Sequential experimental designs for nonlinear regression metamodels in simulation

M. Isabel Reis dos Santos [a,*], Pedro M. Reis dos Santos [b]

[a] Department of Mathematics, Center for Management Studies, Technical University of Lisbon (IST), Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[b] Department of Computer Science and Engineering, Technical University of Lisbon (IST), Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ABSTRACT

The construction of a nonlinear regression metamodel for simulation requires experimental designs that better explore the nonlinearities of the system. The proposed sequential procedure focuses on simulation scenarios in sub-regions where the input–output behavior is more interesting. It takes into account, not only the inputs, but also the output (response) value of the metamodel. The resulting experimental design ensures that the scaled response values are evenly spread over a scaled response surface. Although the focus is on nonlinear regression metamodels, the method may be applied to other types of metamodels. This design procedure is illustrated for one-dimensional and two-dimensional inputs, and the results are compared with space filling designs. The use of bootstrapping for metamodel validation stresses the importance of validation in a metamodeling context.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Computer simulation models are one of the most widely used tools to perform sensitivity analysis of a system response. However, the interpretation of large amounts of data produced by the simulation study can become an intimidating task. The representation of the simulation input–output relationships through a simple mathematical function exposes in a straightforward manner the simulation's behavior. Such a representation of the simulation model, or *metamodel* [5], requires fewer computer resources than the computer simulation model itself. Metamodels may be very useful in answering 'what if' questions and can be used for verifying and validating and optimizing the simulation model as well. In the simulation literature, metamodels are also called response surfaces, emulators, etc. A metamodel approximates the simulation system. Due to their simple construction and use, general linear metamodels have been frequently used by simulation analysts and some researchers [8,23]. However, most real-life systems, such as problems involving queuing systems [40], exhibit nonlinear behavior. The use of linear metamodels (e.g. the intensively studied low-order polynomials) are unable to provide an acceptable global fit to curves of arbitrary shape. To overcome this limitation, some researchers have proposed a number of *nonlinear* metamodels that provide a better and more realistic global approximation of the simulation's input–output relationship; for example, nonlinear regression metamodels [32,34], Kriging metamodels [24], and neural nets [3]. Some nonlinear functions may fit the data well and with fewer and more meaningful parameters than linear approximations, thus making the interpretation of the system and its analysis more intuitive. The construction of a metamodel requires the selection of a design, and the determination of the concomitant metamodel type, the estimation of the metamodel parameters, and the validation of the resulting metamodel.

* Corresponding author. Tel.: +351 936 166 769.
  E-mail addresses: maria.santos@tagus.ist.utl.pt (M. Isabel Reis dos Santos), pedro.reis@tagus.ist.utl.pt (P.M. Reis dos Santos).

The selection of an efficient set of inputs, or experimental design, at which to run the simulation and compute the output can be an important challenge. Analysts need experimental designs capable of efficiently exploring an intricate simulation model characterized by a complex response surface, possibly containing some nonlinearities. A good experimental design should effectively explore an experimental region of a multi-dimensional input space through a small number of design points. The remaining stages of the metamodel's construction rely heavily on a good selection of the inputs used as training data to predict the output. Some sequential experimental design strategies have been proposed for finding inputs that optimize (instead of answering what-if questions) the output of a simulation model [16].

Several specialized methods for building metamodels have been proposed, including input selection based on Latin hypercube experimental designs [27]. Such one-shot designs, where all design points are selected prior to the execution of the simulation experiments, offer little or no escape for exploring the nonlinearities of the metamodel's response. Indeed, if the model is seriously in doubt, the misspecification of the metamodel itself can lead to completely inappropriate forms of design. Linear regression [11,25] and Kriging [30,42] methods have been proposed and compared [2] with mixed results. Other comparisons based on the relative performances of the metamodeling designs [28,30,39] have been presented in order to aid practitioners decide which designs to use.

The dependency on the true values of the underlying unknown model's parameters is one of the main difficulties associated with optimal experimental designs for nonlinear models [13,31]. These optimal experimental designs require strong *a priori* assumptions on the metamodel's type and on the nature of the response, for example, white noise [22]. As a result, this type of designs may be completely inappropriate when the input–output relation is in doubt.

Frequently, simulation experiments are implemented sequentially, except possibly when parallel computers are used [22]. Therefore, new inputs can be selected based on the information gathered in previous stages of the simulation experiment. A careful selection of inputs may require fewer runs than fixed-sample (one-shot) designs, leading to more efficient experimental designs. This characteristic is most critical when we are dealing with complex systems that require expensive simulations. Such designs are particularly useful when the experimenter may be more interested in obtaining a good prediction over a localized region. Recently, in a Kriging metamodeling context, both Kleijnen and van Beers [24] and Sasena et al. [36] developed sequential designs; see also Alam et al. [1], Brantley and Chen [7], Hendrickx and Dhaene [14] and Keys and Rees [16].

In this article, however, we focus on metamodel techniques for the design of experiments, to be analyzed through a possibly nonlinear regression metamodel. If the metamodel response is nonlinear, then the usual measures of performance of a design depend on the parameters being estimated. A sequential approach is then naturally suggested: one should choose design points so as to maximize a measure of performance evaluated at the estimates obtained from observations made at previous design points.

Santos and Santos [35] proposed a sequential design to improve the overall accuracy of a nonlinear regression metamodel with a single input. The procedure adds new points to the design where there are large gaps in the output. It is noted that this technique needs to be conjugated with initial evenly spaced inputs in order to assess regions where the response is regular. The approach improves the accuracy of the resulting metamodel compared with the results of an evenly spaced design. In this paper, we propose an algorithm to address multiple inputs. This algorithm is based on a wire-frame representation of the response surface using a mesh of triangles. The construction of the wire-frame borrows ideas from 3D computer graphics algorithms, where there is a need to expose the details of 3D objects. The values of each factor are scaled to the interval [0; 1] to account for significantly different ranges of each input; the response is also scaled to the same interval in order to normalize distance calculations. In our method the design points are selected to ensure an approximate uniform response behavior. As a result, it concentrates design points in sub-regions where the input–output behavior is of more interest. Subsequent design points are chosen to minimize the response surface gaps. These gaps correspond to triangles with large areas. The results are compared with an evenly spaced experimental design. Our comparisons are motivated by the fact that relatively little attention has been paid to the interaction between the choice of an experimental design and the resulting nonlinear metamodel accuracy.

This paper is organized as follows. In Section 2, the construction procedure for the general nonlinear multi-dimensional input metamodel is presented, excluding the experimental design component; the construction procedure includes the establishing of asymptotic results for the least squares estimators, the use of a nonparametric test for checking the variance heterogeneity in simulation experiments, and the use of bootstrapping for metamodel validation. Section 3 is dedicated to the proposed experimental design for multi-dimensional input, as well as the unidimensional simplification. In Section 4 two application examples are presented to illustrate unidimensional and two-dimensional input experimental designs. Section 5 summarizes the results and suggests directions for future work.

2. Metamodel construction

The metamodeling construction process should follow a scientific approach if meaningful conclusions are to be drawn from the experimental data [10]. First the simulation practitioner should state the objective of the study (prediction, optimization, . . .), choose the controllable input variables (factors), select the response variable (output) and specify the region of interest within a previously determined operating region of the system. To represent the simulation model, an approximating statistical model is postulated for the data generating process. Given a response, Y, and a vector of independent factors $\mathbf{X}$ influencing Y, the relationship between $\mathbf{X}$ and Y is represented by a mathematical transfer function f characterized by a

parameter vector θ. In order to estimate the model's unknown parameters, an experimental design must be performed. This experimental design must be chosen according to a specific criterion, usually the minimization of the mean squared error or the determinant of the variance of the least squares estimator $\hat{\theta}$ of θ. To perform the experiments and collect the data, the analyst must select an appropriate design class, the levels of the factors and the number of replications for each design point, r_i. Finally, an adequacy test must be executed to determine if the resulting metamodel is a valid surrogate of the original simulation model.

2.1. Nonlinear metamodel choice

To represent the simulation model, the following nonlinear regression model in the unknown parameters $\theta = [\theta_1 \ldots \theta_m]^{\mathrm{T}}$ is proposed as a metamodel

$$Y_{ij} = f(\mathbf{X}_{i\cdot}, \theta) + \epsilon_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, r_i, \tag{1}$$

where ϵ_{ij} represents the inaccuracy of the metamodel plus intrinsic noise of the stochastic simulation model and f is an unknown function simpler than the simulation model. We assume that $\mathrm{E}[\epsilon_{ij}] = 0$ and $\mathrm{Var}[\epsilon_{ij}] = \sigma_i^2$, with $\sigma_i > 0$.

The factor settings for the ith simulation configuration, named scenario, are specified by the design point $\mathbf{X}_{i\cdot}$, where the set of such points constitutes the experimental design. For each scenario, r_i independent replications of the simulation model are carried out so the simulation yields r_i observations Y_{ij} ($j = 1, \ldots, r_i$). The variance of Y_{ij} is estimated by

$$\hat{\sigma}_i^2 = \sum_{j=1}^{r_i} (Y_{ij} - \overline{Y}_{i\cdot})^2 / (r_i - 1), \quad i = 1, \ldots, n, \tag{2}$$

where $\overline{Y}_{i\cdot} = \sum_{j=1}^{r_i} Y_{ij}/r_i$, $i = 1, \ldots, n$.

Good metamodeling practice means that the analyst should strive to find the simplest metamodel that captures the essential characteristics of the system. Nonlinear metamodels require the selection of a transfer curve that can be parametrized to mimic the input–output relationship of the simulation model. Whenever some *a priori* insight about the system's behavior is known, a specific curve can be assumed. Otherwise, an initial set of design points must be simulated and the respective values plotted, $(\mathbf{X}_{i\cdot}, Y_{ij})$, i.e. the scatter diagram. A visual check enables the selection of a curve (or curve candidates) that might provide a good fit for the simulation input–output relationship, selected from known functional relationships; see [37, pp. 329, 338, 340].

2.2. Estimation of regression parameters

In metamodel estimation, statistical techniques are used to estimate the unknown parameters of the hypothesized metamodel. If the error variances vary with the experimental points, this estimation may use the weighted nonlinear least squares method. The members of a statistical population of responses Y_{ij} dependent on the input $\mathbf{X}_{i\cdot}$, correspond with all possible pseudo-random number seeds. The classical assumption is that this population is Gaussian: $\epsilon_{ij} \sim \mathrm{NID}(0, \sigma_i^2)$. Under the hypothesis that the ϵ_{ij} are independent and normally distributed, and some further regularity conditions, it is shown that the nonlinear weighted least squares estimator of θ is asymptotically normally distributed as $N \to \infty$ (see the following Proposition).

Proposition 1. *Given appropriate regularity conditions [43], which ensures the asymptotic normality of the weighted least squares estimator in the nonlinear case, and for large N, the weighted least squares estimator of θ, $\hat{\theta}$, in* (1) *satisfies, approximately:*

$$\hat{\theta} \approx \theta + [\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{F}]^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}^{-1}[\overline{\mathbf{Y}} - \mathbf{f}], \tag{3}$$

$$\hat{\theta} \sim \mathrm{N}_m\left(\theta, [\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{F}]^{-1}\right), \tag{4}$$

where $\mathbf{f} = \mathbf{f}(\theta) = [f(\mathbf{X}_{1\cdot}, \theta) \cdots f(\mathbf{X}_{n\cdot}, \theta)]^{\mathrm{T}}$, $\mathbf{F} = \mathbf{F}(\theta)$ *is the Jacobian matrix of* $\mathbf{f}$, *evaluated at* θ; $\mathbf{\Sigma}$ *is the diagonal matrix* $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2/r_1, \ldots, \sigma_n^2/r_n)$.

The verification of these results is shown in the Appendix and are an extension of Santos and Nova [34], with different number of replications at each design point. In order to simplify the notation, we omit that $\mathbf{f}$ and $\mathbf{F}$ are evaluated at θ. In general, we do not know $\mathbf{\Sigma}$, and so it must be replaced by $\hat{\mathbf{\Sigma}} = \mathrm{diag}(\hat{\sigma}_1^2/r_1, \ldots, \hat{\sigma}_n^2/r_n)$; as a result, we have the estimated weighted nonlinear least squares estimator.

In our sequential experimental design, we may consider additional replications at experimental points with larger variances. The selection of the number of replications per experimental point may improve the robustness of the ordinary tests [17, p. 167]. We apply the least squares method to the problem

$$\overline{Y}_{i\cdot} = \mathbf{f}(\mathbf{X}_{i\cdot}, \theta) + \bar{\epsilon}_{i\cdot}, \tag{5}$$

with $\bar{\epsilon}_{i\cdot} \sim \mathrm{N}(0, \sigma_i^2/r_i)$, so $\mathrm{Var}[\overline{Y}_{i\cdot}] = \sigma_i^2/r_i$ tends to be more constant; in the verification of (3) in the Appendix it is shown that the problems (1) and (5) are equivalent. Whenever possible, we realize the hypothesis of, approximately, constant variances:

$$\frac{\sigma_1^2}{r_1} \approx \cdots \approx \frac{\sigma_n^2}{r_n} \approx \sigma^2.$$

Proposition 2. *Let $\sigma_i^2 \approx r_i \sigma^2$ in (1). Then, given appropriate regularity conditions [43] and for large N, the ordinary least squares estimator of θ, $\hat{\theta}$, in (1) satisfies, approximately:*

$$\hat{\theta} \approx \theta + [\mathbf{F}^T\mathbf{F}]^{-1}\mathbf{F}^T[\overline{\mathbf{Y}} - \mathbf{f}], \tag{6}$$

$$\hat{\theta} \sim N_m(\theta, \sigma^2[\mathbf{F}^T\mathbf{F}]^{-1}), \tag{7}$$

$$(N-m)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{N-m}^2 \quad \text{with } \hat{\sigma}^2 = \frac{1}{N-m}\sum_{i=1}^n[\overline{Y}_{i.} - f(\mathbf{X}_{i.}, \hat{\theta})]^2, \tag{8}$$

where **f** *and* **F** *represent the same as in Proposition 1.*

The verification of (6)–(8) is shown in the Appendix, where (6) and (7) are an extension of Santos and Nova [33], with different number of replications at each design point.

The variance heterogeneity may be checked using a distribution-free test based on ranks described in Conover [9]; if populations are assumed to be normal (therefore, the assumptions of the usual parametric test are met), the loss of efficiency applying this method is surprisingly small [9, p. 213]. The application of this test to problem (1) consists of assessing if $\sigma_i^2/r_i = \text{Var}[Y_{ij}/\sqrt{r_i}]$ can be accepted as homogeneous. First, converting each $Y_{ij}/\sqrt{r_i}$ to its absolute deviation from the mean using

$$U_{ij} = \frac{|Y_{ij} - \overline{Y}_{i.}|}{\sqrt{r_i}}, \quad i = 1, \ldots, n, \ j = 1, \ldots, r_i;$$

using $\overline{Y}_{i.}$ as an approximation of the mean in experimental point i, the test is still approximately valid. Then, assign the ranks 1 to N to the combined absolute differences U's in the usual way, that is, rank the combined sample of U's from the smallest to the largest, assigning average ranks when several U_{ij} values are equal to each other. Let T_{ij} be the rank and the average ranks thus assigned. The test statistic is

$$\chi = \frac{1}{D^2}\left(\sum_{i=1}^n \frac{1}{r_i}\sum_{j=1}^{r_i} T_{ij}^2 - N\overline{T}_{..}^2\right),$$

where

$$\overline{T}_{..}^2 = \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^{r_i} T_{ij}^2$$

is the average of all the squared ranks, and

$$D^2 = \frac{1}{N-1}\left(\sum_{i=1}^n\sum_{j=1}^{r_i} T_{ij}^4 - N\overline{T}_{..}^2\right).$$

Finally, the hypothesis of equal variances is rejected if $\chi > \chi^2(n-1; 1-\alpha)$, where $\chi^2(n-1; 1-\alpha)$ is the $(1-\alpha)100$ percentile of the chi-square distribution with $(n-1)$ degrees of freedom. If all U's values are different from each other, then D^2 and $\overline{T}_{..}^2$ simplify to $D^2 = N(N+1)(2N+1)(8N+11)/180$ and $\overline{T}_{..}^2 = (N+1)(2N+1)/6$, respectively.

2.3. Assessing accuracy and adequacy

Once the metamodel parameters have been estimated, it is necessary to investigate if the metamodel is valid, that is, if the metamodel is a representation of the simulation model up to an acceptable degree of accuracy. Furthermore, when comparing metamodels constructed from different experimental designs, it is necessary to evaluate how accurate the resulting representations are. Several possible measures and techniques can be used in validation [23].

For validating the metamodel, we apply the bootstrap technique, based on quantiles, described in [12, Chapter 13]. For each experimental point i, resample with replacement the r_i output values, resulting in the bootstrap data $(\mathbf{X}_{i.}, Y_{i1}^*, \ldots, Y_{ir_i}^*)$ of the original data. Obtain the nonlinear least squares estimator $\hat{\theta}^*$ based on the bootstrap data. Repeat this bootstrapping B times. So, the method will produce B vector values $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

Based on these B vector values, we follow Kleijnen and Deflandre [20] for computing the corresponding coefficients of determination $R_1^{2*}, \ldots, R_B^{2*}$ obtaining a bootstrapped one-sided confidence interval for R^{2*}: $[R_{(B\alpha)}^{2*}; 1]$, where $R_{(B\alpha)}^{2*}$ is the $(1-\alpha)$ bootstrapped quantile of the empirical distribution function of R^{2*}. The metamodel is not rejected if $R^2 \in [R_{(B\alpha)}^{2*}; 1]$, where R^2 is calculated with the original data. Like in Kleijnen and Deflandre [20], R^2 is calculated as a function of $\overline{Y}_{i.}$ instead of the individual Y_{ij}:

$$R^2 = 1 - \frac{\sum_{i=1}^n(\hat{Y}_i - \overline{Y}_{i.})^2 r_i/\hat{\sigma}_i^2}{\sum_{i=1}^n(\overline{Y}_{i.} - \overline{Y}_{..})^2 r_i/\hat{\sigma}_i^2},$$

with $\hat{\sigma}_i^2$ given by (2) and $\overline{Y}_{..} = \sum_{i=1}^n \overline{Y}_{i.}/n$.

3. Experimental design

The experimental region of an experimental design corresponds to the values of the inputs for which the metamodel is useful. The experimental design must choose a set of *n* pilot design points, i.e., combinations of input variables and parameters of the simulation model. These points must be different from each other and must belong to a pre-defined region to explore. The points are chosen to efficiently investigate the relationship between the design factors and the response. Assuming that the interesting features of the metamodel can occur in any part of the experimental region, the design should be based on a selection of evenly distributed points throughout the region.

A uniform design allocates experimental points which are uniformly scattered over the region of interest. A good space-filling design is one in which the design points are scattered throughout the experimental region with minimal unsampled regions; that is, the "empty" regions are relatively small. This means that the design points are not concentrated in clusters or solely at corner points of the region, as can happen with two-level factorial designs. Space-filling designs provide coverage of the entire experimental region, providing a broad exploration of the model and a valuable overview of what the response surface might look like. Space-filling designs are desirable for data analysis methods such as residual plots in regression diagnostics and nonparametric surface fitting [44]. Space-filling designs also provide the best way of exploring surfaces where we do not expect to have smooth metamodels [19].

Whenever some parts of the experimental region exhibit more interesting characteristics, the design should include more points to explore these parts. These hybrid variations of evenly distributed and space-filling designs may provide a better metamodel fit since more detailed information on the relation between the inputs and the response becomes available. In some special cases, where the metamodel belongs to a previously known input–output class, it is possible to achieve an optimal design based on a specific criteria formulation [31]. Anyhow, an experimental design should be easily and quickly generated, if it is to be used in practice.

3.1. Sequential designs

The experimental design should effectively explore an experimental region of a multi-dimensional input space through a small number of design points. Analysts need experimental designs capable of efficiently searching an intricate simulation model characterized by a complex response surface, possibly containing some nonlinearities. To be general, the experimental design should require minimal *a priori* assumptions on the system's response. Experiments run sequentially save experimental cost, since they can be stopped whenever a criterion is met. Several sequential experimental design strategies have been proposed for finding inputs that optimize the output of a computer code [31].

The idea is to start with a small pilot sample within the design space in order to obtain initial information about the entire response surface and fit a metamodel. Based on this metamodel, a set of additional points – the so called infill samples – are selected to be evaluated by the full simulation. Each additional stage takes account of all previous information to obtain an experimental design consistent with the pre-defined experimental goal. The sample set is then updated, and the model refit. A quantitative criterion is implemented for the purpose of deriving the experimental design in each stage. The process of choosing new points continues until the improvement expected from sampling additional points has become sufficiently small [6].

A sequential experimental design should be conducted as follows:

- Postulate an approximating metamodel for the data generating process. Nonlinear functions are good candidates since they are more flexible and explicitly account for nonlinearities in the input–output function and complex interactions.
- Design an initial experiment and collect the required data through simulation. The initial design should provide a good coverage of the entire response surface; Jones et al. [15] recommends at least 10 observations per input variable.
- Estimate the metamodel parameters and calculate the response predictor.
- Compute the termination rule and plot the fitted metamodel. The sequential design is complete whenever a pre-defined accuracy is met or a maximum number of design points is reached. The degree of accuracy of the metamodel is judged by computing a statistical measure:
  - If the accuracy is sufficient, the process is concluded. Then, if the goal is optimization, predict the global optimizer.
  - Otherwise, choose a sub-region for the next experiment and collect more data. Sensitivity analysis, based on output variation, is used as mechanism for quantifying the importance of input variables. Output variation provides guidance for supplying reasonable input variable values to the next stage of the sequential design.

3.2. Experimental design selection

A careful choice of an experimental design may better expose the relation between the inputs and the response. For instance, a bad experimental design may suggest a linear relation between the inputs and the response while the response is in fact nonlinear. The detection of such nonlinearities is important and can be achieved with a larger number of input values or, better still, a careful selection of input values. Also, designs should be easy to construct if they are to be used in practice [22].

Many popular classical designs used in response surface methodology, such as central composite designs and factorial designs, do not spread points out evenly over the experimental region, denoted by Ω. Instead they tend to place points at

the boundary of Ω: corners and midpoints. The most basic experimental design is a full factorial design. The number of design points dictated by a full factorial design is the product of the number of levels for each factor [41]. Most common are 2^k designs for evaluating main effects and interactions for k factors each at two levels. Often two level fractional factorial designs are used to screen factors to identify those with greatest effects [39]. A Central Composite Design is a two level factorial design, augmented by n_0 central points and two star points positioned at $\pm \alpha$ for each factor. For two factors, setting $\alpha = 1$ locates the star points at one coded unit from the center [39].

More demanding experimental designs include the mesh and Latin Hypercube. In a mesh the design points are placed at a grid formed by dividing each axis into intervals of equal length. Latin Hypercube designs are generated by Latin Hypercube sampling, where to produce a design of n points each axis is divided into n equally spaced intervals producing n^d cells of equal size [31], where d is the number of factors. Each cell is given a number from 1 to n, unique in each dimension, or Latin square, i.e., each number appears exactly once in each row and each column of this mesh of cells. Each of the possible designs corresponds to the selection of a middle point from each cell, based on the $\Pi = (\Pi_{jk})$ $n \times d$ matrix having columns which are different randomly selected permutations of $\{1, \ldots, n\}$; the jth row of the Π identifies the cell with $\mathbf{X}_{j\cdot} = [X_{j1} \ldots X_{jd}]$ as the middle point.

3.3. Selecting the design points

An experimental design, $\mathscr{X} = \{\mathbf{X}_{1\cdot}, \mathbf{X}_{2\cdot}, \ldots, \mathbf{X}_{n\cdot}\}$, is a set of multi-dimensional input training sites over an experimental region, $R \equiv [a_1; b_1] \times [a_2; b_2] \times \cdots \times [a_d; b_d] \subset \mathbb{R}^d$, to be evaluated and producing a single-valued output Y. Then a criterion for judging the predictor $\hat{Y} = f(\mathbf{X}, \hat{\theta})$ is used to measure the quality of the prediction in an overall sense. The goal, or global objective, is the determination of $\hat{\theta}$ up to a level set by some pre-defined target accuracy value: $\mathscr{L}(\xi) = \{\mathscr{X} \subset R : M(f(\mathbf{X}, \hat{\theta})) \leqslant \xi\}$. The problem is to obtain an experimental design $\mathscr{X} \in \mathscr{L}(\xi)$ where the accuracy value of the metamodel, $M(f(\mathbf{X}, \hat{\theta}))$, is determined using the statistical measure $\mathrm{MSE}(\hat{\theta}) = \mathrm{SSE}(\hat{\theta})/(n-m)$, where $\mathrm{SSE}(\hat{\theta}) = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_i)^2 r_i/\hat{\sigma}_i^2$; in our case $M(f(\mathbf{X}, \hat{\theta})) = \mathrm{MSE}(\hat{\theta})$. To determine an experimental design $\mathscr{X}$ where $\mathrm{MSE}(\hat{\theta})$ achieves the target value ξ, new data points must be sequentially added.

Consider an initial set of design points $\mathscr{X}^{(0)} = \{\mathbf{X}_{1\cdot}, \ldots, \mathbf{X}_{n_0\cdot}\}$ from the experimental region $R \subset \mathbb{R}^d$, denoted in matrix notation by

$$\mathbf{X}^{(0)} = \begin{bmatrix} \mathbf{X}_{1\cdot} \\ \vdots \\ \mathbf{X}_{n_0\cdot} \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & & \vdots \\ X_{n_0,1} & \cdots & X_{n_0,d} \end{bmatrix}.$$

These design points can be chosen from well known experimental designs that acceptably screen the experimental region; for example, the initial set in Hendrickx and Dhaene [14] is just a full-factorial design. Jones et al. [15]'s rule of thumb is to take at least 10 experimental points per dimension ($n_0 \approx 10d$). Then new points are added until the stop criteria is met.

Suppose that the response and each input variable have been scaled to have domain $[0; 1]$, i.e., consider the scaled points

$$Z_i = \left(\frac{X_{i1} - a_1}{b_1 - a_1}, \ldots, \frac{X_{id} - a_d}{b_d - a_d}, \frac{\overline{Y}_{i\cdot} - a_{d+1}}{b_{d+1} - a_{d+1}} \right) = (z_{i,1}, \ldots, z_{i,d}, z_{i,d+1}), \tag{9}$$

where $\overline{Y}_{i\cdot} \in [a_{d+1}; b_{d+1}]$. Our strategy for selecting the values of the inputs at which to observe the response, during the sequential design, is to choose these values so that the corresponding scaled mean responses, $z_{i,d+1}$, are spread evenly throughout the scaled *response surface*; therefore, the sequential design successively fills the scaled response surface. Intuitively, the resulting design ensures that no two response values in the scaled response surface are "too close", and hence the scaled response values are spread over the scaled response surface. This means that in a region where the output function varies significantly, the inputs may be clustered; i.e. the values of the j component of the design points ($X_{i,j}$, $i = 1, \ldots n$) are not evenly spaced.

In order to determine the location of a new data point, to be added to the sequential design, we construct a wire-frame using a mesh of triangles. The use of triangles, instead of squares, ensures that only one data point is added on each iteration and that at most two new triangles are formed. The initial wire-frame can be build from an arbitrary set of design points using a Delaunay or convex hull algorithm [4]. A Delaunay triangulation in $\mathbb{R}^d$ can be computed from a convex hull in $\mathbb{R}^{d+1}$. Alternatively, the initial triangle wire-frame can be built from a small set of design points where the initial wire-frame is known in advance. For example, in $\mathbb{R}^2$, a simple 2^d factorial design, where for $d = 2$ only two triangles are initially built; see Figs. 1a and 2 (lines 1–3).

The sequential procedure, depicted in Fig. 2, that successively adds new data points to the experimental design, selects the triangle with the largest facet area (line 5). The triangle with the largest area represents the region where the data points are more spread out in the response surface. A new data point is inserted in the middle of the widest side of the selected triangle (lines 6 and 8), $\mathbf{M}_{kl} = (\mathbf{X}_{k\cdot} + \mathbf{X}_{l\cdot})/2$, where $\mathbf{X}_{k\cdot}$ and $\mathbf{X}_{l\cdot}$ are the side vertexes. The selected triangle is then split into two new triangles, using the new data point; see Figs. 1b and 2 (line 11). If that side of the triangle connects with another adjacent triangle, the other triangle is also split in two, using the same new data point, to ensure a coherent mesh representation (lines 12–14). In order to determine the area of the new triangles, the response of the simulation model must be evaluated at the new point (line 9).
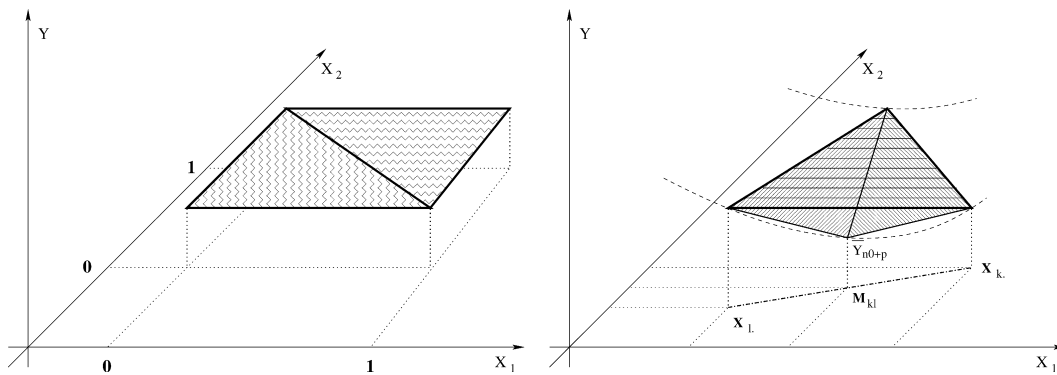
Fig. 1. (a) Initial experimental design; (b) $\mathbb{R}^3$ triangle splitting procedure.

Fig. 2. The proposed sequential experimental design procedure.

The new experimental design is $\mathscr{X}^{(p)} = \mathscr{X}^{(p-1)} \cup \{\mathbf{M}_{k'l'}\}$, or

$$\mathbf{X}^{(p)} = [\mathbf{X}_{1\cdot}^{\mathrm{T}} \ldots \mathbf{X}_{n_0\cdot}^{\mathrm{T}} \mathbf{X}_{n_0+1\cdot}^{\mathrm{T}} \ldots \mathbf{X}_{n_0+p-1\cdot}^{\mathrm{T}} \mathbf{X}_{n_0+p\cdot}^{\mathrm{T}}]^{\mathrm{T}}$$

in matrix notation, where $p = 1, 2, \ldots$ represents the pth step in the sequential design strategy.

Now execute r_i replications of the new design point, obtain the corresponding $\overline{Y}^{(p)} = [\overline{Y}_{1\cdot} \ldots \overline{Y}_{n_0+p\cdot}]^{\mathrm{T}}$. Fit the metamodel with this data, obtaining $\hat{\boldsymbol{\theta}}^{(p)}$. To compute the termination rule we check whether the metamodel accuracy fitted with a newly added design point has been achieved, or if the maximum number of design points is reached (lines 15 and 16). The degree of accuracy is judged by computing the statistical diagnostic $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{(p)})$ and comparing this value with a pre-defined target precision value ξ. The experimental design is complete if $\mathrm{MSE}(\hat{\boldsymbol{\theta}}^{(p)}) \leqslant \xi$ or a maximum number of points is reached. Otherwise, a new design point must be added.

The procedure defined in Fig. 2 uses a set of triangles, or a triangle-mesh, where each vertex refers to an experimental design point. Although not depicted, the triangle areas are associated with each triangle and computed once, for efficiency, and the triangle-mesh is, in fact, a list ordered by areas.

If the design is unidimensional ($d = 1$) no triangles are needed; it suffices to evaluate the distances between consecutive points. This distance is equivalent to the largest area of the triangle (line 5). A new point X_{n_0+p} is inserted in the middle of the largest distance between two consecutive points (line 8).

3.4. Collect the simulation data

The collection of simulation data requires some tactical decisions: perform a terminating or a steady-state simulation; determine the initial conditions; choose the final conditions such as run time or number of events completed; and decide on an appropriate balance between run length and number of replications.

Terminating simulations are those that run until a specific event has occurred (including the event of simulating a fixed amount of time). For terminating simulations, it may be necessary to censor results if we are simulating rare events. Steady-state simulations have no natural termination point, and can keep generating data for their analysis. For steady-state simulations, the warm-up period must be chosen carefully, and the length of the warm-up period affects the total experimentation time.

In steady-state simulations, the initial conditions may generate rare sequences of events, that introduce a warm-up period, until the system achieves statistical equilibrium. This warm-up creates bias in the simulation output values. For example, when all queues are empty, the simulation program begins with a sequence of events that may have a low probability of occurring. So, the simulation output can be contaminated with an initial bias [26, Section 9.5]. To control the initial bias, that can be the greatest source of error in the estimation of the metamodel parameters, an initial data deletion is performed. However, the initial data deletion selection involves a compromise, since in a small deletion the bias can still affect the result, while a large one may produce a large variance and it is not efficient. A good bias detection is important in the context of the independent replications method, because usually many runs are needed.

In this paper, we use Welch's procedure to detect the initial bias [26]. For each replication of the ith experimental point, we ignore the observations until the corresponding truncation point and collect the remaining simulation results; it turns out that approximately 85% of the observations are collected.

The selection of the number of replications in each design point is important since the response may be affected by other factors, besides the selected input values. These factors introduce effects that can be referred to as unsystematic (random error or noise) and as systematic (bias). Replication allows the estimation of the magnitude and distribution of random error, and the sample means of the replicated responses have smaller variances than the individual responses.

The number of replications can be selected such that the variances of the average response become approximately equal; see Kleijnen and van Groenendaal [21]. As in [18], we perform a *pilot sample* of $r_0 \geqslant 2$ replications for each design point and estimate the variances σ_i^2 through (2) with $r_i = r_0$ and compute $\hat{\sigma}_{\min} = \min_{i=1,\dots,n}\hat{\sigma}_i^2$. Then we use the two-stage procedure to determine an approximate number of additional replications, based on $\hat{r}_i = \lfloor r_i\hat{\sigma}_i^2/\hat{\sigma}_{\min}\rfloor$, and perform the additional $\hat{r}_i - r_i$ replications, where $\lfloor x \rfloor$ denotes the largest integer not bigger than x. The sequential sampling procedure is, in our case, performed independently on each data point. The procedure adds one replication at a time and determines the resulting value of $\hat{\sigma}_i^2$ until a value smaller than $\hat{\sigma}_{\min}$ is achieved. If additional replications are needed, a new value for $\hat{r}_i$ is determined, using the new r_i and the initial $\hat{\sigma}_{\min}$ computed with r_0. When, finally, the number of replications produces a value of $\hat{\sigma}_i^2$ smaller than the threshold limit $\hat{\sigma}_{\min}$, the final number of replications for data point i is r_i if $|\hat{\sigma}_i^2 - \hat{\sigma}_{\min}| < |\hat{\sigma}_{i-1}^2 - \hat{\sigma}_{\min}|$, or r_{i-1} otherwise. However, if the number of replications is too large or each run takes a long time, this procedure may become unrealistic and prohibitive. As a consequence, an estimated weighted least squares method must be used, with the largest number of replications available.

4. Application examples

All the models presented in this section were simulated using AweSim version 3.0 Pritsker et al. [29], and the metamodel was built in MATLAB 6.5 using some custom made routines. We obtained the least squares estimators $\hat{\theta}$ using the Levenberg–Marquart method, implemented in MATLAB 6.5, with a termination tolerance of 10^{-6} and a maximum number of function evaluations equal to 600.

An automobile parts factory model, depicted in Fig. 3, is used to illustrate the proposed sequential experimental design procedure. First, a unidimensional design is compared with a classical equally spaced design. Then, a two-dimensional experimental design is presented. The factory contains two drills and a finishing area, and processes two types of parts. Type I and type II parts arrive according to a gamma$(2,\lambda_1)$ and a gamma$(2,\lambda_2)$ distributions, respectively. After arriving, types I and II parts take 2 and 5 min to be routed to the drill area, respectively. Both parts require drilling and finishing operations and assume no time delays between these two operations. We assume that operators are always available if a machine is available. Drilling time is triangularly distributed between 10 and 30 min, with a mode of 15 min. The time to perform the finishing operations has a uniform distribution on the interval [2,5] minutes and only one part can be finished at a time. On the average, finishing operations have to be repeated 20% of the time. If a part has been routed through finishing twice but still needs to be refinished, it must be drilled again.
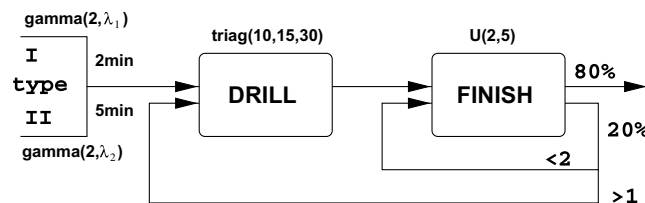

Fig. 3. The automobile parts factory diagram.

The purpose of the simulation experiment is to express the average time in the system (response), Y, as a function of the mean time between arrivals of the parts of types I and II, corresponding to $X_1 = 2\lambda_1$ and $X_2 = 2\lambda_2$ (decision variables). The unidimensional experimental designs consider $\lambda_2 = 20$ and use an equal number of design points (19) to provide comparable values.

The evenly spaced experimental design, denoted by *fixed*, uses $\mathscr{X}_{\text{fixed}}^{(0)} = \{0.5, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90\}$. This *fixed* experimental design is used as a reference to measure the relative improvement of the proposed procedure. At each design point, we run Welch's procedure in order to determine the length of each simulation and the initial-data deletion. Welch's moving average is based on $r_i = 10$ replications of the simulation model, where each replication contains 2000 observations. For example, at the design point $X_{8,1} = 35$ of the *fixed* experimental design, we delete 200 observations from the beginning of the run and we use only the next 1134 observations to estimate the response Y (the remaining 666 observations are ignored). In order to reduce the variation of the variances, we carry out a different number of replications of each of the $n = 19$ design points. An initial 10 replications are performed at each design point and, based on the procedure describe in Section 3.4, additional replications were performed. While data point $X_{14,1} = 65$ has the smallest variance, and no additional replications are performed, the data point $X_{8,1} = 35$ requires a total of 13 583 replications, corresponding to the highest number required.

The proposed sequential experimental design, denoted by *fill*, starts with two points $\mathscr{X}_{\text{fill}}^{(0)} = \{0.5, 90\}$, corresponding to the experimental region boundary or a 2^1 factorial design. Additional design points are inserted, one by one, in the middle of the largest gap between existing consecutive points, up to a total of 19 data points. As in the *fixed* experimental design, an initial set of 10 replications are performed at each design and then additional replications are performed using the procedure describe in Section 3.4. Since the *fill* experimental design searches for nonlinearities, many of the additional design points reside in the nonlinear region of the example that correspond to higher variances. Therefore, the highest number of replications (55 937) is obtained for $X_{17,1} = 33.7129$ ($\hat{\sigma}_{17}^2/r_{17} = 0.00659717$) and the lowest (10) at $X_{2,1} = 90$ ($\hat{\sigma}_2^2/r_2 = 0.00707208$) and $X_{15,1} = 56.4375$ ($\hat{\sigma}_{15}^2/r_{15} = 0.00659716$), in a total of 298 218 replications for all 19 design points. The resulting variances are approximately constant, so the null-hypothesis of equal variances is not rejected by test described in Section 2.2 ($28.8693 = \chi < \chi^2(18; 0.95) = 201.2721$), with a maximum of $\hat{\sigma}_7^2/r_7 = 0.00653807$ and a minimum of $\hat{\sigma}_2^2/r_2 = 0.00707208$, corresponding to $(\hat{\sigma}_7^2/r_7)/(\hat{\sigma}_2^2/r_2) = 1.082$, As a result, the method of nonlinear ordinary least squares is used.

To identify a curve that might fit the input–output relationship of the simulation model (computer program), we performed a visual check, observing the simulation results plots of each experimental design, in Fig. 4. Comparing the plots with graphical representations of known functional relationships, a good candidate seams to be the arc-tangent function

$$y = \theta_1 + \theta_2 \arctan(\theta_3 + \theta_4 x).$$

The final metamodel's estimated parameters obtained for the two experimental designs are presented in Table 1. To check the accuracy of the resulting metamodels, we observe that, with a bootstrap sample of size $B = 200$ gives $R^2 = 0.999476$ which falls into the acceptance interval $[0.999421, 1]$. So, the metamodel is not rejected.

In Table 2 the MSE and MSPE values for both designs are displayed, where MSE is defined in Section 3.3. The MSPE's *fixed* value is given by

$$\text{MSPE} = \frac{r_i}{n^* \hat{\sigma}_i^2} \sum_{i=1}^{n^*} [\overline{Y}_{i\cdot} - f(\mathbf{X}_{i\cdot}, \hat{\boldsymbol{\theta}})]^2,$$
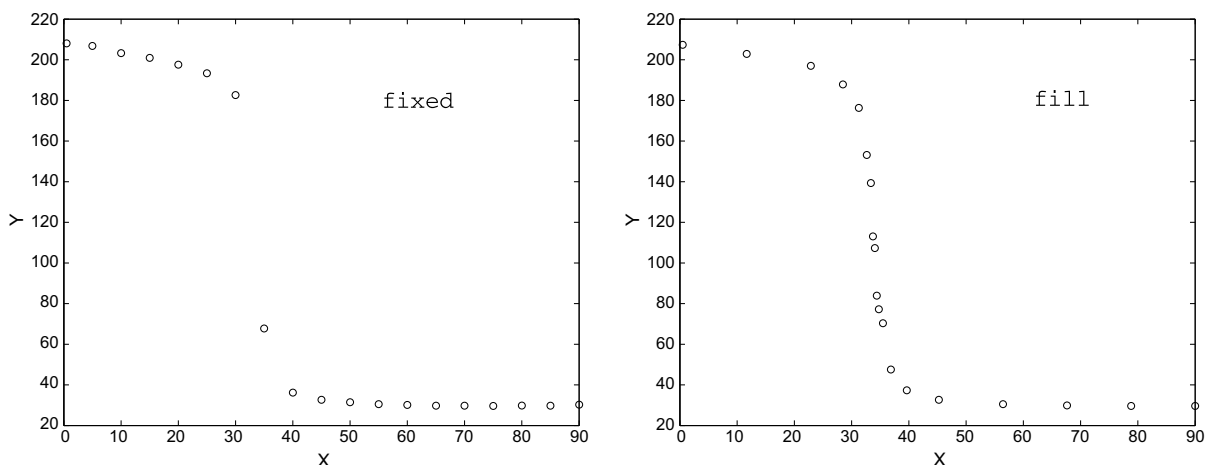

Fig. 4. *Fixed* and *fill* experimental designs, with 19 design points.

Table 1
Metamodel's estimated parameters for the experimental designs

Designs	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
Fixed	116.557	−57.0318	−23.0443	0.686365
Fill	116.854	−56.8379	−24.8619	0.736861

Table 2
Metamodel's MSE and MSPE

Design	MSE	MSPE	MSE + MSPE
Fixed	211.456	1556.70	1768.16
Fill	474.400	204.147	678.548

where $\hat{\theta}$ is obtained using the *fixed* experimental design and $\overline{Y}_{i\cdot}$ is the response value at $\mathbf{X}_{i\cdot}$ of the *fill* experimental design; and, vice-versa, the MSPE's *fill* value is determined using $\hat{\theta}$ from the *fill* design and $\mathbf{X}_{i\cdot}$ of the *fixed* design. It can be observed that the *fixed* design presents a lower MSE value than the *fill* design since it ignores most of the nonlinear behaviour. However, if we use the data available from the other experimental design as predictive data we conclude that the *fill* design performs much better in terms of MSPE. Even if we add both measures, corresponding to the use of all available data, the *fill* design still provides a significantly better adjustment.
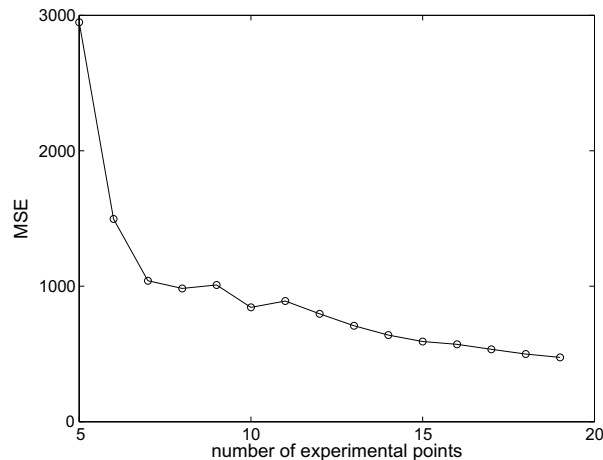
Fig. 5. The evolution of MSE as new points are added.
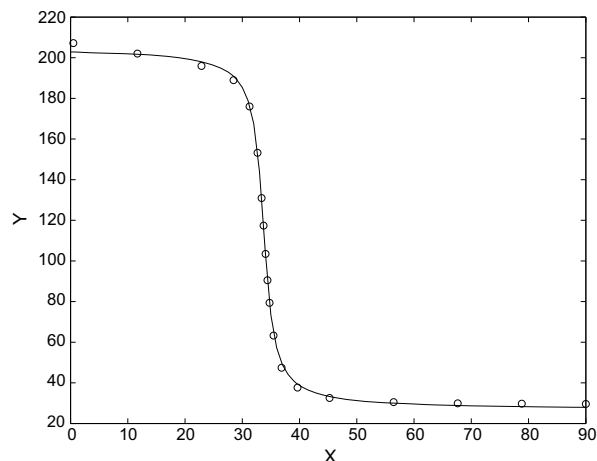
Fig. 6. The adjusted metamodel with 19 data points from the *fill* experimental design.

The use of a statistical diagnostic as a termination criterion implies a refinement of the fitted metamodel and, as a result, an increasingly smaller value of MSE as new design points are added to the *fill* experimental design; see Fig. 5. Fig. 6 depicts the resulting adjusted metamodel with superimposed design points.
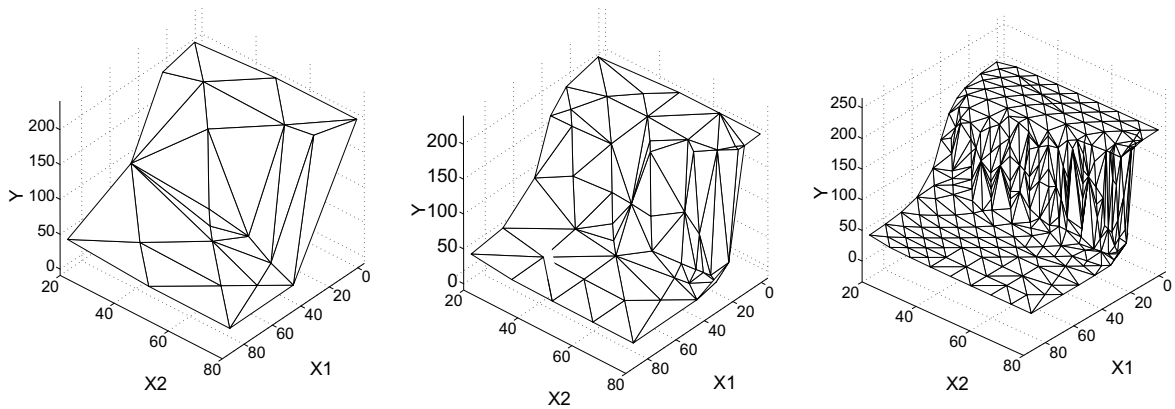

Fig. 7. 20, 50 and 248 data points experimental designs.

Table 3
Metamodel's estimated parameters and diagnostics

| Design	20	50	112	248
$\hat{\theta}_1$	117.720	117.022	116.900	116.635
$\hat{\theta}_2$	−57.9560	−57.5629	−57.4514	−58.3365
$\hat{\theta}_3$	−0.322912	−0.400339	−0.439303	−0.448109
$\hat{\theta}_4$	0.0215550	0.0240407	0.0260403	0.0261604
$\hat{\theta}_5$	−11.2109	−6.28069	−4.96593	−5.11790
$\hat{\theta}_6$	−0.0122992	−0.251322	−0.348255	−0.331101
$\hat{\theta}_7$	−0.00280678	−0.00114813	−0.000605079	−0.00106710
MSE	112.729	115.128	82.3239	100.539
max r_i	70140	70140	535320	535320
Total r_i	89900	374500	4075560	7643840
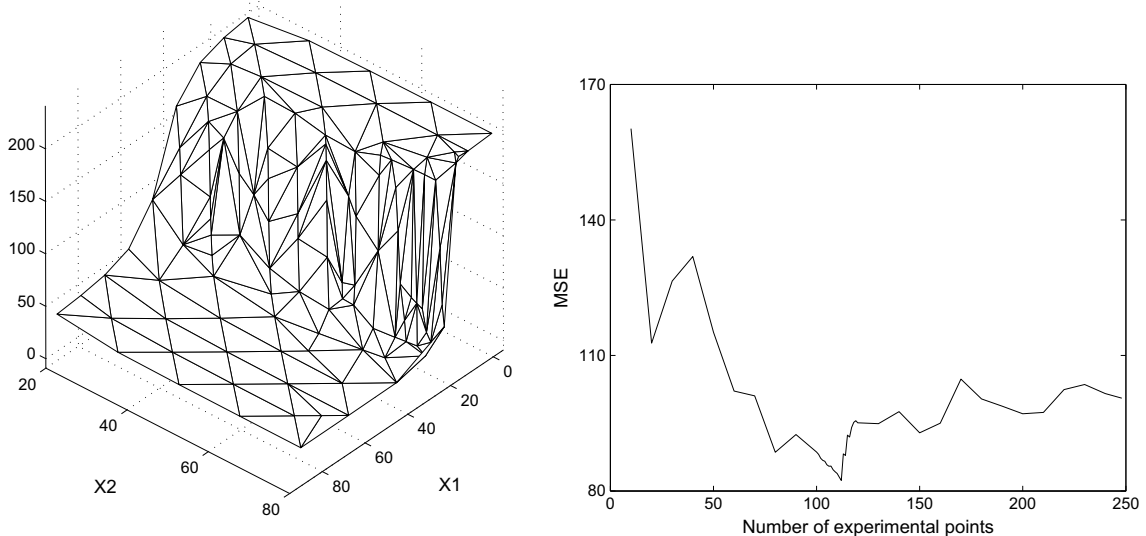
Fig. 8. Final experimental design with 112 data points and evolution of MSE.

The two-dimensional experimental designs use the same X_1 interval and X_2 values between 20 and 80. The experimental designs start with a 2^2 factorial design, corresponding to the region boundary points $\mathscr{X}^{(0)} = \{(0.5, 20), (0.5, 80), (90, 20), (90, 80)\}$. All subsequent design points are determined by the proposed sequential design procedure. The proposed procedure was executed upto 248 data points and partial data was collected for 20, 50 and 112 experimental points; see Fig. 7.

After performing an initial $r = 20$ replications for each design point, we concluded that the number of additional replications, requested by the two-stage procedure, for the worst case (max r_i, in Table 3) as well as the total number of additional replications for all design points (total r_i, in Table 3), were too expensive to perform. Therefore, the estimated weighted least squares method is used, since the variances of the various data points can not be considered constant, after the 20 replications. To relate the average time in system with the arrival of both types of parts we used the unidimensional metamodel expression combined with a polynomial component for X_2, resulting in

$$Y = \theta_1 + \theta_2 \arctg(\theta_3 X_1 + \theta_4 X_1 X_2 + \theta_5 + \theta_6 X_2 + \theta_7 X_2^2).$$

The adjusted resulting values for the unknown θ parameters are presented in Table 3, as well as the degree of accuracy determined by the statistical measure MSE. Based on these results and the evolution of MSE as new data points are added, the experimental design with 112 data points proved to be the best choice (see Fig. 8). The resulting metamodel is not rejected by the bootstrapped R^2 accuracy test since $0.999566 = R^2 \in [0.999521; 1]$, with $B = 200$.

5. Conclusions

Metamodels can be used, as simulation model surrogates, to expose the fundamental nature of the input–output relationships. Represented as a simple mathematical function, the metamodel can also be used for verifying and validating the original simulation model. Queuing systems, as many other real-life systems, exhibit nonlinear behavior. Nonlinear regression metamodels provide a better and more realistic global fit than polynomials since they are able to fit curves of arbitrary shape.

In order to provide a good fit, the sample design points must be chosen to efficiently investigate the input–output relationship of the simulation model. A careful choice of an experimental design may better expose the relation between the input and the response, such as subtle nonlinearities of the simulation model. This choice should be complemented with an accurate and statistically sound estimation of the metamodel parameters using a small set of design points. In order to reduce the number of design points, without compromising its accuracy, an experimental design strategy is proposed.

The design should be generic, in spite of the fact that the paper concentrates on nonlinear regression metamodels, and should be easy to construct if they are to be commonly used in practice. The proposed approach is a sequential design where an initial set of design points, possibly from a well known classical experimental design, is augmented by a surface filling approach. The surface filling process considers both the inputs and the output, and is based on a wire-frame of using a mesh of triangles built on a scaled response surface. A new point is added in the triangle with the largest area, splitting it into two equally sized triangles. The number of steps in the sequential design depends on the interesting features of the metamodel. The main objective of the proposed design strategy is to provide a better fit of the metamodel through a careful choice of design points.

A statistical diagnostic, MSE, is employed to ensure a sufficient number of design points, working as an indicator of the metamodel's accuracy. The investigation of the improved experimental design included a comparative study between an equally spaced classical design and the proposed design, using the same number of design points. The same example is extended to two inputs and the number of design points is successively increased. The accuracy of each design is then evaluated and a metamodel is built and validated.

This paper stresses that a careful choice of an experimental design can lead to better metamodels, with the same simulation effort, and proposes a sequential design that improves the accuracy of the constructed nonlinear simulation metamodels.

The construction of multivariate metamodels presents new challenges that includes the use of already existing Delaunay tesselations algorithms from N-dimensional convex hulls. For the specific case of two responses and a single input, the adaptation of the proposed algorithm is not complex. Additionally, a new estimation procedure must be addressed using, for example, the application of multivariate nonlinear regression statistical techniques.

Future tunning of nonlinear experimental designs should include, for instance, screening techniques when the number of factors is high. The presented designs may require very large initial simulations before the sequential approach can be applied. Screening techniques should be applied to detect the few really important factors.

Appendix

(i) Verification of (3). The least squares estimate of θ of the problem (1), denoted by $\hat{\theta}$, minimizes the error sum of squares

$$\mathrm{SSE}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{1}{\sigma_i^2} [Y_{ij} - f(\mathbf{X}_{i.}, \boldsymbol{\theta})]^2$$

over $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}$. Since we have the usual decomposition

$$\sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{1}{\sigma_i^2} [Y_{ij} - f(\mathbf{X}_{i.}, \boldsymbol{\theta})]^2 = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \frac{1}{\sigma_i^2} (Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{n} \frac{r_i}{\sigma_i^2} [\overline{Y}_{i.} - f(\mathbf{X}_{i.}, \boldsymbol{\theta})]^2,$$

to find the least squares estimate $\hat{\theta}$ of θ is equivalent to minimizing

$$\sum_{i=1}^{n} \frac{r_i}{\sigma_i^2} [\overline{Y}_{i.} - f(\mathbf{X}_{i.}, \theta)]^2, \tag{10}$$

i.e., a weighted least squares analysis weights r_i/σ_i^2 correspondent to the problem (5); consequently the problem (1) and (5) are equivalent. Consequently, the normal equations for $\hat{\theta}$ are

$$-2\sum_{i=1}^{n} \frac{r_i}{\sigma_i^2} [\overline{Y}_{i.} - f(\mathbf{X}_{i.}, \theta)] \frac{\partial f(\mathbf{X}_{i.}, \theta)}{\partial \theta_k}\bigg|_{\theta=\hat{\theta}} = 0, \quad k = 1, 2, \ldots, m.$$

These equations correspond to the problem

$$\mathbf{W} = \mathbf{g}(\mathbf{X}, \theta) + \boldsymbol{\eta}, \tag{11}$$

where $\mathbf{W} = \mathbf{R}\overline{\mathbf{Y}}$, $\mathbf{g}(\mathbf{X}, \theta) = \mathbf{R}\mathbf{f}(\mathbf{X}, \theta)$ and $\boldsymbol{\eta} = \mathbf{R}\boldsymbol{\epsilon}$, with $\mathbf{R} = (\mathbf{U}^\mathrm{T})^{-1}$ where $\mathbf{U}$ is obtained from the decomposition $\boldsymbol{\Sigma} = \mathbf{U}^\mathrm{T}\mathbf{U}$ and $\boldsymbol{\epsilon} = (\bar{\epsilon}_{1.}, \ldots, \bar{\epsilon}_{n.})^\mathrm{T}$; we consider the assumption that $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix, that accepts the Cholesky decomposition, $\boldsymbol{\Sigma} = \mathbf{U}^\mathrm{T}\mathbf{U}$. Therefore, we have $\mathrm{E}[\boldsymbol{\eta}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\eta}] = \mathbf{R}\,\mathrm{Var}[\boldsymbol{\epsilon}]\mathbf{R}^\mathrm{T} = \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^\mathrm{T} = \mathbf{R}\mathbf{U}^\mathrm{T}\mathbf{U}\mathbf{R}^\mathrm{T} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order n. Applying Eq. (2.15) of Seber and Wild [37] to the ordinary least squares problem (11) we find that, for large n and certain regularity conditions, the least squares estimator of θ satisfies $\hat{\theta} \approx \theta + [\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1}\mathbf{G}^\mathrm{T}[\mathbf{W} - \mathbf{g}]$, where $\mathbf{G}$ is the Jacobian matrix of $\mathbf{g}$ both evaluated at θ. Therefore, using $\mathbf{g}(\mathbf{X}, \theta) = \mathbf{R}\mathbf{f}(\mathbf{X}, \theta)$ (and so, $\mathbf{G}(\theta) = \mathbf{R}\mathbf{F}(\theta)$), $\mathbf{W} = \mathbf{R}\overline{\mathbf{Y}}$ and $\mathbf{R}^\mathrm{T}\mathbf{R} = \boldsymbol{\Sigma}^{-1}$, we obtain

$$\hat{\theta} \approx \theta + [\mathbf{F}^\mathrm{T}\boldsymbol{\Sigma}^{-1}\mathbf{F}]^{-1}\mathbf{F}^\mathrm{T}\boldsymbol{\Sigma}^{-1}[\overline{\mathbf{Y}} - \mathbf{f}(\mathbf{X}, \theta)].$$

(ii) Verification of (4). The application of Seber and Wild [37, Theorem 2.1(i)] to the problem (11) results on $\hat{\theta} \sim \mathrm{N}_p[\theta, (\mathbf{G}^\mathrm{T}\mathbf{G})^{-1}]$. Using $\mathbf{G} = \mathbf{R}\mathbf{F}$ and $\mathbf{R}^\mathrm{T}\mathbf{R} = \boldsymbol{\Sigma}^{-1}$, we obtain

$$\hat{\theta} \sim \mathrm{N}_p[\theta, (\mathbf{F}^\mathrm{T}\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}].$$

(iii) Verification of (6). The result (6) is obtained applying (7) of Santos and Nova [33] to the problem (5), with the approximation $\mathrm{Var}[\overline{Y}_{i.}] = \sigma_i/r_i = \sigma^2 = $ constant.

(iv) Verification of (7). The result (7) is obtained applying (12.23) of Seber and Wild [37] to the problem (5), using the approximation $\mathrm{Var}[\overline{Y}_{i.}] = \sigma_i/r_i = \sigma^2 = $ constant.

(v) Verification of (8). The estimator $\hat{\sigma}^2$ in (8) can be written in the form

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}(\hat{\theta})}{N - m} \tag{12}$$

with $N = \sum_{i=1}^{n} r_i$, $\mathrm{SSE}(\hat{\theta}) = \|\mathbf{Y} - \tilde{\mathbf{f}}(\hat{\theta})\|_2^2$, and where $\mathbf{Y} = (Y_{11}, \ldots, Y_{1r_1}, \ldots, Y_{n1}, \ldots, Y_{nr_n})^\mathrm{T}$ and $\tilde{\mathbf{f}}(\hat{\theta}) = (f(\mathbf{X}_{1.}, \hat{\theta}), \ldots, f(\mathbf{X}_{1.}, \hat{\theta}), \ldots, f(\mathbf{X}_{n.}, \hat{\theta}), \ldots, f(\mathbf{X}_{n.}, \hat{\theta}))^\mathrm{T}$. Using (2.18) of Seber and Wild [37], we obtain

$$\mathrm{SSE}(\hat{\theta}) = \boldsymbol{\epsilon}^\mathrm{T}(\mathbf{I}_N - \mathbf{P}_{\widetilde{F}})\boldsymbol{\epsilon}, \tag{13}$$

where $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1r_1}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{nr_n})^\mathrm{T}$ and $\mathbf{P}_{\widetilde{F}}$ is a projection matrix $\mathbf{P}_{\widetilde{F}} = \widetilde{\mathbf{F}}(\widetilde{\mathbf{F}}^\mathrm{T}\widetilde{\mathbf{F}})^{-1}\widetilde{\mathbf{F}}^\mathrm{T}$, with $\widetilde{\mathbf{F}}$ the Jacobian matrix of $\tilde{\mathbf{f}}$ calculated in $\hat{\theta}$. Since $\mathbf{I}_N - \mathbf{P}_{\widetilde{F}}$ is a symmetric and idempotent matrix (that is, $\mathbf{I}_N - \mathbf{P}_{\widetilde{F}} = (\mathbf{I}_N - \mathbf{P}_{\widetilde{F}})^\mathrm{T}$ and $(\mathbf{I}_N - \mathbf{P}_{\widetilde{F}})^2 = \mathbf{I}_N - \mathbf{P}_{\widetilde{F}}$) and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, \ldots, r_i$, we apply the Theorem B.4 of [38] to (13) and conclude that $\sigma^{-2}\mathrm{SSE}(\hat{\theta})$ has a chi-square distribution with $\mathrm{tr}(\mathbf{I}_N - P_{\widetilde{F}})$ degrees of freedom ('tr' represents the trace of a matrix). But $\mathrm{tr}(\mathbf{I}_N - \mathbf{P}_{\widetilde{F}}) = \mathrm{tr}(\mathbf{I}_N) - \mathrm{tr}(\mathbf{P}_{\widetilde{F}}) = N - \mathrm{tr}[\widetilde{\mathbf{F}}(\widetilde{\mathbf{F}}^\mathrm{T}\widetilde{\mathbf{F}})^{-1}\widetilde{\mathbf{F}}^\mathrm{T}] = N - \mathrm{tr}[\widetilde{\mathbf{F}}^\mathrm{T}\widetilde{\mathbf{F}}(\widetilde{\mathbf{F}}^\mathrm{T}\widetilde{\mathbf{F}})^{-1}] = N - \mathrm{tr}(\mathbf{I}_m) = N - m$, that is, $\sigma^{-2}\mathrm{SSE}(\hat{\theta}) \sim \chi_{N-m}^2$. Moreover, $\mathrm{SSE}(\hat{\theta}) = (N - m)\hat{\sigma}^2$ (see (12)), so the result (8) is verified.

References

[1] F.M. Alam, K.R. McNaught, T.J. Ringrose, A comparison of experimental designs in the development of a neural network simulation metamodel, Simulation Modeling Practice and Theory 12 (2004) 559–578.

[2] T.T. Allen, M.A. Bernshteyn, K.K. Yu Kabiri, A comparison of alternative methods for constructing meta-models for computer experiments, Journal of Quality Technology 35 (2) (2003) 1–17.

[3] A.B. Badiru, D.B. Sieger, Neural network as a simulation metamodel in economic analysis of risky projects, European Journal of Operational Research 105 (1998) 130–142.

[4] C.B. Barber, D.P. Dobkin, H.T. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Transactions on Mathematical Software 22 (4) (1996) 469–483.

[5] R.R. Barton, Simulation metamodels, in: Proceedings of the Winter Simulation Conference, Piscataway, IEEE, NJ, 1998, pp. 167–174.

[6] M.C. Bernardo, R.J. Buck, L. Liu, W.A. Nazaret, J. Sacks, W.J. Welch, Integrated circuit design optimization using a sequential strategy, IEEE Transactions on Computer-Aided Design 11 (1992) 361–372.

[7] M.W. Brantley, C.H. Chen, A moving mesh approach for simulation budget allocation on continuous domains, in: Proceedings of the Winter Simulation Conference, Piscataway, IEEE, NJ, 2005, pp. 699–707.

[8] R.C.H. Cheng, J.P.C. Kleijnen, Improved design of queueing simulation experiments with highly heteroscedastic responses, Operations Research 47 (5) (1999) 762–777.

[9] W.J. Conover, Practical Nonparametric Statistics, John Wiley and Sons, NY, USA, 1971.

[10] J.M. Donohue, Experimental designs for simulation, in: Proceedings of the Winter Simulation Conference, Piscataway, IEEE, NJ, 1994, pp. 200–206.
[11] J.M. Donohue, E.C. Houck, R.H. Myers, Simulation designs for quadratic response surface models in the presence of model misspecification, Management Science 38 (12) (1992) 1765–1791.
[12] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman-Hall, London, UK, 1993.
[13] I. Ford, B. Torsney, C.F.J. Wu, The use of a canonical form in the construction of locally optimal designs for non-linear problems, Journal of the Royal Statistical Society Series B 54 (2) (1992) 569–583.
[14] W. Hendrickx, T. Dhaene, Sequential design and rational metamodeling, in: Proceedings of the Winter Simulation Conference, Piscataway, IEEE, NJ, 2005, pp. 290–298.
[15] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, Journal of Global Optimization 13 (1998) 455–492.
[16] A.C. Keys, L.P. Rees, A sequential-design metamodeling strategy for simulation optimization, Computers and Operational Research 31 (2004) 1911–1932.
[17] J.P.C. Kleijnen, Statistical Tools for Simulation Practitioners, Marcel Dekker, NY, 1987.
[18] J.P.C. Kleijnen, White noise assumptions revisited: Regression metamodels and experimental designs in practice, in: Proceedings of the Winter Simulation Conference, Piscataway, IEEE, NJ, 2006, pp. 107–117. December.
[19] J.P.C. Kleijnen, R.C.H. Cheng, B. Bettonvil, Validation of trace-driven simulation models: Bootstrap tests, Management Science 47 (11) (2001) 1533–1538. November.
[20] J.P.C. Kleijnen, D. Deflandre, Validation of regression metamodels in simulation: Bootstrap approach, European Journal of Operational Research 170 (1) (2006) 120–131.
[21] J.P.C. Kleijnen, W. van Groenendaal, Two-stage versus sequential sample-size determination in regression analysis of simulation experiments, American Journal of Mathematical and Management Sciences 15 (1) (1995) 83–114.
[22] J.P.C. Kleijnen, S.M. Sanchez, T.W. Lucas, T.M. Cioppa, A user's guide to the brave new world of designing simulation experiments, INFORMS Journal on Computing 17 (3) (2005) 263–289.
[23] J.P.C. Kleijnen, R.G. Sargent, A methodology for fitting and validating metamodels in simulation, European Journal of Operational Research 120 (2000) 14–29.
[24] J.P.C. Kleijnen, W.C.M. van Beers, Application-driven sequential designs for simulation experiments: Kriging metamodeling, Journal of the Operational Research Society 55 (8) (2004) 876–883.
[25] M. Koc, T. Allen, S. Jiratheranat, T. Altan, The use of fea and design of experiments to establish design guidelines for simple hydroformed parts, International Journal of Machine Tools and Manufacture 40 (2000) 2249–2266.
[26] A.M. Law, Simulation Modeling and Analysis, fourth ed., Mc-Graw-Hill, NY, USA, 2007.
[27] M.D. McKay, R.J. Beckman, W.J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 42 (1) (2000) 55–61.
[28] K. Palmer, K.L. Tsui, A minimum bias Latin hypercube design, IIE Transactions 33 (2001) 793–808.
[29] A. Pritsker, J. O'Reilly, D. LaVal, Simulation with Visual SLAM and AweSim, John Wiley and Sons, NY, USA, 1997.
[30] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, Statistical Science 4 (4) (1989) 409–435.
[31] T.J. Santner, B.J. Williams, W.I. Notz, The Design and Analysis of Computer Experiments, Springer-Verlag, NY, USA, 2003.
[32] M.I. Santos, A.M. Porta Nova, The main issues in nonlinear simulation metamodel estimation, in: Proceedings of the Winter Simulation Conference, Piscataway, IEEE, NJ, 1999, pp. 502–509.
[33] M.I. Santos, A.M. Porta Nova, Estimation of nonlinear simulation metamodels using control variables, Journal of Statistical Computation and Simulation 75 (12) (2005) 959–973.
[34] M.I. Santos, A.M. Porta Nova, Statistical fitting and validation of nonlinear simulation metamodels: a case study, European Journal of Operational Research 171 (1) (2006) 53–63.
[35] M.I. Santos, P.M. Santos, Sequential designs for simulation experiments: Nonlinear regression metamodeling, in: Proceedings of the 26th IASTED International Conference Modelling, Identification and Control, IASTED, February 2007, pp. 88–93.
[36] M.J. Sasena, P. Papalambros, P. Goovaerts, Exploration of metamodeling sampling criteria for constrained global optimization, Engineering Optimization 34 (3) (2002) 263–278.
[37] G.A.F. Seber, C.J. Wild, Nonlinear Regression, John Wiley and Sons, NY, USA, 1989.
[38] A. Sen, M. Srivastava, Regression Analysis, Theory, Methods, and Applications, Springer-Verlag, NY, 1990.
[39] T.W. Simpson, J.D. Peplinski, P.N. Koch, J.K. Allen, Metamodels for computer-based engineering design: Survey and recommendations, Engineering with Computers 17 (2001) 129–150.
[40] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, A. Juditsky, Nonlinear black-box modeling in system identification: a unified overview, Automatica 31 (12) (1995) 1691–1724.
[41] J.G. Voelkel, The efficiencies of fractional factorial designs, Technometrics 47 (4) (2005) 488–501.
[42] W.J. Welch, R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell, M.D. Morris, Screening, prediction, and computer experiments, Technometrics 34 (1) (1992) 15–25.
[43] H. White, Nonlinear regression on cross-section data, Econometrica 48 (3) (1980) 721–746.
[44] K.Q. Ye, Orthogonal column Latin hypercubes and their application in computer experiments, Journal of the American Statistical Association 93 (444) (1998) 1430–1439.