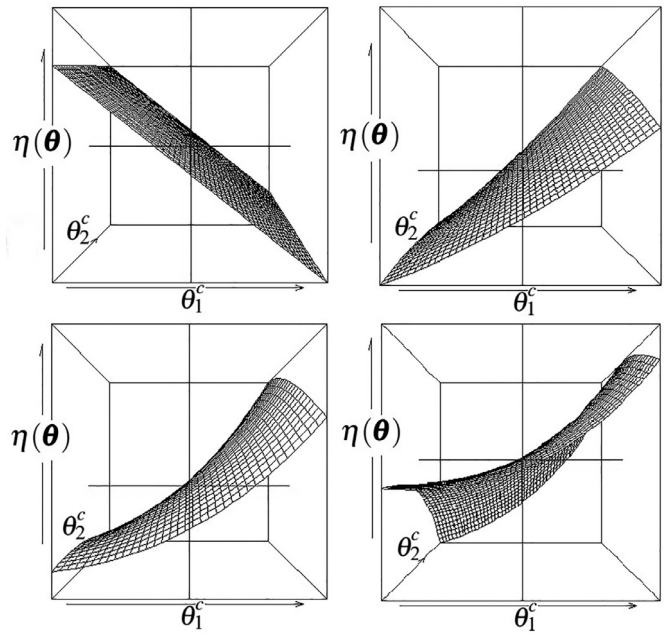


Fig. 2. The true response surfaces plotted over the CCD design space. Top left: linear, Eq. (12); top right: quadratic, Eq. (13); bottom left: cubic, Eq. (14); and bottom right: cubic, Eq. (15). The point(θ_1^c, θ_2^c) is marked in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

demonstrate the use of the bias detection test in the NHS 111 call centre setting in Section 4.2.

4.1. Monte Carlo evaluation of the method

Recall that the bias due to input modelling is caused when the error in the estimation of the input models that drive the simulation is passed through a non-linear response function. We therefore evaluate how well our bias detection test works when there is no bias due to input modelling i.e., the response is linear; when the response surface is truly quadratic; and finally when the underlying quadratic assumption does not hold.

We consider a stochastic simulation model with two unknown input parameters, $\boldsymbol{\theta}^c = \{\theta_1^c, \theta_2^c\} = \{3, 2\}$. These input parameters are the means of two independent exponentially distributed random variables, $W_1 \sim \text{Exp}(1/\theta_1^c)$, $W_2 \sim \text{Exp}(1/\theta_2^c)$.

Within this setting we consider the following functional forms for the response surface $\eta(\boldsymbol{\theta})$: linear, Eq. (12); quadratic, Eq. (13); and two cubic functions, Eqs. (14) and (15), as displayed in Fig. 2,

$$\eta(\boldsymbol{\theta}) = 3 - 10\theta_1 + 4\theta_2 \tag{12}$$

$$\eta(\boldsymbol{\theta}) = 3 - 10\theta_1 + 4\theta_2 + 8\theta_1\theta_2 + 2.5\theta_1^2 - 2.5\theta_2^2 \tag{13}$$

$$\eta(\boldsymbol{\theta}) = 3 - 10\theta_1 + 4\theta_2 + 8\theta_1\theta_2 + 2.5\theta_1^2 - 2.5\theta_2^2 + 0.4\theta_1^3 - 0.8\theta_2^3 \tag{14}$$

$$\eta(\boldsymbol{\theta}) = 3 - 10\theta_1 + 4\theta_2 + 8\theta_1\theta_2 + 2.5\theta_1^2 - 2.5\theta_2^2 + 0.8\theta_1^3 - 3\theta_2^3. \tag{15}$$

In this carefully constructed experiment the input parameters and the response functions are known. We also chose our input distributions so that the third moment of the MLE could be calculated exactly and we were therefore able to quantify, b, the bias due to input modelling from each function as well as the delta approximation of bias, b^{approx}; see Table 1. We set the size of the bias detection test to $\alpha_1 = 0.05$ and the power to $1 - \alpha_2 = 0.8$; the size for the lack-of-fit test is also 0.05.

Table 1
Bias test results varying the form of $\eta(\cdot)$, the amount of input data, m, and number of replications, r. Here $\hat{p}$ and LOF are the fraction out of $G = 1000$ macroreplications that the bias test and lack-of-fit test, respectively, rejected their null hypothesis, and $\bar{b}$ is the average bias estimate.

	m	r	b	b^{approx}	$\bar{b}$	$\hat{p}$	LOF
Linear (12)							
	10	50	0.00	0.00	-0.01	0.06	0.04
	100	50	0.00	0.00	0.00	0.05	0.05
	1000	50	0.00	0.00	0.00	0.04	0.06
	10	500	0.00	0.00	0.00	0.05	0.05
	100	500	0.00	0.00	0.00	0.05	0.05
	1000	500	0.00	0.00	0.00	0.04	0.05
Quadratic (13)							
	10	50	1.25	1.25	1.36	0.64	0.05
	100	50	0.13	0.13	0.13	0.71	0.06
	1000	50	0.01	0.01	0.01	0.80	0.05
	10	500	1.25	1.25	1.42	0.63	0.06
	100	500	0.13	0.13	0.13	0.72	0.05
	1000	500	0.01	0.01	0.01	0.80	0.06
Cubic 1 (14)							
	10	50	2.66	2.57	3.01	0.70	0.06
	100	50	0.26	0.26	0.26	0.65	0.06
	1000	50	0.03	0.03	0.03	0.75	0.05
	10	500	2.66	2.57	3.33	0.69	0.06
	100	500	0.23	0.26	0.27	0.70	0.06
	1000	500	0.03	0.03	0.03	0.78	0.06
Cubic 2 (15)							
	10	50	0.48	0.53	0.08	0.96	0.62
	100	50	0.05	0.05	0.05	0.92	0.22
	1000	50	0.01	0.01	0.01	0.74	0.09
	10	500	0.48	0.53	0.90	0.97	0.36
	100	500	0.05	0.05	0.06	0.92	0.10
	1000	500	0.01	0.01	0.01	0.78	0.07

To evaluate the bias detection test the value of the relevant bias γ is set equal to the delta approximation of bias b^{approx} in both the quadratic and cubic scenarios. In choosing $\gamma = b^{approx}$ we expect the power to hold at the pre-set value $1 - \alpha_2$. In the linear experiment $b = b^{approx} = 0$, so we use $\widehat{\kappa}$, the estimate of IU variance, found using the method of Cheng and Holland (1997), to guide the choice of γ where $\gamma = \sqrt{0.3\widehat{\kappa}}$.

Since the true bias, b, is known in these examples we set σ^2/r to be 5 times larger than b in the quadratic and cubic experiments, implying that there is still significant simulation noise in the evaluation of each design point. In all of the linear experiments σ^2 was set to 0.1. Given σ^2 and the response functions, we simulated by adding normally distributed noise, $N(0, \sigma^2)$, to Eqs. (12)–(15). From here on we assume the response functions are unknown and require estimation for the bias detection test.

We complete $G = 1000$ macro-replications of the bias detection test. To do this we collect m observations from each input distribution by generating observations, $\{w_{11}, w_{12}, \ldots, w_{1m}\}$ and $\{w_{21}, w_{22}, \ldots, w_{2m}\}$ from the true input distributions. This is our "real-world" data from which we estimate the input parameters using maximum likelihood. Given these estimates we run the nominal experiment, and in the linear case estimate the IU variance in the model. We then apply the bias detection test.

To quantify how well the bias detection test performs we estimate the power of the test by recording the empirical power, the proportion of times we reject the null hypotheses over $G = 1000$ macro-replications; we call this estimate $\hat{p}$. We then observe how close the empirical estimate $\hat{p}$ gets to the nominal power, $1 - \alpha_2 = 0.8$, for $\gamma = b^{approx}$, given the functional form of $\eta(\cdot)$, m and r. We also record the average of the estimates of the bias due to input modelling, $\hat{b}$, over the G replications, $\bar{b}$, for comparison with the true bias, b. The results are presented in Table 1.

In the linear system, Eq. (12), there is no bias. In Table 1 it can be seen that we reject the null hypothesis of no bias and the lack-of-fit test in approximately 5% of all the linear cases corresponding to the pre-set size of the tests, 0.05, as required.

In the quadratic system, Eq. (13), the delta approximation of bias is exact, so $b^{approx} = b$, and centering the CCD at $\boldsymbol{\theta}^{mle}$ rather the $\boldsymbol{\theta}^{c}$ does not matter since the response is globally quadratic. We would therefore expect the power hold at $1-\alpha_2$ plus or minus sampling error. In Table 1 we see this for $m = 1000$ and it is close for $m = 100$ where the error in $\widehat{p}$ is roughly ± 0.04. When $m = 10$ however, we see a lower power than expected and a discrepancy between $b = b^{approx}$ and $\bar{b}$. When the quantity of real-world input data is so exceptionally small, use of the plug-in estimate $\widehat{\Omega}$ without accounting for its variance is likely the cause.

Two cubic functions were also considered. When the response surface is cubic the locally quadratic assumption of our response surface not strictly correct, but it may be reasonable depending on the cubic function. Here b, the true bias due to input modelling, contains the third moment of the MLEs of the input distributions, $\mathbb{E}[(\boldsymbol{\theta}^{mle})^3]$; these can be calculated using the skewness of the MLEs: $\text{Skew}(\theta_i^{mle}) = 2/\sqrt{m}$, for $i = 1, 2$. The delta approximation of the bias due to input modelling, b^{approx}, is a second-order approximation and therefore does not take the higher moments into account. However, in results Table 1 it can be seen that as m increases $b^{approx} \to b$ since $2/\sqrt{m} \to 0$ as $m \to \infty$.

The first cubic function, Eq. (14), was selected such that the quadratic approximation was reasonable over the space covered by the CCD design. In Table 1 we see that, when the smallest values of m and r were used, the lack-of-fit test is passed approximately the same proportion of times as the quadratic function, and we see similar results to the quadratic experiment. As m and r increase we see the power get increasingly close to 0.8 and the delta approximation, b^{approx}, converges to b. Overall our method works well for this example.

The second cubic function, Eq. (15), was chosen so that the quadratic assumption was a poor approximation over the CCD space for the smallest values of m and r considered. When $m = 10$ and $r = 50$ the lack-of-fit test rejected the the quadratic model in approximately 60% of the $G = 1000$ macro replications; this was the best case, but overall this test was not very sensitive to the lack of fit. In Table 1 we see that the power of the bias test is often higher than our nominal value of 0.8 for small values of m and r even when the average estimated bias, $\bar{b}$, differs substantially from b and b^{approx}; this is good, but we should not expect it to be a general phenomenon. Increasing m or r has the effect of shrinking the width of the CCD making the quadratic assumption over our CCD space a better approximation.

This experiment shows the importance of the locally quadratic assumption over the CCD space. When the quadratic assumption does not hold our estimate of the bias, $\widehat{b}$, can be quite different from b when m is small. Using the lack-of-fit test to validate the quadratic assumption is therefore advised, but is not a panacea; recall this requires no additional simulation effort. Another problem is that, for small m, the distance between $\boldsymbol{\theta}^{mle}$ and $\boldsymbol{\theta}^{c}$ may be quite large, implying that we estimate the Hessian of the response surface at the wrong point which could impact both the estimate of the bias and the power of the test.

4.2. A realistic example - NHS 111 healthcare call centre

We 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 arrival counts over 96, 15-minute intervals, from an NHS 111 healthcare call centre in the UK. As previously described, the NHS 111 system was designed to remove some of the strain from other healthcare services, for example emergency departments, by advising callers on which service they should access. Of the 6 months of data we had we decided to consider Wednesdays only as UK public holidays mid-week are rare and therefore we would expect no outliers in the arrival rates.

After checking the Poisson assumptions were satisfied by the arrival data, this system was simulated as an $M(t)/G/S(t)$ queueing model with a non-stationary Poisson arrival process having a piecewise-constant rate. Based on data from the NHS 111 system we conducted two experiments with different levels of input data. Let s denote the number of days of observations of the arrival process. Figs. 3a and 3 b show the average rates over $s = 10$ and $s = 26$ days of arrival count data, respectively. In both scenarios change-point analysis for Poisson data, as discussed in Chen and Gupta (2011), was used to distinguish between intervals with significantly different arrival counts. This pre-processing technique was used because the IU variance in each small interval may be large, especially in intervals with low arrival rates where we would not expect to observe many arrivals. The change-point analysis reduced the arrival rate process to 7 and 8 intervals of varied length for the two scenarios; see the blue intervals in Figs. 3a and 3 b. Using the methods discussed by Morgan et al. (2016) we were then able to estimate the total IU variance, $\widehat{\kappa}$, of the expected waiting time of callers, $\mathbb{E}(\text{WTime})$, in both cases.

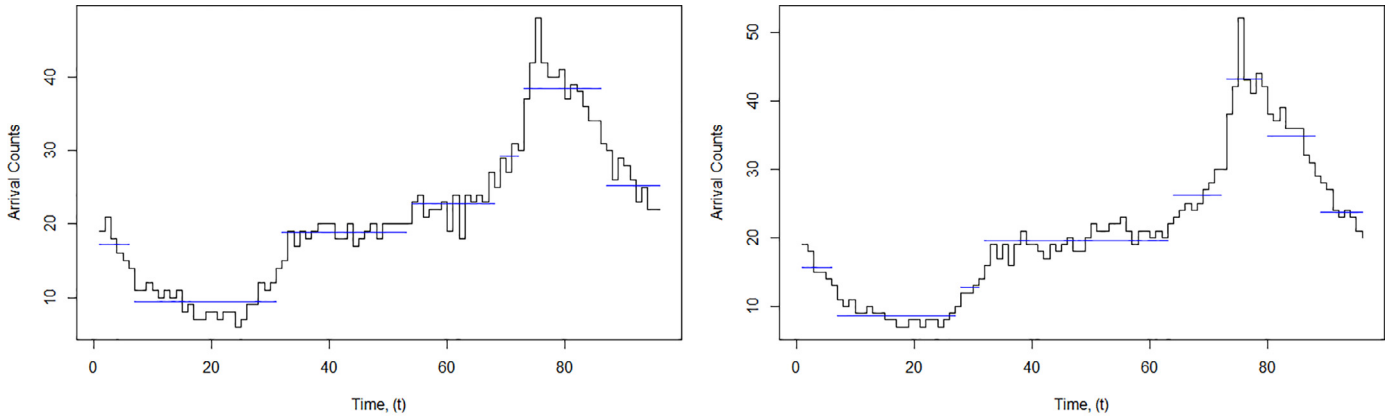
From two months of service-time data the mean service time was 8.00 minutes and the standard deviation was 4.33 minutes. A moment matching approach was used to fit a Gamma distribution with shape parameter $\phi_1 = 3.408$ and scale parameter $\phi_2 = 2.347$. Since we wanted to mimic having observed a service time for each arrival, we created a synthetic "observed" data set of service-time observations of size m, corresponding to the expected number of arrivals in each scenario, and treated this as the real-world service time data.

To generate a realistic scenario we used approximately proportional staffing to meet the NHS target level of service, $\text{P}(\text{WTime} > 1 \text{ minute}) < 0.05$. This corresponded to server utilisation of 62% in the model with $s = 10$ days of arrival data and 65% in the system with $s = 26$ days of arrival data. In the nominal experiment estimates of the expected waiting time of callers were found to be $\text{E}(\text{WTime}) = 0.0756$ minutes and $\mathbb{E}(\text{WTime}) = 0.0674$ minutes, respectively; this is our performance measure of interest.

For both systems we carry out the bias diagnostic test, as described in Section 3, and within this we run the lack-of-fit diagnostic test to investigate our quadratic approximation. An estimate $\widehat{\kappa}$ of IU variance is used to guide our choice of the relevant bias, γ. Note that, γ will therefore reduce with m, the amount of input data, because IU variance is also reduced. We want a high power of rejecting the null if the true bias is larger than $\gamma = \sqrt{\upsilon \times \widehat{\kappa}}$ where $0 < \upsilon < 1$. This gives us a threshold of the bias deemed to have an important effect on the MSE. Estimates of $\boldsymbol{\theta}^{c}$ and Ω were obtained from the input data, and σ^2 from the nominal experiment.

The desired power of the bias detection test was set equal to $1-\alpha_2 = 0.8$ and the size to $\alpha_1 = 0.05$; the size for the lack-of-fit test is also 0.05. For these experiments the relevant bias, γ, was set using $\upsilon = 0.3$, meaning we consider bias squared higher than 30% of the value of IU variance to be relevant.

For the two scenarios the number of input parameters driving the simulations are $k = 9$ and $k = 10$, respectively. This comes from the piecewise-constant arrival rate process having 7 or 8 distinct intervals, which are treated as independent input distributions; the final two parameters describe the service-time distribution. We conducted experiments employing both the full-factorial CCD and the reduced fraction CCD design proposed by Sanchez and Sanchez (2005). The latter design reduced the number of factorial points in both experiments to $n_F = 128$ from $n_F = 512$ and $n_F = 1024$, respectively. Note that in all experiments we repeat the centre point



(a) The arrival count function given $s = 10$ days of observations.

(b) The arrival count function given $s = 26$ days of observations.

Fig. 3. The average arrival counts over 96, 15 minutes, intervals given s days of arrival data. Intervals post pre-processing of the data using change-point analysis are shown in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The bias detection test in a NHS 111 system considering the expected waiting time of callers, $\mathbb{E}(\text{WTime})$, with $s = 10$ and $s = 26$ days of arrival data. Results for both the bias and the lack-of-fit tests are presented.

Design	Exp	s	m	n	r	γ	a	$\widehat{b}$	Bias	LOF
Full	1	10	20068	550	500	0.0035	0.577	0.0014	**Accept**	*Reject*
				550	1000	0.0035	0.485	0.0019	**Accept**	**Accept**
Frac	2	10	20068	166	500	0.0035	0.603	0.0013	**Accept**	*Reject*
				166	1000	0.0035	0.507	0.0005	**Accept**	**Accept**
Full	3	26	52711	1064	500	0.0024	0.699	0.015	*Reject*	*Reject*
				1064	1000	0.0024	0.588	0.011	*Reject*	*Reject*
Frac	4	26	52711	168	500	0.0024	0.737	0.005	*Reject*	**Accept**

$n_C = 20$ times. The results of the bias detection test are displayed in Table 2.

Before we analyse the results of our bias detection test note that in Table 2 for experiments 1, 2 and 3 the result of the lack-of-fit test in the initial experiment with $r = 500$ replications at each design point was to reject the quadratic model. For this reason we repeated these experiments, increasing the number of replications at each design point from $r = 500$ to $r = 1000$. This did not change the conclusion of the bias detection test, but did result in experiments 1 and 2 passing the lack-of-fit test. Thus, in these two experiments with $r = 1000$ we have no strong evidence that our quadratic approximation is inadequate. In experiment 3, even with $r = 1000$, the lack-of-fit test rejects the null, suggesting a more complicated model is required to describe the response surface. Note that, although we doubled the number of replications at each design point the scaling factor of the design, a, only decreased by a small amount. Acquiring a scaling factor small enough for the quadratic approximation to hold may take a much larger number of replications; recall that $r = O(1/a^4)$.

In experiments 1 and 3 we use the full-factorial CCD and in experiments 2 and 4 we use the reduced fractional CCD by Sanchez and Sanchez (2005). In Table 2 we see that the conclusion of the bias detection test given the full CCD agrees with the conclusion when the reduced fractional design is used for both levels of arrival data. The scalability of our method was an issue of concern to us. Here we see a great reduction in the number of design points, n, and thus computational effort, required to estimate the bias due to input modelling when using the reduced experimental design, yet we are still able to gain an estimate $\widehat{b}$ reasonably close to the estimate from the full CCD and make the same conclusion using the bias detection test.

In Table 2 we also see that, given a larger number of days of observations of the NHS 111 system γ, our relevant value of the bias, decreases from $\gamma = 0.0034$ to $\gamma = 0.0024$. This is because we used IU variance to guide our value of γ and the estimate of IU variance, $\widehat{\kappa}$, is smaller in the system with more days of input data. Our bias detection test is set up so that when $|b| \geq \gamma$ we have high power of detecting the bias. Since γ is higher in experiments 1 and 2 with $s = 10$ days of observations we require a larger departure from H_0 than we do in the experiments where $s = 26$ to have a high probability of rejecting the null. Further, given a large amount of input data the variability of the MLE's will be small. With our method this causes a smaller variance about the bias due to input modelling, $\text{Var}(\widehat{b})$, which in turn increases the power of our bias detection test.

Turning our attention to the conclusions of the bias detection tests in Table 2, we see that in experiments 1 and 2, with $s = 10$ days of arrival data, we accept the null hypothesis, so there is insufficient evidence to suggest $b \neq 0$ in these experiments. Since we set our threshold for relevant b^2 to 30% of the input uncertainty variance, and controlled the power to detect a bias larger than this size, our conclusion is more practically stated as the bias is making a small contribution to overall MSE due to input modelling.

In experiments 3 and 4, with $s = 26$ days of observations, we reject the null hypothesis; that is, we have sufficient evidence to suggest that $b \neq 0$. At this point we may wish to spend additional computational effort on estimating $\widehat{b}$, to get a more precise estimate of the bias due to input modelling about our performance measure estimate. Alternatively, at this point the practitioner may wish to reduce the bias to a level that does not concern them by collecting more input data and repeating the bias detection test.

We have now illustrated our bias detection test on a realistic example. This example had a non-stationary piecewise-constant rate Poisson arrival process that we pre-processed using change-point analysis. Note that the location of the change-points will have had an effect on the bias due to input modelling. Change-point analysis aids the choice of arrival intervals but does not guarantee an arrival function that represents the true arrival process perfectly propagating minimal error due to input modelling to our simulation output.

5. Conclusion

This paper presents a test with controlled power for detecting a bias of a relevant size caused by input distributions with parameters estimated from real-world data. Previously this form of error has been virtually ignored. The test is built on the assumption that close to $\boldsymbol{\theta}^c$ the true response can be approximated by a quadratic model. We fit the quadratic response surface using a CCD experimental design, which is constructed in a novel way allowing the practitioner to control the power of the bias detection test through the scaling of CCD width or the number of replications at each design point.

We explored and evaluated the bias detection test using a controlled experiment investigating the functional form of the response surface, the amount of input data and the number of replications completed at each design point. This experiment highlighted the importance of the validity of our quadratic assumption over the CCD space for our power to hold. We were also able to show that by increasing the number of replications of the experiment at each design point or the number of observations used to estimate our input models we achieved our target power. Also influential was the distance between the estimated input model parameters, $\boldsymbol{\theta}^{mle}$, and the true input model parameters, $\boldsymbol{\theta}^c$, which was seen to affect both the estimate of the power and the estimate of the bias. We also demonstrated the bias detection test in a realistic NHS 111 system example. This included the use of the IU variance to guide our choice of the relevant value of the bias.

From our exploration of quantifying and detecting the bias due to input modelling there still remain open questions that may be of interest. One of these is the study of other performance measures beyond the mean response. In this contribution our focus was on detecting the bias caused by input modelling in the expected value of a performance measure of interest; in future this could be extended to other measures such as the variance or the quantiles. Another question is how we might optimally set n_C the number of centre points in our model. Currently n_C is set in an ad hoc manner dependent on the number of factorial and axial points in the CCD. Also of interest is how we might optimally set r, the number of replications of the simulation at each design point. Recall that r controls a, the scaling factor for the width of the CCD.

We need r large enough to ensure our quadratic assumption holds sufficiently closely but do not wish to waste unnecessary simulation budget. In the experiments in this paper we chose r to be suitably large to satisfy our quadratic assumption.

In the NHS 111 example we used change-point analysis to form the arrival-process input model, which introduces its own error, but more generally input model misspecification is a source of model risk not captured here (e.g., if the arrival process is not actually Poisson). Similarly, we found that the lack-of-fit test was not as strong an indicator as one might like of approximation error. This could be due to the assumption of constant variances over the CCD, or the assumption of normally distributed simulation responses. This is an important problem for future study.

Note that our method can be used alongside current IU variance quantification techniques, allowing us to express the total error due to input modelling of our performance measures of interest. Current techniques allow IU variance quantification for simulation models with time-homogeneous distributions and piecewise-constant rate non-stationary Poisson processes. Estimation and detection of error due to input modelling in simulation with more complex arrival processes is something we leave for future work.

In conclusion, this paper offers the first method for estimation and detection of the bias due to input modelling. In doing so it allows a practitioner to consider the total error due to input modelling that may impact their performance measures of interest.

Acknowledgements

We gratefully acknowledge the support of the EPSRC funded EP/L015692/1 STOR-i Centre for Doctoral Training, NSF Grant CMMI-1634982 and GOALI sponsor Simio LLC. We also thank Bruce Ankenman for insightful discussion and suggestions. An earlier version of this paper was published in the *Proceedings of the 2017 Winter Simulation Conference* as Morgan et al. (2017).

Appendix A. Variability of the Jackknife estimator of bias

The jackknife method is an alternative to the delta-method that can be used for bias estimation. Usually when quantifying the bias we refer to the bias of a statistic of interest, for example a population parameter given a sample of data; in this case let us denote the jackknife estimator of bias $\widehat{b}_{JK}$. In stochastic simulation the statistic we would like to examine is the expected value of the simulation response, $\eta(\cdot)$. However, we can only observe this in the presence of simulation noise. In this appendix we investigate the effect of simulation noise on the variability of the jackknife estimator of the bias.

As a simplification, consider a stochastic simulation model with a single input parameter, θ^c from a single input model. Let θ^{mle} be the maximum likelihood estimator (MLE) of θ^c based on m observations of the input distribution and $\theta^{mle}_{(i)}$ is the "reduced information" MLE based on all but the ith observation. The jackknife estimate of the bias is

$$\widehat{b}_{JK} = (m-1)\left[\frac{1}{m}\sum_{i=1}^{m}\eta(\theta^{mle}_{(i)}) - \eta(\theta^{mle})\right].$$

Since we cannot evaluate $\eta(\cdot)$ directly, the natural extension to simulation output is,

$$\widehat{b}_{JK+noise} = (m-1)\left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{r}\sum_{j=1}^{r}Y_j(\theta^{mle}_{(i)}) - \frac{1}{r}\sum_{k=1}^{r}Y_k(\theta^{mle})\right] \quad (16)$$

which requires r independent replications of the simulation at each reduced information MLE, $\theta^{mle}_{(i)}$, and independent of this r replications of the simulation at the MLE, θ^{mle}. Within (16) the output of a replication of the simulation can be decomposed into the expected simulation response plus simulation noise

$$\begin{aligned}
&\widehat{b}_{JK+noise}\\
&= (m-1)\left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{r}\sum_{j=1}^{r}(\eta(\theta^{mle}_{(i)}) + \epsilon_{ij}) - \frac{1}{r}\sum_{k=1}^{r}(\eta(\theta^{mle}) + \epsilon_k)\right]\\
&= (m-1)\left[\frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m}\eta(\theta^{mle}_{(i)}) - \frac{1}{r}\sum_{k=1}^{r}\eta(\theta^{mle})\right.\\
&\quad \left. + \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m}\epsilon_{ij} - \frac{1}{r}\sum_{k=1}^{r}\epsilon_k\right], \qquad (17)
\end{aligned}$$

where $\epsilon_{ij} \sim$ i.i.d$(0,\ \sigma^2)$ and $\epsilon_k \sim$ i.i.d$(0,\ \sigma^2)$. Here (17) can be thought of as breaking $\widehat{b}_{JK+noise}$ into $\widehat{b}_{JK}$, the jackknife estimator of the bias without simulation noise, and $\widehat{b}_{noise}$, the additional variability in the estimator of the bias caused by simulation noise.

The key to this investigation is the variance of $\widehat{b}_{noise}$

$$\begin{aligned}\mathrm{Var}(\widehat{b}_{noise}) &= \mathrm{Var}\left((m-1)\left[\frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m}\epsilon_{ij} - \frac{1}{r}\sum_{k=1}^{r}\epsilon_k\right]\right)\\ &= (m-1)^2\left[\frac{1}{r^2m^2}\sum_{j=1}^{r}\sum_{i=1}^{m}\mathrm{Var}(\epsilon_{ij}) + \frac{1}{r^2}\sum_{k=1}^{r}\mathrm{Var}(\epsilon_k)\right]\\ &= (m-1)^2\left[\frac{\sigma^2}{rm} + \frac{\sigma^2}{r}\right]\\ &= (m-1)^2\frac{(m+1)\sigma^2}{rm} \end{aligned} \tag{18}$$

which is, for large m, approximately equal to $m^2\sigma^2/r$. This says that, in the presence of simulation noise, the number of simulation replications per reduced information MLE, r, required to maintain a constant level of error as m grows is $r = O(m^2)$, and the total number of simulation replications to compute the jackknife with constant error grows as $O(m^3)$. For stochastic simulation models with more than one input parameter this effect would be even greater. Thus, it is clear that significant simulation effort may be required; otherwise the jackknife estimate of this bias could be obscured by the presence of simulation noise.

Appendix B. Asymptotics of b and b^{approx}

Using Taylor series we show that, under certain assumptions, as $m \to \infty$ the bias, $b = \mathbb{E}[\eta(\boldsymbol{\theta}^{mle})] - \eta(\boldsymbol{\theta}^c)$, coincides with the delta approximation of the bias, b^{approx}.

Assumption B.1. Let the expected simulation response, $\eta : \mathbb{R}^k \to \mathbb{R}$,

1. Be three times continuously differentiable in a closed ball G centred at $\boldsymbol{\theta}^c$.
2. Have bounded above, third-order partial derivatives such that in the closed ball G, there exists some $M > 0$, for all $\boldsymbol{s} \in G$, $\frac{\partial^3\eta(\boldsymbol{s})}{\partial\theta_i\partial\theta_j\partial\theta_p} \le M$ for $i, j, p = 1, 2, \ldots, k$.

Assumption B.2. Let the simulation be driven by L independent, parametric input distributions, with $k \ge L$ input parameters. Assume we have m observations for each of the L distributions. Now let $\boldsymbol{\theta}^{mle} \in \mathbb{R}^k$ be the vector of MLEs given the m observations of each input distribution. We assume the MLEs satisfy standard conditions implying that

1. The MLEs converge in mean, $\mathbb{E}(\theta_i^{mle} - \theta_i^c) \to 0$ as $m \to \infty$ for $i = 1, 2, \ldots, k$.
2. The MLEs are asymptotically normal, $\sqrt{m}(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c) \xrightarrow{\mathrm{D}} \mathrm{MVN}_k(\mathbf{0},\ \mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}) = \boldsymbol{Z}$.
3. For some $\varepsilon > 0$, $|\theta_i^{mle} - \theta_i^c|^{3+\varepsilon}$ are uniformly integrable for all $m \in \mathbb{N}$, and $i = 1, 2, \ldots, k$.

Theorem B.1. *Let Assumptions B.1 and B.2 hold. Then as $m \to \infty$ the scaled bias, mb, and the scaled delta approximation, mb^{approx}, both converge to*

$$\frac{1}{2}\mathrm{tr}(\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}\mathrm{H}(\boldsymbol{\theta}^c)).$$

Proof. Convergence of the MLEs implies that for m large enough we will have $\boldsymbol{\theta}^{mle} \in G$. Therefore, under Assumption B.1.1, the expected simulation response at $\boldsymbol{\theta}^{mle} \in G$ can be expanded via a Taylor series as

$$\begin{aligned}\eta(\boldsymbol{\theta}^{mle}) &= \eta(\boldsymbol{\theta}^c) + \nabla\eta(\boldsymbol{\theta}^c)^T(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)\\ &\quad + \frac{1}{2}(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)^T\mathrm{H}(\boldsymbol{\theta}^c)(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c) + \Upsilon_3(\boldsymbol{\theta}^{mle}), \end{aligned} \tag{19}$$

where $\Upsilon_3(\boldsymbol{\theta}^{mle})$ is the remainder, made up of higher-order terms of the Taylor series. For $k \ge 3$ there exists $\boldsymbol{\rho} \in G$ such that

$$\begin{aligned}\Upsilon_3(\boldsymbol{\theta}^{mle}) &= \frac{1}{6}\sum_{i=1}^{k}(\theta_i^{mle} - \theta_i^c)^3\frac{\partial^3\eta(\boldsymbol{\rho})}{\partial\theta_i{}^3}\\ &\quad + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,j\ne i}^{k}(\theta_i^{mle} - \theta_i^c)^2(\theta_j^{mle} - \theta_j^c)\frac{\partial^3\eta(\boldsymbol{\rho})}{\partial\theta_i{}^2\partial\theta_j}\\ &\quad + \frac{1}{6}\sum_{i=1}^{k}\sum_{j=1,j\ne i}^{k}\sum_{p=1,p\ne i,\ j}^{k}(\theta_i^{mle} - \theta_i^c)(\theta_j^{mle}\\ &\quad - \theta_j^c)(\theta_p^{mle} - \theta_p^c)\frac{\partial^3\eta(\boldsymbol{\rho})}{\partial\theta_i\partial\theta_j\partial\theta_p}. \end{aligned}$$

By taking the expectation of (19) we may write bias due to input modelling as

$$\begin{aligned}b &= \mathbb{E}[\eta(\boldsymbol{\theta}^{mle})] - \eta(\boldsymbol{\theta}^c) = \nabla\eta(\boldsymbol{\theta}^c)^T\mathbb{E}(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)\\ &\quad + \frac{1}{2}\mathbb{E}\left[(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)^T\mathrm{H}(\boldsymbol{\theta}^c)(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)\right] + \mathbb{E}[\Upsilon_3(\boldsymbol{\theta}^{mle})]. \end{aligned}$$

Note that, the delta approximation of bias only takes into account the second-order term in this expansion

$$b^{approx} = \frac{1}{2}\mathbb{E}\left[(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)^T\mathrm{H}(\boldsymbol{\theta}^c)(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c)\right] = \frac{1}{2}\mathrm{tr}\left(\Omega\mathrm{H}(\boldsymbol{\theta}^c)\right)$$

where $\Omega = \mathrm{Var}(\boldsymbol{\theta}^{mle})$, and under Assumption B.2.2, $\lim_{m\to\infty} m\Omega = \mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}$ the inverse Fisher information matrix. We can therefore write $b = b^{approx} + c(\boldsymbol{\theta}^{mle})$; that is, the bias due to input modelling is equal to the delta approximation of bias, b^{approx}, plus a function $c(\cdot)$ containing the expectation of the additional terms of the Taylor expansion evaluated at $\boldsymbol{\theta}^{mle}$. Clearly $mb^{approx} \to \mathrm{tr}(I_0(\boldsymbol{\theta}^c)^{-1}\mathrm{H}(\boldsymbol{\theta}^c))/2$; we will show that $mc(\boldsymbol{\theta}^{mle}) \to 0$.

Consider the expectation of the first order term of the Taylor series expansion. By Assumption B.2.1, $\mathbb{E}(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c) \to 0$ as $m \to \infty$ and therefore $\nabla\eta(\boldsymbol{\theta}^c)\mathbb{E}(\boldsymbol{\theta}^{mle} - \boldsymbol{\theta}^c) \to 0$ as $m \to \infty$.

Next consider the expectation of the remainder term, $\mathbb{E}[\Upsilon_3(\boldsymbol{\theta}^{mle})]$. Under Assumption B.1.2 the third-order partial derivatives are bounded above at $\boldsymbol{\rho} \in G$ by $M > 0$ for $i, j, p = 1, 2, \ldots, k$. Thus by linearity of expectation we have,

$$\begin{aligned}\mathbb{E}\left[\Upsilon_3(\boldsymbol{\theta}^{mle})\right] &\le \frac{1}{6}\sum_{i=1}^{k}\mathbb{E}[(\theta_i^{mle} - \theta_i^c)^3]M\\ &\quad + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,j\ne i}^{k}\mathbb{E}[(\theta_i^{mle} - \theta_i^c)^2(\theta_j^{mle} - \theta_j^c)]M\\ &\quad + \frac{1}{6}\sum_{i=1}^{k}\sum_{j=1,j\ne i}^{k}\sum_{p=1,p\ne i,\ j}^{k}\mathbb{E}[(\theta_i^{mle}\\ &\quad - \theta_i^c)(\theta_j^{mle} - \theta_j^c)(\theta_p^{mle} - \theta_p^c)]M. \end{aligned} \tag{20}$$

We will now show that $m \times (20)$ converges to 0 as $m \to \infty$ and thus, by sandwich rule, the scaled expectation of the remainder, $m\mathbb{E}\left[\Upsilon_3(\boldsymbol{\theta}^{mle})\right]$, converges to 0. Here the behaviour of the RHS of (20) depends on the behaviour of $\mathbb{E}[(\theta_i^{mle} - \theta_i^c)(\theta_j^{mle} - \theta_j^c)(\theta_p^{mle} - \theta_p^c)]$ for $i, j, p = 1, 2, \ldots, k$. Taking the modulus of this expectation and applying Holder's inequality, (Hardy, Littlewood, & Pólya, 1952), followed by the arithmetic mean - geometric mean inequality (Abramowitz & Stegun, 1964), we have

$$
\begin{aligned}
&|\mathbb{E}[(\theta_i^{mle}-\theta_i^c)(\theta_j^{mle}-\theta_j^c)(\theta_p^{mle}-\theta_p^c)]|\\
&\le \mathbb{E}\Big[|(\theta_i^{mle}-\theta_i^c)||(\theta_j^{mle}-\theta_j^c)||(\theta_p^{mle}-\theta_p^c)|\Big]\\
&= \mathbb{E}\Big[\sqrt[3]{|(\theta_i^{mle}-\theta_i^c)|^3|(\theta_j^{mle}-\theta_j^c)|^3|(\theta_p^{mle}-\theta_p^c)|^3}\,\Big]\\
&\le \frac{1}{3}\mathbb{E}[|(\theta_i^{mle}-\theta_i^c)|^3]+\frac{1}{3}\mathbb{E}[|(\theta_j^{mle}-\theta_j^c)|^3]\\
&+\frac{1}{3}\mathbb{E}[|(\theta_p^{mle}-\theta_p^c)|^3]. \qquad (21)
\end{aligned}
$$

By Assumption B.2.2 and B.2.3, $\sqrt{m}\,\mathbb{E}[|(\boldsymbol{\theta}^{mle}-\boldsymbol{\theta}^c)|^3]\to\mathbb{E}[|\boldsymbol{Z}|^3]$; that is, the third absolute moment of the MLE converges to the third absolute moment of the multivariate normally distributed random variable $\boldsymbol{Z}$ (Osius, 1989). Thus,

$$
m^{\frac{3}{2}}\mathbb{E}[|(\theta_i^{mle}-\theta_i^c)|^3]\to\frac{1}{\sqrt{\pi}}\left(2\,\mathrm{I}_0(\boldsymbol{\theta}^c)_{ii}^{-1}\right)^{\frac{3}{2}},
$$

as $m\to\infty$ for $i=1,2,\ldots,k$, (Winkelbauer, 2012). Here $\mathrm{I}_0(\boldsymbol{\theta}^c)_{ii}^{-1}$ is the ith diagonal element of the Fisher information matrix of the joint distribution of the k input parameters. This says that as $m\to\infty$, $m\mathbb{E}[|(\theta_i^{mle}-\theta_i^c)|^3]\to 0$ for $i=1,2,\ldots,k$ and therefore $m\times(21)$ converges to 0 as well.

By applying the sandwich rule we have $m|\mathbb{E}[(\theta_i^{mle}-\theta_i^c)(\theta_j^{mle}-\theta_j^c)(\theta_p^{mle}-\theta_p^c)]|\to 0$ as $m\to\infty$ for $i,j,p=1,2,\ldots,k$. Thus, $m\mathbb{E}[(\theta_i^{mle}-\theta_i^c)(\theta_j^{mle}-\theta_j^c)(\theta_p^{mle}-\theta_p^c)]\to 0$ as $m\to\infty$ for $i,j,p=1,2,\ldots,k$. Therefore $m\times(20)$ converges to 0 and thus the scaled remainder $m\mathbb{E}[\Upsilon_3(\boldsymbol{\theta}^{mle})]\to 0$ as $m\to\infty$. All components of $mc(\boldsymbol{\theta}^{mle})$ converge to 0 as $m\to\infty$ as required. □

Appendix C. Asymptotics of $\widehat{b}$

Our delta approximation of the bias is $b^{approx}=\frac{1}{2}\,\mathrm{tr}(\Omega\mathrm{H}(\boldsymbol{\theta}^c))$, where $\mathrm{H}(\boldsymbol{\theta}^c)$ is the Hessian matrix of the second-order partial derivatives of $\eta(\cdot)$ evaluated at $\boldsymbol{\theta}^c$ and $\Omega=\mathrm{Var}(\boldsymbol{\theta}^{mle})$, the variance-covariance matrix of the MLEs. Due to the unknowns in b^{approx} we estimate it by $\widehat{b}=\frac{1}{2}\mathrm{tr}(\widehat{\Omega}\widehat{\mathrm{H}}(\boldsymbol{\theta}^{mle}))$. We now show that, under certain assumptions, $m\widehat{b}$ converges to $mb^{approx}=\frac{1}{2}\mathrm{tr}(\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}\mathrm{H}(\boldsymbol{\theta}^c))$.

Assumption C.1. The expected simulation response, $\eta:\mathbb{R}^k\to\mathbb{R}$, is quadratic; i.e.,

$$
\eta(\boldsymbol{\theta})=\beta_0+\boldsymbol{\theta}^T\boldsymbol{\beta}+\boldsymbol{\theta}^T\mathbf{B}\boldsymbol{\theta}. \qquad (22)
$$

Assumption C.2. Except for the point at which it is centered, the CCD is fixed and sufficient to support Model (22) such that $\widehat{\mathrm{B}}_{ij}\in\mathbb{R}$, the least squares estimator of B_{ij} is a consistent estimator for $i,j=1,2,\ldots,k$. That is, $\widehat{\mathrm{B}}_{ij}\xrightarrow{P}\mathrm{B}_{ij}$ as $r\to\infty$ for $i,j=1,2,\ldots,k$.

Assumption C.3. Let the simulation be driven by L independent parametric input distributions, with $k\ge L$ input parameters. Assume we have m observations from each of the L distributions. Now let $\boldsymbol{\theta}^{mle}\in\mathbb{R}^k$ be the vector of MLEs given the m observations of each input distribution. We assume that

1. The MLEs are consistent, $\theta_i^{mle}\xrightarrow{P}\theta_i^c$ as $m\to\infty$ for $i=1,2,\ldots,k$.
2. The scaled variance of the MLEs $m\Omega$ tends to the inverse Fisher information at $\boldsymbol{\theta}^c$, $\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}$, as $m\to\infty$, $m\Omega\to\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}$ as $m\to\infty$.
3. The inverse Fisher information, $\mathrm{I}_0(\cdot)^{-1}$, is continuous.

Theorem C.1. *Let Assumptions C.1, C.2 and C.3 hold. Then the scaled estimate of the delta approximation of bias, $m\widehat{b}$, converges to the scaled delta approximation of bias; that is, as $m,r\to\infty$*

$$
m\widehat{b}\xrightarrow{P}\frac{1}{2}\mathrm{tr}(\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}\mathrm{H}(\boldsymbol{\theta}^c)).
$$

Proof. First consider the Hessian. Under Assumption C.1 the expected simulation response is globally quadratic; therefore the Hessian does not depend on where we evaluate it since

$$
\mathrm{H}(\boldsymbol{\theta})=\begin{pmatrix}2B_{11} & B_{12} & \cdots & B_{1k}\\ B_{21} & 2B_{22} & & \\ \vdots & & \ddots & \\ B_{k1} & & & 2B_{kk}\end{pmatrix}.
$$

Thus $\widehat{b}=\frac{1}{2}\mathrm{tr}(\widehat{\Omega}\mathrm{H}(\boldsymbol{\theta}^{mle}))$ and this proof is equivalent to showing that $m\widehat{\Omega}\mathrm{H}(\boldsymbol{\theta}^{mle})\xrightarrow{P}\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}\mathrm{H}(\boldsymbol{\theta}^c)$.

Further, the least-squares estimators of the second-order terms are unchanged by shifting the center point of the design. Thus, under Assumption C.2, by completing r replications of the simulation at each of the design points of the CCD we gain the consistent estimators of the second-order partial derivatives, $\widehat{\mathrm{B}}_{ij}\xrightarrow{P}\mathrm{B}_{ij}$ for $i,j=1,2,\ldots,k$, such that $\mathrm{H}(\boldsymbol{\theta})\xrightarrow{P}\mathrm{H}(\boldsymbol{\theta}^c)$ as $r\to\infty$ for any $\boldsymbol{\theta}$. Therefore, $\mathrm{H}(\boldsymbol{\theta}^{mle})\xrightarrow{P}\mathrm{H}(\boldsymbol{\theta}^c)$ as $r\to\infty$.

Now consider $\widehat{\Omega}=\widehat{\mathrm{Var}}(\boldsymbol{\theta}^{mle})$. In practice we use the plug in estimator $\widehat{\Omega}=\mathrm{I}_0(\boldsymbol{\theta}^{mle})^{-1}/m$. Under Assumption C.3.1 and C.3.3, using continuous mapping theorem, $\mathrm{I}_0(\boldsymbol{\theta}^{mle})^{-1}\xrightarrow{P}\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}$ as $m\to\infty$ thus $m\widehat{\Omega}\xrightarrow{P}\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}$ as $m\to\infty$

Finally, by applying Slutsky's theorem we have $m\widehat{\Omega}\mathrm{H}(\boldsymbol{\theta}^{mle})\xrightarrow{P}\mathrm{I}_0(\boldsymbol{\theta}^c)^{-1}\mathrm{H}(\boldsymbol{\theta}^c)$ as $m,r\to\infty$ as required. □

Remark 1. The results of Theorem B.1 and Theorem C.1 can be extended to the case where $m_1\ne m_2\ne\cdots\ne m_L$ provided that $m_i/\sum_{j=1}^L m_j\to c_i>0$, for some fixed values c_i.

References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*: 55. Courier Corporation.

Barton, R. R. (2012). Tutorial: Input uncertainty in outout analysis. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, & A. Uhrmacher (Eds.), *Proceedings of the 2012 winter simulation conference* (pp. 1–12). Piscataway, New Jersey: IEEE.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis and model building*. New York: Wiley.

Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media.

Cheng, R. C., & Holland, W. (1997). Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation, 57*(1–4), 219–241.

Dean, A., & Voss, D. (1999). *Response surface methodology*. New York: Springer-Verlag.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*: 38. Siam.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1952). *Inequalities*. Cambridge university press.

Kleijnen, J. P. (1983). Cross-validation using the t statistic. *European Journal of Operational Research, 13*(2), 133–141.

Montgomery, D. C. (2013). *Design and analysis of experiments*. John Wiley & Sons.

Morgan, L. E., Titman, A. C., Worthington, D. J., & Nelson, B. L. (2016). Input uncertainty quantification for simulation models with piecewise-constant non-stationary poisson arrival processes. In *Proceedings of the 2016 winter simulation conference* (pp. 370–381). IEEE Press.

Morgan, L. E., Titman, A. C., Worthington, D. J., & Nelson, B. L. (2017). Detecting bias due to input modelling in simulation models. In *Proceedings of the 2017 winter simulation conference* (pp. 1974–1985). IEEE Press.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (1995). Response surface methodology: Process and product optimization using designed experiments (Wiley series in probability and statistics). *Applied Probability and Statistics, 1*, 43–47.

Nelson, B. (2013). *Foundations and methods of stochastic simulation: a first course*. Springer Science & Business Media.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician, 46*(1), 27–29.

Osius, G. (1989). *Some results on convergence of moments and convergence in distributions with applications in statistics*. Universität Bremen.

Sanchez, S. M., & Sanchez, P. J. (2005). Very large fractional factorial and central composite designs. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 15*(4), 362–377.

Song, E., Nelson, B. L., & Pegden, C. D. (2014). Advanced tutorial: Input uncertainty quantification. In A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, & J. A. Miller (Eds.), *Proceedings of the 2014 winter simulation conference* (pp. 162–176). Piscataway, New Jersey: IEEE Press.

Winkelbauer, A. (2012). Moments and absolute moments of the normal distribution. Institute of Telecommunications, Vienna University of Technology https://arxiv.org/pdf/1209.4340.pdf.

Withers, C. S., & Nadarajah, S. (2014). Bias reduction: The delta method versus the jackknife and the bootstrap. *Pakistan Journal of Statistics, 30*(1), 143–151.