2011

An empirical study of the prediction performance of space-filling designs

Johnson, Rachel T.; Montgomery, Douglas C.; Jones, Bradley

Inderscience Enterprises, Ltd.

International Experimental Design and Process Optimisation, v. 2, no. 1, 2011, pp. 1-18.
http://hdl.handle.net/10945/47617

An empirical study of the prediction performance of space-filling designs

Rachel T. Johnson*

Operations Research Department,
Naval Postgraduate School,
1411 Cunningham Road, Monterey, CA 93943, USA
E-mail: rtjohnso@nps.edu
*Corresponding author

Douglas C. Montgomery

School of Computing, Informatics and Decision Systems Engineering,
Arizona State University,
P.O. Box 878809, Tempe, AZ 85287-8809, USA
E-mail: doug.montgomery@asu.edu

Bradley Jones

SAS Institute,
SAS Campus Drive, Cary, NC 27513, USA
E-mail: Bradley.Jones@jmp.com

Abstract: The Gaussian process (GASP) model has found widespread use as a surrogate model for results from deterministic computer model output. In this paper, we compare the fits of GASP models to specific space-filling designs based on their accuracy in predicting responses at previously unsampled locations. This is done empirically using several test functions. We demonstrate that no one space-filling design outperforms another with respect to prediction accuracy. We also found that while the GASP model is substantially easier to fit using the cubic correlation function than with the Gaussian correlation function, its prediction accuracy is not quite as good as the Gaussian correlation function for the chosen test functions especially for larger sample sizes. The best way to improve prediction accuracy is to increase the number of simulation runs, which suggests that the efficient augmentation of space filling designs is an important area for further research.

Keywords: correlation functions; CFs; Gaussian process models; jackknife plots; sample size; computer experiments.

Reference to this paper should be made as follows: Johnson, R.T., Montgomery, D.C. and Jones, B. (2011) ‘An empirical study of the prediction performance of space-filling designs’, *Int. Experimental Design and Process Optimisation*, Vol. 2, No. 1, pp.1–18.

Biographical notes: Rachel T. Johnson is an Assistant Professor in the Operations Research Department at the Naval Postgraduate School. She received her PhD and MS in Industrial Engineering from Arizona State University and BS in Industrial Engineering from Northwestern University. Her research interests are in the design and analysis of both physical and computer experiments and simulation methodology. She is a recipient of the Mary G. and Joseph Natrella Scholarship for Excellence in Statistics. She is a member of the American Society for Quality and the Institute for Operations Research and Management Sciences.

Douglas C. Montgomery is a Regents' Professor of Industrial Engineering and Statistics, ASU Foundation Professor of Engineering, and Co-director of the Graduate Program in Statistics at Arizona State University. He received his PhD in Engineering from Virginia Tech. His professional interests are in statistical methodology for problems in engineering and science. He is a recipient of the Shewhart Medal, the George Box Medal, the Brumbaugh Award, the Lloyd S. Nelson Award, the William G. Hunter Award, and the Ellis Ott Award. He is one of the current chief editors of *Quality and Reliability Engineering International*.

Bradley Jones is the Principal Research Fellow at SAS Institute and a Guest Professor at the University of Antwerp. He is the inventor of the prediction profile plot, an interactive graph for exploring multi-dimensional response surfaces. At SAS Institute, he is responsible for the design of experiments capabilities in the JMP software package. He is a Fellow of the American Statistical Association and a winner of the Brumbaugh Award for 2009.

1 Introduction

Computer experiments are experimental designs used to study a deterministic computer simulation model. Sacks et al. (1989b) introduced the Gaussian process (GASP) model for fitting the response variable in a computer experiment. Subsequently the GASP model has found widespread use in the computer experiments literature. Examples of theoretical and empirical studies of the GASP model can be found with applications in calibration (Kennedy and O'Hagan 2001), validation (Bayarri et al., 2007), screening (Welch et al., 1992; Linkletter et al., 2006), and response fitting/prediction (Currin et al., 1991). The GASP model is an attractive modelling choice for because it:

1 is an exact interpolator

2 typically provide fairly accurate predictions of the response at unobserved factor settings.

Because of its flexibility in approximating complex response surfaces, Kleijnen and Beers (2005), and Ankenman et al. (2008) also used the GASP model to fit stochastic simulation models.

Space-filling designs appear almost exclusively in the computer experiments literature. The goal of space-filling designs is to explore the entire region of interest. This is in contrast to screening designs for physical experiments that tend to place set factors at the extremes of their ranges of interest. Johnson et al. (2010) presented a theoretical comparison of space-filling designs based on the integrated prediction variance with respect to the GASP model. The results indicated that the GASP integrated mean square error (GP IMSE) design had the lowest theoretical integrated prediction variance across a range of potential responses. The range of responses was generated by comparing designs with respect to a variety of parameter vectors, θ, and a range of sample sizes for experiments with two, three, and four factors. In this work, we are interested in two main topics: experimental design choice and sample size. Specifically, we investigate the following questions:

- Does it matter in practice what design you choose? That is, is there a dominating experimental design that performs better in terms of model fitting and prediction?
- Does the choice of the form of the correlation function (CF) matter? That is, does the choice of one CF over another provide a fitted model with better predictive power?
- What is the role of sample size in experimental designs used to fit the GASP model? At what point do prediction error variance and other measures of prediction performance become reasonably small with respect to N, the sample size chosen?

Work by Hussain et al. (2002) and Allen et al. (2003) address some questions about the prediction performance of surrogate models used for computer simulation output in their respective papers. Both of these papers focus on the fit of the surrogate model. The focus of this paper is on the design, CF, and sample size and is entirely empirical. In the first part of the paper, we focus on experimental design choice. Specifically we investigate the class of space-filling designs. Our comparisons are made by investigating the prediction quality of fitted GASP models to the selection of design points prescribed by the space-filling design. This is followed by a section that investigates two different CF choices for the GASP model in order to determine if one correlation structure provides any advantages in terms of prediction accuracy. In the final section, we consider sample size. Work by Loeppky et al. (2008) indicate that $N = 10\,p$, where N is the sample size and p is the number of factors in the experiment, holds as an adequate rule of thumb for determining the sample size for a computer experiment. We show that the complexity of the response has an effect on recommended sample size values when precision of prediction of the response surface is of interest to the modeller.

Section 2 provides a brief introduction to the GASP model. Section 3 compares design types empirically. Section 4 compares two different CF s used in the GASP model. Section 5 presents a comparison of sample size. Section 6 presents our conclusions and future work.

2 GASP model

The GASP model operates on the notion that a realisation of a random function can mimic a deterministic response, $y(\mathbf{x})$. Typically, the random function is a multivariate normal distribution. So the system is being modelled as a stochastic process where the simulation model output is viewed as a deterministic realisation of that process. The output response is an $n \times 1$ data vector $\mathbf{y}(\boldsymbol{x})$, where $\mathbf{y}(\boldsymbol{x})$ is $\mathrm{N}(\mu 1_{\mathrm{n}}, \sigma^2\, R(\boldsymbol{X}, 0))$. $R(\boldsymbol{X}, \boldsymbol{\theta})$ is an $n \times n$ correlation matrix that can take a variety of forms (see Sacks et al., 1989a;

Santner et al., 2003). In this paper, we consider two different forms of the CF. The first is a special case of the power exponential CF and is sometimes referred to as the Gaussian correlation function (GCF) with the form

$$R_{ij}(\boldsymbol{X},\boldsymbol{\theta}) = \exp\left(-\sum\nolimits_{k=1}^{p}\theta_k\left(x_{ik}-x_{jk}\right)^2\right) \tag{1}$$

where $\theta_k \geq 0$. Another choice of CF is a one-dimensional cubic correlation function (CCF). The specific form of the CCF that we have used is

$$R_{ij}(\boldsymbol{X},\boldsymbol{\theta}) = \prod\nolimits_{k=1}^{p}\begin{cases} 1-6\left(\theta\left(x_k-x_{jk}\right)\right)^2+6\left(\theta\left|x_k-x_{jk}\right|\right)^3, & \theta \leq .5\left|x_k-x_{jk}\right|^{-1} \\ 2\left(1-\left(\theta\left|x_k-x_{jk}\right|\right)^3\right), & .5\left|x_k-x_{jk}\right|^{-1} < \theta \leq \left|x_k-x_{jk}\right|^{-1} \\ 0, & \left|x_k-x_{jk}\right|^{-1} < \theta \end{cases} \tag{2}$$

Note that if $\theta_k = 0$, then the correlation is 1.0 across the range of the kth factor and the fitted surface will be flat in that direction. Large θ_k corresponds to low correlation in the kth factor and the fitted surface will exhibit strong curvature in the direction of the kth variable.

Using either CF, the fitted GASP prediction equation is

$$\hat{y}(\boldsymbol{x}) = \hat{\mu} + \boldsymbol{r}'\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)\boldsymbol{R}^{-1}\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{y}-\hat{\mu}1_{\boldsymbol{n}}\right) \tag{3}$$

where the fitted mean, variance, and θ_j's are represented by $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$. These parameters are estimated via maximum likelihood. In the fitted equation, $\boldsymbol{r}\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)$ is an $n \times 1$ vector of estimated correlations of the unobserved $\boldsymbol{y}(\boldsymbol{x})$ at a new value of the explanatory variables with the observations in the original data. The choice of CF is determined by the user and the elements of $\boldsymbol{r}\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)$ depends on the choice of the CF. Using either of the CFs, $\hat{y}(\boldsymbol{x})$ interpolates the data.

The relative prediction variance can be calculated from

$$\frac{Var\left(\hat{y}(\boldsymbol{x})\right)}{\sigma^2} = 1-\boldsymbol{r}'\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)R^{-1}\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)\boldsymbol{r}\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)+\frac{\left(1-1'\boldsymbol{R}^{-1}\left(\boldsymbol{X},\hat{\boldsymbol{\theta}}\right)\boldsymbol{r}\left(\boldsymbol{X},\hat{\boldsymbol{\theta}}\right)\right)^2}{1'\boldsymbol{R}^{-1}\left(\boldsymbol{X},\hat{\boldsymbol{\theta}}\right)1} \tag{4}$$

The variance of the predicted response at a new point depends on the design, $\boldsymbol{X}$, and the unknown parameter vector, $\boldsymbol{\theta}$ It also depends implicitly on the sample size (number of rows in $\boldsymbol{X}$) and the dimensionality (number of columns in $\boldsymbol{X}$).

3 An empirical comparison of experimental designs

In this section, we compare the prediction quality of several space-filling designs fit by the GASP model using the GCF presented in equation (2). The generic procedure to compare the designs follows:

Step 1 Choose a test function.

Step 2 Choose a sample size and space-filling design.

Step 3 Create the design with the number of factors equal to the number in the test function chosen in step 1 and the specifications set in step 2.

Step 4 Using the test function in 1 find the values that correspond to each row in the design.

Step 5 Fit the GASP model.

Step 6 Generate a set of 40,000 uniformly random selected points in the design space and compare the predicted value (generated by the fitted GASP model) to the actual value (generated by the test function) at these points. Using equation (5) below, compute the root mean square error (RMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}} \qquad n = 40,000 \tag{5}$$

In Step 6 of the generic procedure, we make empirical comparisons – using RMSE – based on the difference between the fitted GASP model and the actual response, which are known because we are using test functions which act as surrogates for deterministic computer simulation code. Test functions are ideal for making comparisons because they allow the researcher the ability to know every ‘true’ value within the design region, simply by solving the mathematical equation for a given set of input values. The main purpose in this section is to evaluate prediction performance of space-filling designs with respect to the GASP model fits in order to answer the question: is there a difference in prediction performance of the fitted GASP model with respect to experimental design chosen? If the answer is, yes, this would imply that it would be better to choose one design over another.

The design types that we compare in this paper are the maximin Latin hypercube design (LHD), the uniform design (U), the sphere packing design (SP), and the GP IMSE design. The LHD was developed by McKay et al. (1979) and the maximin LHD was explored in Morris and Mitchell (1995). The U design was developed by Fang (1980). The SP design was proposed in Johnson et al. (1990). The GP IMSE design was presented in Sacks et al. (1989b). For examples of two dimensional plots of all the designs investigated see Johnson and Jones (2009). In Section 3.1, we present the four test functions we use to make the comparisons and in Section 3.2, we provide results of experimental design comparisons. In Section 3.3, we present an ANOVA using the results.

3.1 Test functions

We use four different test functions to compare design types. Test functions 1 and 2 both have two factors. Test function 3 has three factors and test function 4 has eight factors, but only four are significant and for the purpose of this paper we only vary four in the experiment.

3.1.1 Test function 1

The first test function was introduced by Welch et al. (1992). Allen et al. (2003) also employed this function to compare the performance of several design types with respect to the Gaussian process and linear regression models. The function is

$$y(x_1, x_2) = \left[30 + x_1 \sin(x_1)\right]\left(4 + \exp\{-x_2\}\right) \tag{6}$$

where $0 \le x_1, x_2 \le 5$. A surface plot of test function 1 is shown in Figure 1.

Figure 1 Surface plot of test function 1 (see online version for colours)

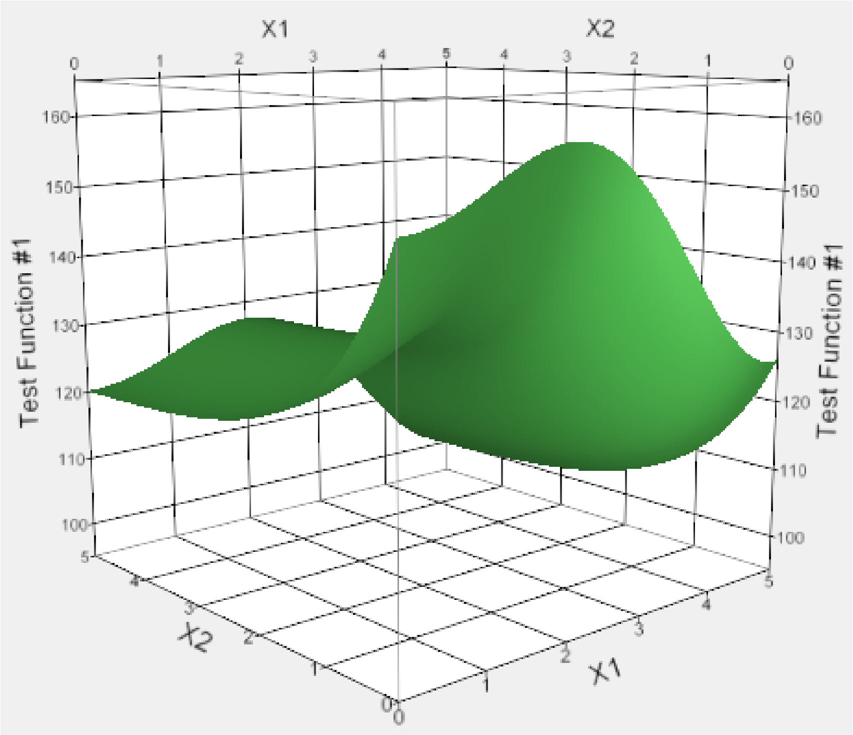
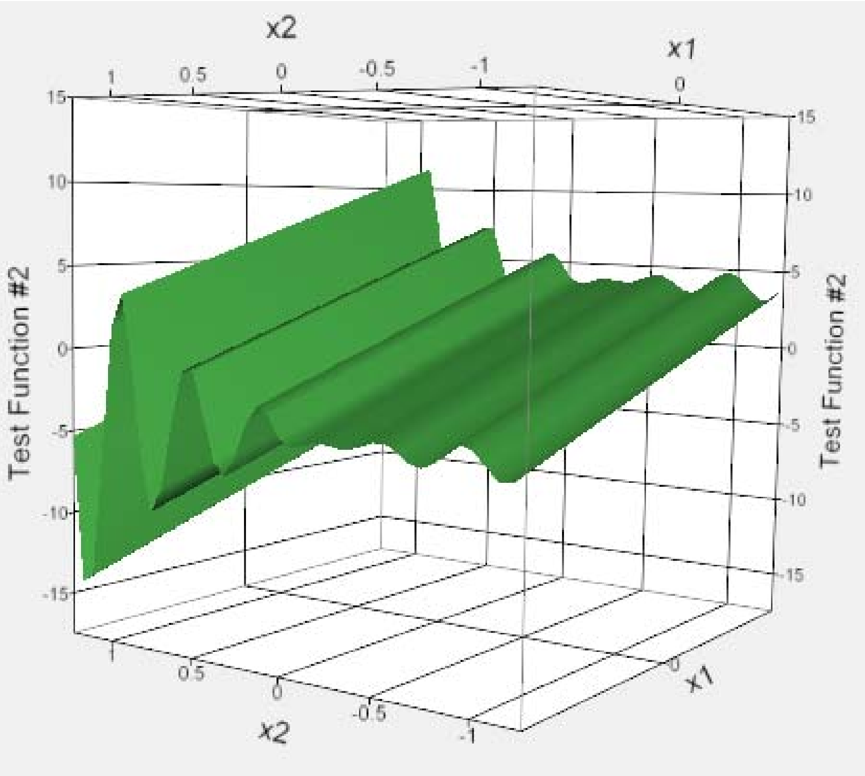


Figure 2 Surface plot of test function 2 (see online version for colours)



3.1.2 Test function 2

The next test function used as a surrogate simulation model is the following:

$$y(x_1,x_2)=5x_1+\left[\sin\square 15x_2\times\left(\sqrt[3]{x_2}\right)\square+\exp\left(2x_2\right)\right)\right] \tag{7}$$

where $-1 \leq x_1$, $x_2 \leq 1$. Figure 2 illustrates that this test function is quite complex. That is, the surface is wavy and irregular. This function allows investigation of the relationship between sample size requirements and response surface complexity.

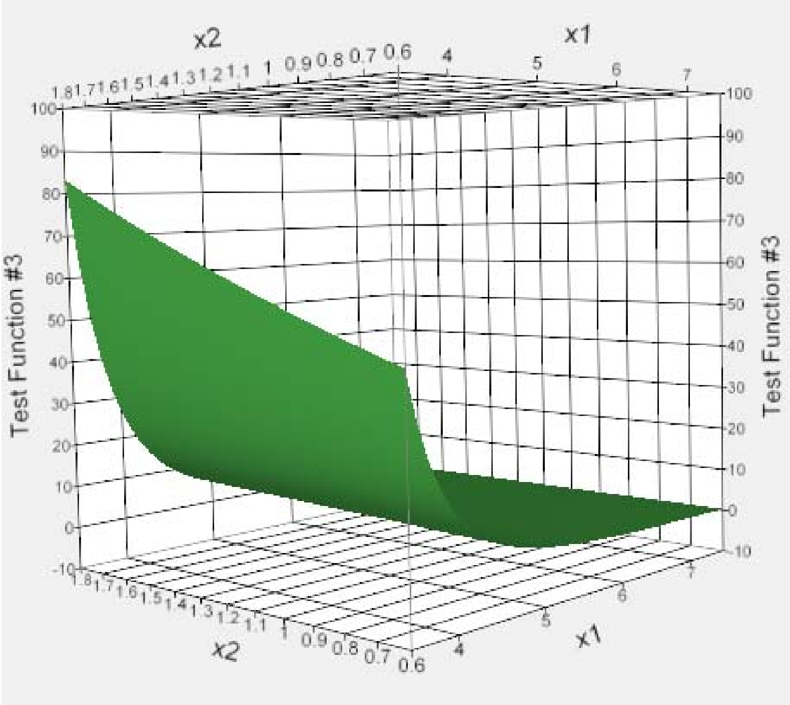
3.1.3 Test function 3

The third test function is also found in Allen et al. (2003) and is designed to act as a surrogate model for a plastic seal design. The approximate analytical function is given as

$$y(x_1,x_2,x_3)=\left(105\left[0.58\left(x_2+x_3-0.85\right)+3.0\right]^3\right)\times\left(\frac{\sin\left[\dfrac{1.5x_3}{x_1-2.0}\right]}{\left(x_1-2.0\right)^3}\right) \tag{8}$$

where x_1, x_2, and x_3 represent input parameter dimensions on the plastic seal. The bounds for the parameters are (in millimetres): $4 \le x_1 \le 7$, $0.7 \le x_2 \le 1.7$, and $0.055 \le x_3 \le 0.500$. A surface plot of test function 3 is shown in Figure 3 for variables x_1 and x_2 at a fixed value of $x_3 = 0.2225$.

Figure 3 Surface plot of test function 3 with $x_3 = 0.2225$ (see online version for colours)



3.1.4 Test function 4

Our final test function was first published in Morris et al. (1993) and subsequently used for comparing meta-models in Allen et al. (2003). The function is

$$y\left(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\right) = \frac{2\pi x_3\left(x_4 - x_6\right)}{\ln\left(\frac{x_2}{x_1}\right)\left[1 + \frac{2x_7 x_3}{\ln\left(\frac{x_2}{x_1}\right)x_1{}^2 x_8} + \frac{x_3}{x_5}\right]} \tag{9}$$

where y predicts water flow – in cubic meters per year – as a function of eight design dimensions. As in Allen et al. (2003), we only vary x_1, x_4, x_6, and x_7 and set the other four variables at their midpoint of the specified ranges from the experiment demonstrated in Morris et al. (1993). The ranges and fixed values for each of the variables are presented in Table 1.

Table 1 The ranges and fixed values for the experimental and fixed variables in test function 4

Experimental variables			*Fixed variables*	
Variable	*Low*	*High*	*Variable*	*Fixed value*
x_1	0.05	0.15	x_2	25,050
x_4	990	1,110	x_3	89,335
x_6	700	820	x_5	89.6
x_7	1,120	1,680	x_8	9,855

3.2 Empirical comparison

The results are presented in the form of box plots. Figure 4 to Figure 7 displays the RMSE values for each of the four space-filling designs [maximin LHD, uniform (U), SP, and GP IMSE] fit to the responses using the GP model with the GCF for sample sizes 10, 20, and 40, for each of the four test functions, respectively.

Figure 4 RMSE for test function 1 in Section 3.1 (see online version for colours)

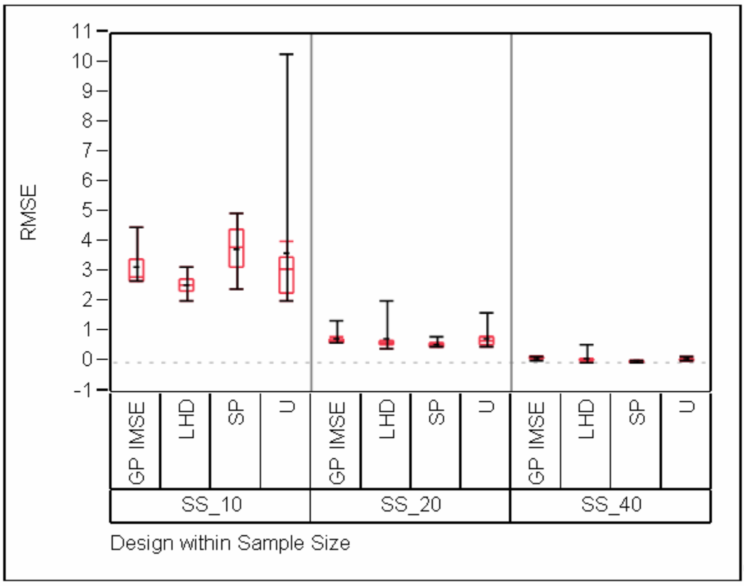


Figure 5 RMSE for test function 2 in Section 3.1 (see online version for colours)

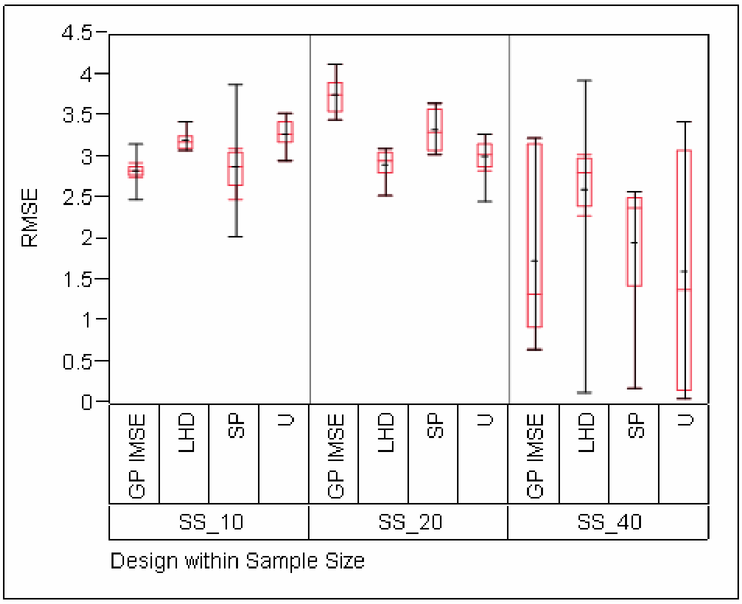


Figure 6 RMSE for test function 3 in Section 3.1 (see online version for colours)

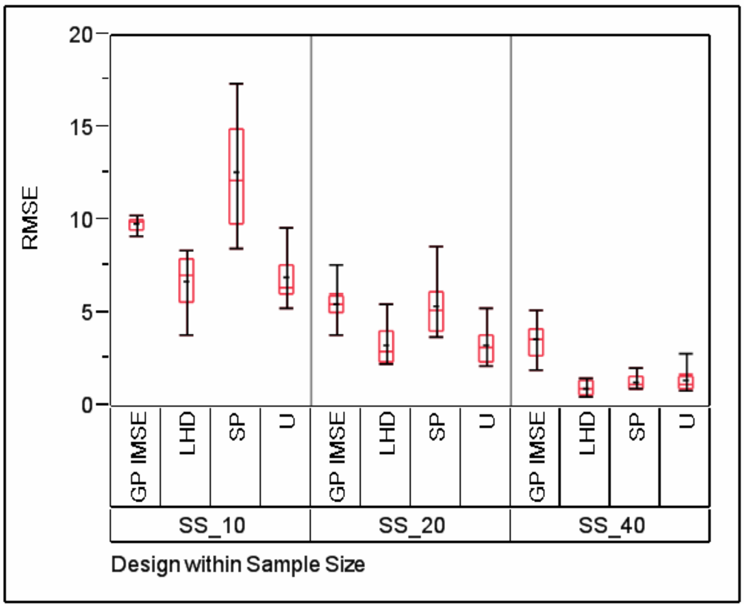
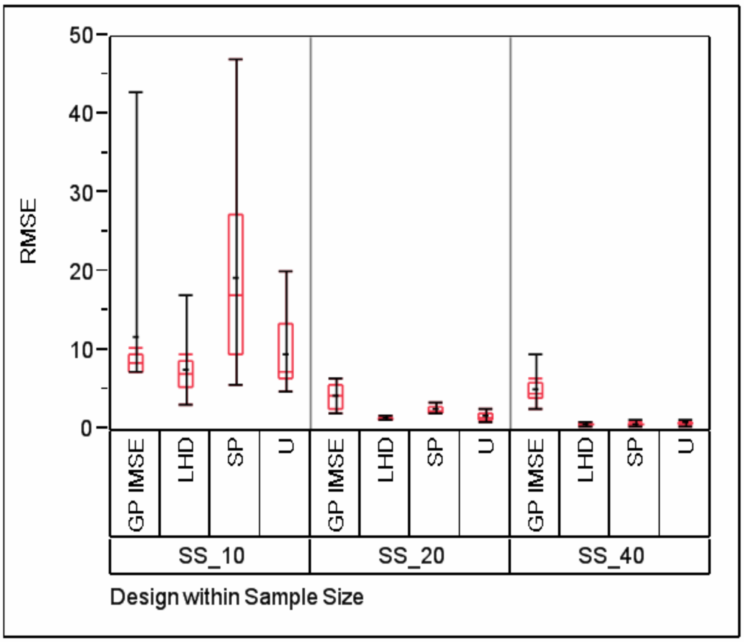


Figure 7 RMSE for test function 4 in Section 3.1 (see online version for colours)



For a given design type, sample size and number of factors the design is not unique. For example, there are many different maximin LHDs for two factors and 20 runs. For each design type, test function and sample size, we generated ten different designs. The box plots show the variability in the RMSE measures over the ten designs.

From the results illustrated in Figure 4 to Figure 7, we observe that:

- no design type outperforms the others in terms of RMSE (Note: low RMSE is preferable)
- increasing the sample size lowers the RMSE
- for a given design type the variability in results for the 10 designs decreases with increasing sample size, except in the case of test function #2
- for test function#2 with small sample sizes all the design types perform poorly – variability across the ten design increases with the largest sample size because a few realisations of each design type are starting to do a better job of picking up the irregular waviness of this complex function.

3.3 ANOVA

In the previous section, we presented the results from the design comparison. We used four different test functions, a range of sample size values, and four different space-filling design types. Combining these data, we performed an ANOVA to see if any of these (main) effects or two factor interactions has an impact on the RMSE. The analysis in Table 2 shows that all of the main effects and two-factor interactions have an impact on the RMSE. The two-factor interaction involving test function and design type leads to the conclusion no design type outperforms the others across all test functions.

Table 2 Effects test from the ANOVA analysis

Source	*DF*	*Sum of squares*	*F ratio*	*Prob > F*
TF	3	259.53939	360.2849	< .0001*
Design	3	18.21060	25.2794	< .0001*
Sample size	1	317.51322	1322.287	< .0001*
TF*Design	9	17.12007	7.9219	< .0001*
TF*Sample size	3	75.50168	104.8092	< .0001*
Design*Sample size	3	16.28325	22.6039	< .0001*

A profiler plot from the ANOVA for the RMSE is shown in Figure 8. We modelled the log of the RMSE, since the fit was better than for the untransformed RMSE response. For the particular setting in Figure 8, one notices that the SP design has the lowest RMSE for test function 1 and a sample size of 20. However, Figure 9 and Figure 10 show that for other sample sizes and test functions other designs are better. Thus, it is apparent that no one design dominates the others. In Figure 9, the uniform design has the lowest average RMSE – this is for the case of test function 2 and a sample size of 30. In Figure 10, the LHD has the lowest RMSE – for the case of test function 3 and a sample size of 25.

Figure 8 Profiler from the ANOVA on RMSE with sphere packing design best (see online version for colours)

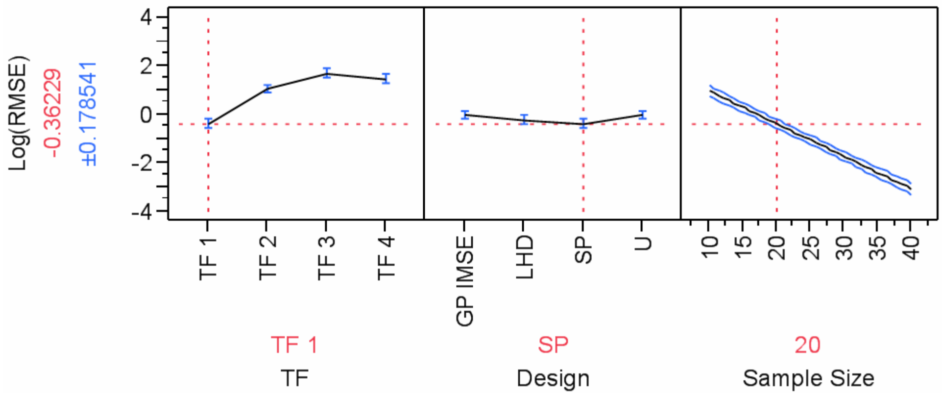


Figure 9 Profiler from the ANOVA on RMSE with uniform design best (see online version for colours)

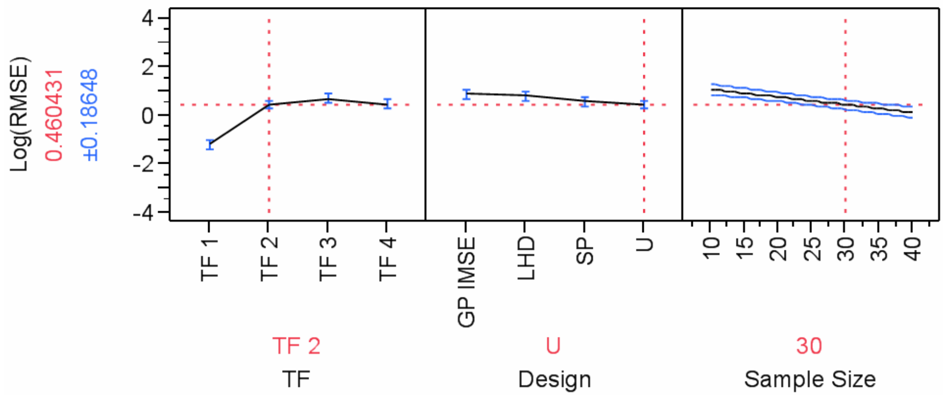
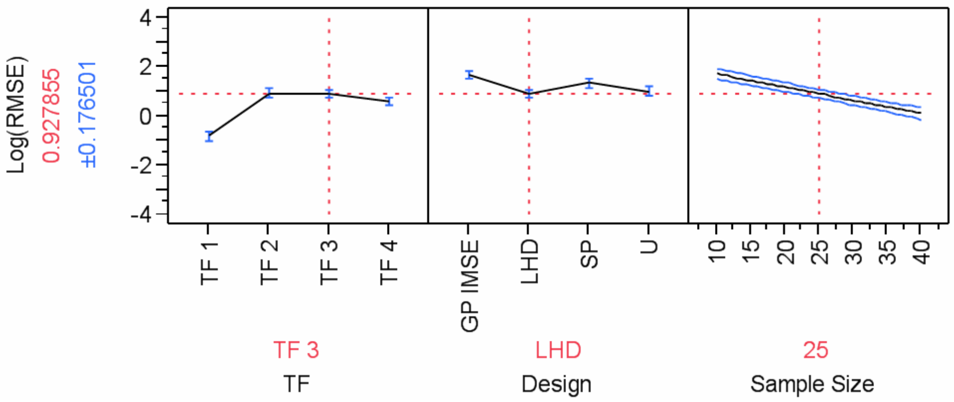


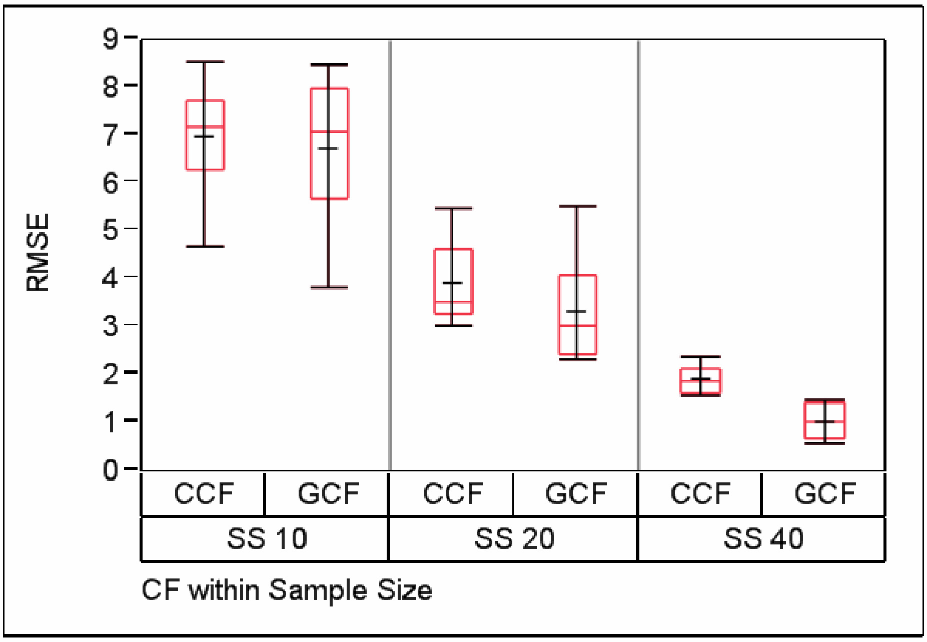
Figure 10 Profiler from the ANOVA on RMSE with Latin hypercube design best (see online version for colours)



4 An empirical comparison of GASP CFs

This section compares the prediction quality of an experimental design fit by the GASP model using two different CFs. Determining which CF to use is a choice that must be made by the researcher. The use of the GCF allows all of the points in the design space to effect one another, while the use of the CCF will allow for zero correlation between two design points that are far away from each other in terms of distance. We compare the two CFs to see if either CF produced GASP model fits with better accuracy and precision as measured by RMSE. Here, we only present the results using test function 3 using the LHD. Figure 11 displays box plots generated from ten LHDs that were used to generate the GASP model fits with either the GCF or the CCF.

Figure 11 Box plot of RMSE for ten LHDs fit using the GASP model with two different correlation functions (see online version for colours)



The box plot illustrated in Figure 11 indicates that the GCF has a slightly lower RMSE average than the CCF. An ANOVA was performed to test for significant main effect (sample size and CF) and also the two factor interaction between sample size and CF type. The results from the ANOVA are presented in Table 3. Sample size and the sample size by CF interaction are significant at an alpha 0.05 level. The CF however is not significant as a main effect. While the GCF is better (in terms of RMSE) for larger sample sizes one might still consider the CCF as an option in order to exploit its faster model fitting ability.

Table 3 ANOVA Results from the first order model with interactions

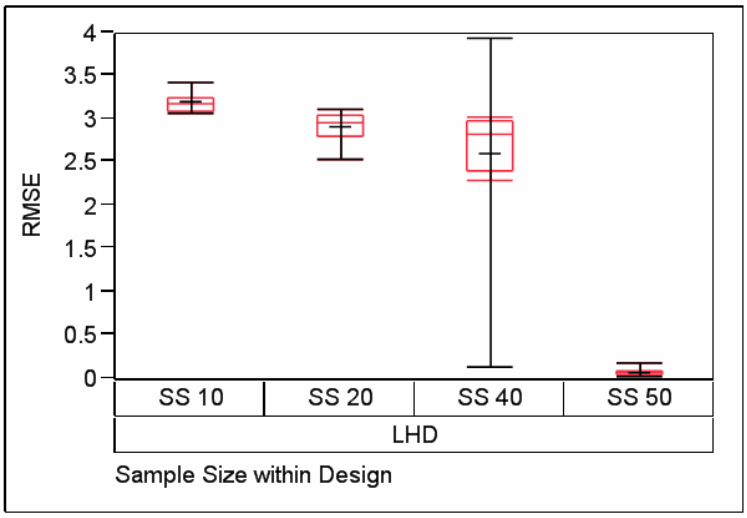
Source	*DF*	*Sum of squares*	*F ratio*	*Prob > F*
CF	1	0.12	2.03	0.1594
Sample size	1	25.32	416.47	< .0001*
CF*Sample size	1	1.00	16.51	0.0002*

5 Sample size study

The previous section demonstrated that empirical comparisons do not favour one design type over another. However, it does appear from Figure 4 to Figure 7 that sample size has an impact on the prediction performance of a design.

We described earlier the $N = 10\,p$ rule of thumb for sample size. For test functions 1 and 4 it would appear that this rule is adequate. However, for test function 2 and 3, the rule of thumb is not adequate. The largest sample size used in the previous two sections was 40, but now we demonstrate the impact of increasing the sample size to 50 for the case of test function 2. Figure 12 illustrates box plots of RMSE for the test function in equation (5) generated by 10 LHDs for each of four different sample sizes (10, 20, 40, and 50) which correspond to ($N = 5\,p$, $N = 10\,p$, $N = 20\,p$ and $N = 40\,p$), respectively. We use only the LHD case here to illustrate our point.

Figure 12 RMSE box plots for ten LHDs over a range of sample sizes for test function 2 (see online version for colours)



With an increase in sample size from 40 to 50, there is a remarkable drop in the RMSE of the fitted GASP model to the LHD. This is due to the complexity of the response surface. A sample size of 40 was still not adequate to capture the entire surface.

Our research indicates that when the surface is complex, the sample size requirements are much greater than when the response surface is relatively flat and smooth. The problem with sample size recommendations it that is not often known what type of surface will result from the simulation model until after the design has been run. We find that a useful indicator of accuracy and precision in the fitted GASP model is the jackknife prediction plot.

Usually, jackknife predictions are created by the leave-one-out method. That is, a single data point is left out, the mathematical model of interest (in this case the GASP model) is fit and then the left out point is predicted using that model. This is done for each data point. The jackknife plot is created by plotting the actual observed response vs. the predicted response using the mathematical model fit without that data point. The jackknife plots in Figure 13 and Figure 14 are generated in a slightly different way. Because GASP model fits are often time consuming, we modify the leave-one-out method is created by removing the row and column from the correlation matrix corresponding to the data point without re-estimating the parameters. Then we use the formula in equation (3) to calculate the predicted response for the removed point.

Figure 13 Jackknife plot of a LHD with 20 runs fit with the GASP model

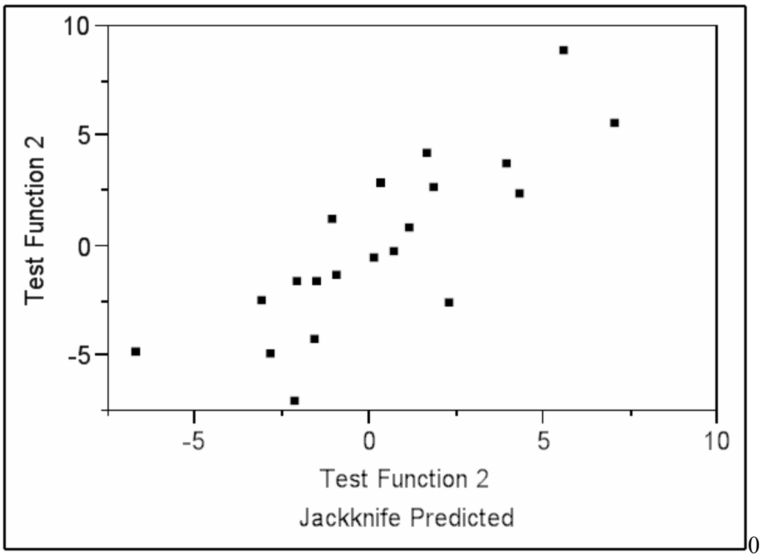
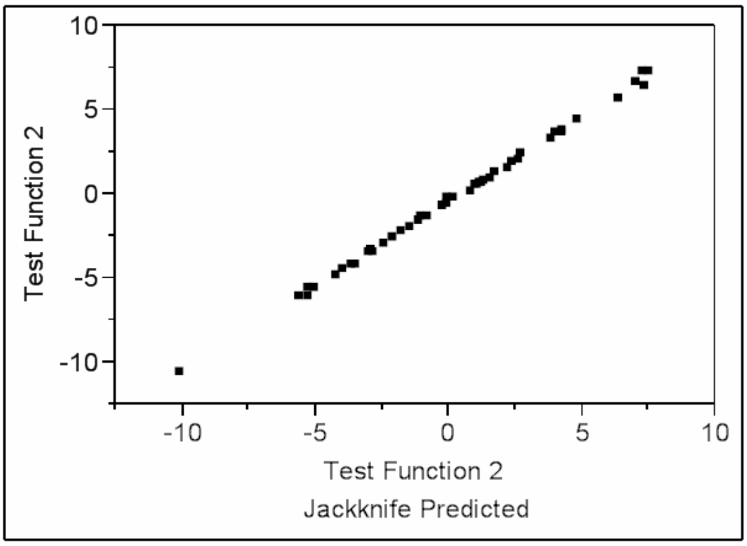


Figure 14 Jackknife plot of a LHD with a sample size of 50 fit to test function 2 using the GASP model



A jackknife plot with points on the x = y line indicates a good fit. If the jackknife prediction plot does not follow a 45 degree line, there is a strong indication that more runs are needed. As an example, Figure 13 shows the jackknife plot for a LHD with a sample size of 20 fit with the GASP model (using test function 2). This plot indicates a poor fit. Figure 13 can be compared with Figure 14 which illustrates the jackknife plot for a LHD with a sample size of 50 fit with the GASP model (using test function 2). This plot indicates an excellent fit. This method suggests that the idea of sequential designs is very powerful. However, more work is needed to decide how and where in the design space to add additional runs.

6 Conclusions

We have demonstrated that no one space-filling design approach outperforms the others in terms their accuracy in predicting responses at previously unsampled locations. We also found that while the GASP model is substantially easier to fit using CCF than with the GCF, its accuracy is not quite as good as the GCF for the test functions illustrated in this paper especially for larger sample sizes. The best way to improve prediction accuracy is to increase the number of simulation runs. Unfortunately, the number of runs necessary to adequately model any function is heavily dependent upon the complexity of its response surface. This suggests that the efficient augmentation of space filling designs is an important area for further research. We have shown that the jackknife plot is a good qualitative indicator of how well the GASP model is working but further work is necessary to fully quantify this useful diagnostic.

References

Allen, T.T., Bernshteyn, M.A. and Kabiri-Bamoradian, K. (2003) 'Constructing meta-models for computer experiments', *Journal of Quality Technology*, Vol. 35, No. 3, pp.264–274.

Ankenman, B., Nelson, B.L. and Staum, J. (2008) 'Stochastic Kriging for simulation metamodeling', *Proceedings of the 2008 Winter Simulation Conference*, pp.362–370.

Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H. and Tu, J. (2007) 'A framework for validation of computer models', *Technometrics*, Vol. 49, No. 2, pp.138–154.

Currin, C., Mitchell, T.J, Morris, M.D. and Ylvisaker, D. (1991) 'Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments', *Journal of the American Statistical Association*, Vol. 86, pp.953–963.

Fang, K.T. (1980) 'The uniform design: application of number-theoretic methods in experimental design', *Acta Math. Appl. Sinica.*, Vol. 3, pp.363–372.

Hussain, M.F., Barton, R.R. and Joshi, S.B. (2002) 'Metamodeling: radial basis functions, versus polynomials', *European Journal of Operational Research*, Vol. 138, pp.142–154.

Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1990) 'Minimax and maxmin distance design', *Journal of Statistical Planning and Inference*, Vol. 26, pp.131–148.

Johnson, R.T., Montgomery, D.C., Jones, B. and Parker, P.A. (2010) *Comparing Computer Experiments for the Gaussian Process Model Using Integrated Prediction Variance*, Submitted for publication.

Jones, B. and Johnson, R.T. (2009) 'The design and analysis of the Gaussian process model', *Quality and Reliability Engineering International*, Vol. 25, pp.515–524.

Kennedy, M.C. and O'Hagan, A. (2001) 'Bayesian calibration of computer models (with discussion)', *Journal of the Royal Statistical Society B*, Vol. 63, pp.425–464.

Kleijnen, J.P.C. and Beers, W.C.M. (2005) 'Robustness of Kriging when interpolating in random simulation with heterogeneous variances', *European Journal of Operational Research*, Vol. 165, pp.826–834.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Ye, K.Q. (2006) 'Variable selection for Gaussian process models in computer experiments', *Technometrics*, Vol. 48, pp.478–490.

Loeppky, J.L., Sacks, J. and Welch, W. (2008) 'Choosing the sample size of a computer experiment: a practical guide', Technical Report Number 170, National Institute of Statistical Sciences.

McKay, N.D., Conover, W.J. and Beckman, R.J. (1979) 'A comparison of three methods for selecting values of input variables in the analysis of output from a computer code', *Technometrics*, Vol. 21, pp.239–245.

Morris, M.D. and Mitchell, T.J. (1995) 'Exploratory designs for computational experiments', *Journal of Statistical Planning and Inference*, Vol. 43, pp.381–402.

Morris, M.D., Mitchell, T.J. and Ylvisaker, D. (1993) 'Bayesian design and analysis of computer experiments: use of derivatives in surface prediction', *Technometrics*, Vol. 35, pp.243–255.

Sacks, J., Schiller, S.B. and Welch, W.J. (1989a) 'Designs for computer experiments', *Technometrics*, Vol. 31, No. 1, pp.41–47.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989b) 'Design and analysis of computer experiments', *Statistical Science*, Vol. 4, No. 4, pp.409–423.

Santner, T.J., Williams, B.J. and Notz, W.I. (2003) *The Design and Analysis of Computer Experiments*, Springer Series in Statistics, Springer-Verlag, New York.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J. and Morris, M.D. (1992) 'Screening, predicting, and computer experiments', *Technometrics*, Vol. 34, No. 1, pp.15–25.