Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

DOES MORE UNIFORMLY DISTRIBUTED SAMPLING GENERALLY LEAD TO MORE ACCURATE PREDICTION IN COMPUTER EXPERIMENTS?

Longjun Liu

Gunderson Inc.
4350 NW Front Avenue
Portland, OR 97210, U.S.A.

Wayne Wakeland

Systems Science Ph.D. Program
Portland State University
Portland, OR 97225, U.S.A.

ABSTRACT

Sampling uniformity is one of the central issues for computer experiments or metamodeling. Is it generally true that more uniformly distributed sampling leads to more accurate prediction? A study was conducted to compare four designs for computer experiments, based on simulation tests and statistical analysis. Maximin Latin hypercube design (LHMm) nearly always generated more uniform sampling in two- and three- dimensional cases than does random sampling (Rd), Latin hypercube design (LHD), or Minimized centered L_2 discrepancy Latin hypercube design (LHCL2). But often there was no significant difference among the means of the prediction errors by employing LHMm versus the other designs. Occasionally, even the opposite was seen. More uniform sampling did not generally lead to more accurate prediction unless sampling included extremely nonuniform cases, especially when the sample size was relatively small.

1 INTRODUCTION

It is generally agreed that, by intuition, sampling for computer experiments should be uniformly distributed in the design region (e.g. Koehler and Owen, 1996; Santner, et al, 2003). The more uniformly distributed the sampling, the better the experimental design. However, different or even opposite results were observed in a study to compare eighteen designs for computer experiments, based on simulation tests and statistical analyses via ANOVA. The design types included random sampling (Rd), Latin hypercube design (LHD), Maximin Latin hypercube design (LHMm) based on the Φ_q criterion, Minimized centered L_2 discrepancy Latin hypercube design (LHCL2), etc. The results showed that more often there was no significant difference between the approximation error means resulted from applying different designs for sampling. Where there were significant differences, LHD and LHCL2 often outperformed LHMm. However, LHMm almost always generated more uniformly distributed samples in 2D and 3D designs than other methods. For more details of the study, please see Liu (2004) and Liu (2005).

It seems that the results are contrary to the general belief that better designs should yield more uniformly distributed sampling. Since uniformity has been taken as one of the fundamental issues for experimental design in computer experiments, and thus has significant impact on research and application, more tests were conducted for further investigation and are described below, in order to address the central research question that is the title.

2 RESEARCH DESIGN

NOMENCLATURE

d_{ij}	the Euclidean distance between points i and j
n (N)	the number of design variables
m (M)	the number of runs or sample size
q	a parameter to be determined
X_{ij}	the j^{th} component of the i^{th} sampled point
U_{ij}	the j^{th} element of i^{th} independent random variable U [0, 1] (uniform distribution in [0,1]), independent of the π_{ij}
Φ_q	a design criterion
$\boldsymbol{\pi}_{ij}$	the j^{th} element of i^{th} independent uniform random permutations of the integers 1 through n

This section provides the test design, the measures for uniformity, the four design (sampling) types, and the twenty test functions.

2.1 Test scheme

There were two test groups. In the first group, the four designs (LHMm, Rd, LHD, and LHCL2) were employed and compared simultaneously. Twenty test functions were approximated using each specific sampling. There were five levels of the sample size. The second group focused on one function, one level of sample size, and pairwise comparison each time.

In the first group, for the same sample size, the four designs were employed for sampling as four different "treatments." For each "treatment" that is a particular sampling, there were 20 "observations" that corresponded to 20 functions. For each function, a kriging model was built as the approximation model to fit the sampled points and the responses. To test the prediction accuracy, 10,000 validation points were generated by LHD within the domain [-30,30; -30,30]. The same domain was also used for sampling to build the models. The domain is specified in the literature for some of the test functions and was applied to most of the functions in this study. At each point, the values of the test function and of the approximation model were compared to find the difference. From all the validation points, a relative error called relative root mean square error (RRMSE) was found by the following formula:

$$RRMSE = \sqrt{\frac{1}{10000}\sum_{i=1}^{10000}\left(\frac{Ft(i)-Fa(i)}{|Ft(i)|+\varepsilon}\right)^2} \quad (1)$$

Ft: function response; Fa: approximation response; $\varepsilon = 10^{-4}$: to guard against possible Ft = 0.

And, the maximum relative error was found by the following formula:

$$Max-rel-error = Max\left(\frac{|Ft(i)-Fa(i)|}{|Ft(i)|+\varepsilon}\right) \quad i = 1,2,... \quad (2)$$

With exactly the same sampling, nineteen more RRMSE observations and nineteen more Max-rel-error observations were generated by the same procedure for approximating nineteen more test functions. This process was repeated for the other three design methods or "treatments." In all, there were twenty observations per each specific sampling to compare RRMSE and Max-rel-error respectively. One-way ANOVA and error distribution were used to compare the designs by comparing the means of RRMSE and Max-rel-error respectively. The whole process outlined above was repeated for another sample size.

The sample sizes tested were 8, 10, 12, 14, and 16 to allow for visual comparison of the distribution uniformity. With too few or too many points, it was difficult to tell if the distributions were uniform by visual inspection. The relative error was employed because it was likely to be more meaningful for prediction. Further, since twenty test functions were used, using RMSE might inflate within group variance, making it difficult to detect small differences in the error means resulted from different "treatments."

In the second group, many pairwise comparison tests between LHMm and one of the other three designs were conducted for one test function each time. After sampling, a kriging model was built and RMSE was calculated based on 10,000 validation points. The same process was repeated four more times for the same design type and the same sample size. Then, the whole process was repeated for the other design type. Next, ANOVA was conducted with each treatment having five observations of RMSE. Finally, the whole process was repeated on another pair of designs and another function.

2.2 Measure for Uniformity

There seems to be no generally agreed upon definition for sampling uniformity. Many criteria have been proposed for reaching space filling or uniformly distributed designs, e.g. maximin, Fi (Φ_q) criterion, several discrepancy criteria, etc. We conducted many 2D and 3D tests with LHMm. Those tests showed that LHMm resulted in more uniformly distributed samplings than other Latin hypercube designs. Thus, for the purpose of comparing the sampling uniformity of different designs, the Fi criterion was used as an indirect measure for uniformity, as follows. The measure is called relative uniformity (UNIF) shown by Eq. 3. The larger the UNIF, the larger the sum of the distances between the points, which results in a more uniform distribution within a stratified sampling domain such as LHD. It also seemed to be a good measure for Rd as well. More study is needed for the possibly better definition of (relative) uniformity.

$$UNIF(i) = FIinv(i)/MaxFIinv \quad (3)$$

UNIF: relative uniformity; FIinv (i): i^{th} inversed Fi value corresponding to i^{th} design; MaxFIinv: the maximum inversed Fi value among those of the designs under comparison.

2.3 Four Types and Twenty Test Functions

Four design types

1. Random design (Rd)
2. Latin hypercube design (LHD) (McKay, et al, 1979): The j^{th} component of the i^{th} sampled point is

$$X_{ij} = \frac{\pi_{ij} - U_{ij}}{m}. \quad (4)$$

3. Maximin Latin hypercube design (LHMm)

Criterion: Φ_q criterion (slightly modified in the form from that of Morris and Mitchell, 1995); q: Pilot tests show better values: 1 for most 2D cases, 75 or 45 for 5D cases, 75 for most 10D cases; testing with q = 2 or other values were also conducted with similar or a little worse performance.

$$\Phi_q = \left[\sum_{i=1}^{m}\sum_{j=i+1}^{n} d_{ij}^{-q}\right]^{1/q} \quad (5)$$

4. Minimized CL2 Latin hypercube design (LHCL2). Criterion: Centered L_2 discrepancy CL_2 (Hickernell,1998): Eq.6.

$$[CL_2(P_m)]^2 = (\tfrac{13}{12})^2 - \tfrac{2}{n}\sum_{k=1}^{m}\prod_{j=1}^{n}[1+\tfrac{1}{2}|x_{kj}-0.5| - \tfrac{1}{2}|x_{kj}-0.5|^2] + \tfrac{1}{n^2}\sum_{k=1}^{m}\sum_{j=1}^{m}\prod_{i=1}^{n}[1+\tfrac{1}{2}|x_{ki}-0.5| + \tfrac{1}{2}|x_{ji}-0.5| - \tfrac{1}{2}|x_{ki}-x_{ji}|]. \quad (6)$$

Most of the functions tested are popular functions for testing global optimization methodologies. Many of them have high nonlinearity or multi-modes, whereas others are quite simple or smooth. The details of the functions are presented in the Appendix.

3 RESULTS OF THE SIMULATION TESTS, ANOVA, AND ERROR RANGE ANALYSIS

The test results for the 2D and 3D cases are listed here.

3.1 Two-dimension Cases

1. Comparing the four designs simultaneously: The sample sizes tested were: 8, 10, 12, 14, and 16. For validation, only one sampling group of 10,000 points was generated which was used by all the tests. The sampling plots and the ANOVA plots together with the p-values are shown in Figures 1-5.

2. Pairwise comparison between LHMm and Rd: The pairwise comparison between LHMm and one of the other three designs showed that again, in most cases, there was no significant difference in the RMSE means resulting from using LHMm versus another design. To show less uniformly distributed sampling can sometimes result in higher prediction accuracy, one pairwise comparison between LHMm and Rd is provided in Fig. 6. The test function was AC. The sample size was 10. In Rd sampling, the points cluster around one corner one time, around another corner at another time. In spite of this, “bad” Rd sampling still outperformed “better” LHMm sampling.

3.2 Three-dimension Cases

The sample sizes tested were 12, 15, 18, 21, and 24. For validation, only one sampling group of 10,000 points was generated which was used by all the tests. The sampling plots, shown as the projections onto XY, XZ, ZY planes, and the ANOVA plots together with the p-values are shown in Fig. 7 and many more are available upon request.

It has been shown that LHMm generated more uniformly distributed sampling than did RD, LHD, or LHCL2. In most cases, however, there was no statistically significant difference in the means of RRMSE, Max-rel-error, or RMSE, regardless of the critical p-value selected as 0.01 or 0.05. In some cases, most of the points were along one straight line or cluster in a corner, but still there was no significant difference in the RRMSE means. Occasionally, LHMm was significantly worse than Rd despite its being more uniformly distributed. Note that in the first group, twenty different functions, instead of only one function, were approximated using exactly the same specific sampling.

In terms of the variance or the range of the data, it seemed that uniformity did not have large impact either, but sample size did matter. As the sample size increased, the range usually decreased, as shown in Figures 1-5. Note that the figures have different axis limits for the errors. It is noticed in Fig.5 when the size was relatively large, the least-uniform sampling did resulted in relatively larger range and higher level of errors.

4 SOME EXTREME CASES WITH A VERY NONUNIFORM SAMPLING: PROPORTIONAL SAMPLING (PS)

Santner, Williams, and Notz (2003) described a sampling along a straight line, resulting in higher accuracy at the line but poor prediction elsewhere. It is possible for Rd or LHD to end up with some very nonuniform designs like this one. We refer to such a sampling as "proportional sampling" (PS) — one in which all sample points uniformly distribute along the diagonal line of the space. Some comparison results among PS, LHD, LHMm, and LHCL2 are shown in Figures 8-12. The test scheme was the same as mentioned in Section 3, but with PS replacing Rd.

When the sample size was relatively small, there were not apparent differences in the error means and ranges by employing different designs. When the size increased, however, while other designs produced more widely spread samplings and decreasing levels of errors, PS stayed along the diagonal line and kept the error level almost unchanged or even higher in many cases. In the latter cases when the error level increased as the sample size increased, the kriging models were misled by "over-fitting." It is shown in ANOVA plots that, when the size was relatively large, the error range by PS was much larger and the error level was higher than those by other design types. The results for even larger sizes (not shown) presented larger differences.

It is clear that very nonuniform designs should be avoided. Since it is possible for Rd and LHD to create such extreme samples, it is necessary to check the resulting design to guard against such extreme cases. The author evaluated the Fi criterion as a possible measure for performing this check. When its inverse FIINV was very low, the sampling was quite nonuniform. The FIINV values for sampling are shown in Table 1.

In most cases, FIINV values by PS were lower than those by other design types. Unfortunately, FIINV became smaller when the sample size became larger. After some experimentation, it was determined that FIINV*$m^{2.5}$ resulted in relatively stable values vs. sample sizes, as is shown in Table 2. It appears that SFIINV = FIINV*$m^{2.5}$ might be a better tentative measure for "absolute" uniformity (vs. UNIF as described in Section 5).

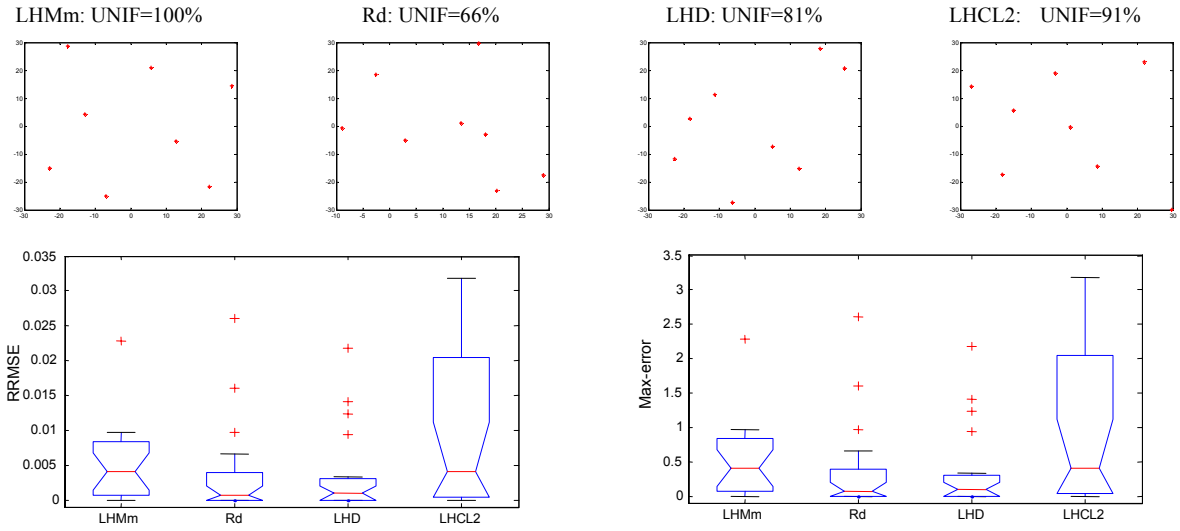


Figure 1: Comparing four designs based on comparing RRMSE, m=8 (m: sample size), p-v = 0.3975

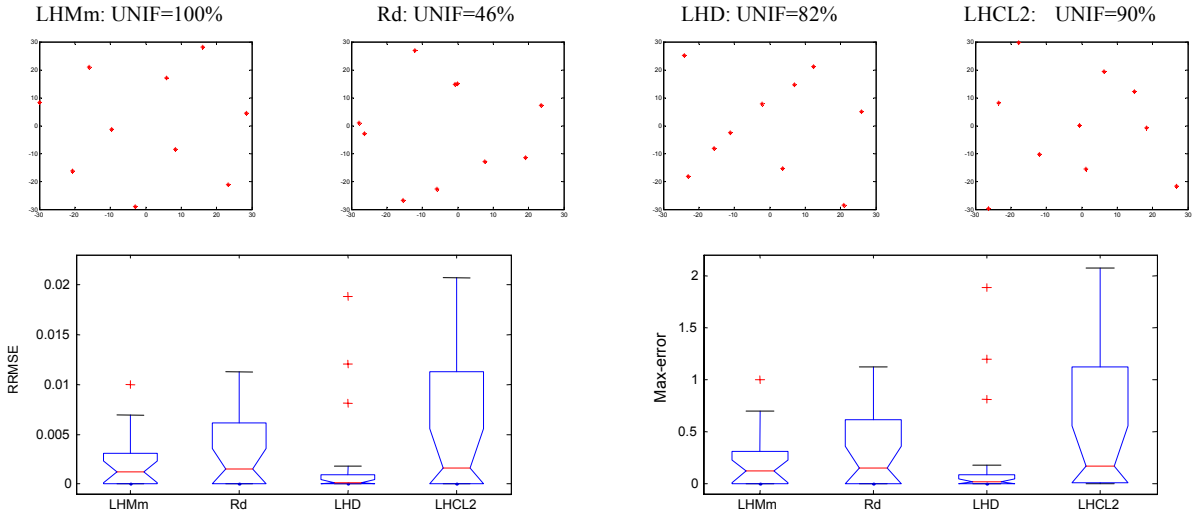


Figure 2: Comparing four designs based on comparing RRMSE, m=10 (m: sample size), p-v = 0.3974

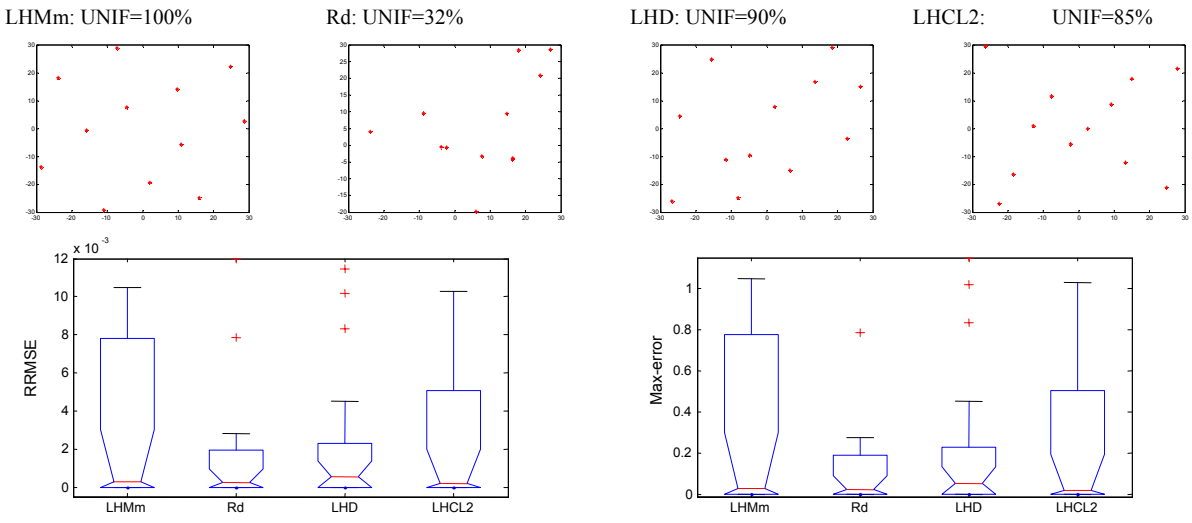


Figure 3: Comparing four designs based on comparing RRMSE, m=12 (m: sample size), p-v = 0.4104

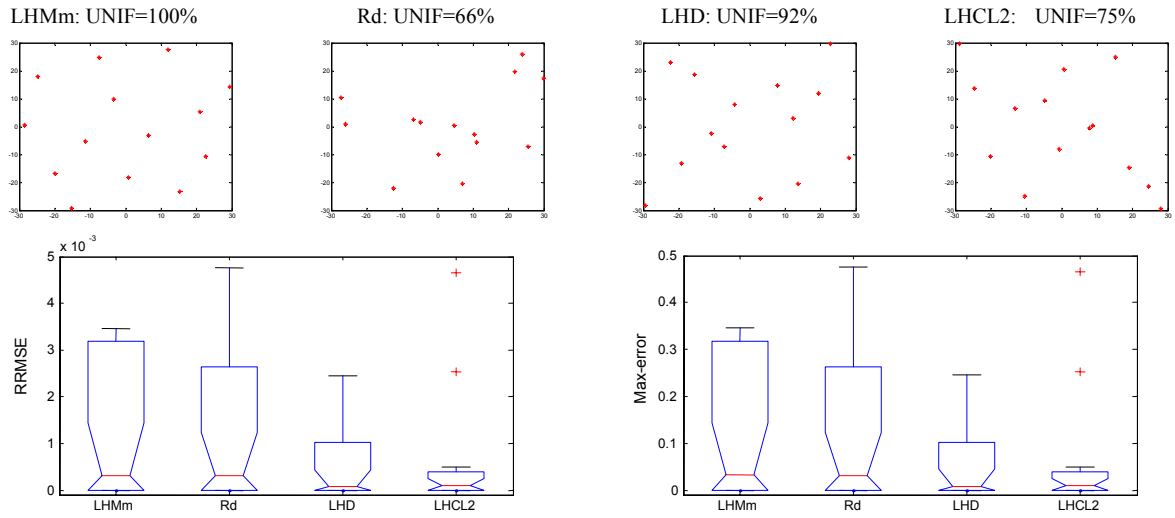


Figure 4: Comparing four designs based on comparing RRMSE, m=14 (m: sample size), p-v = 0.5340

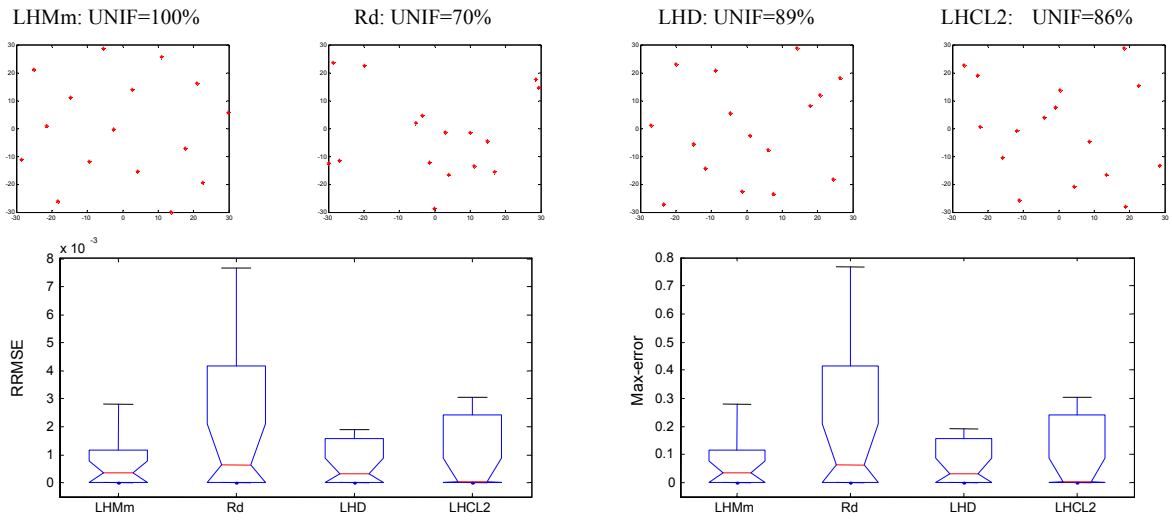
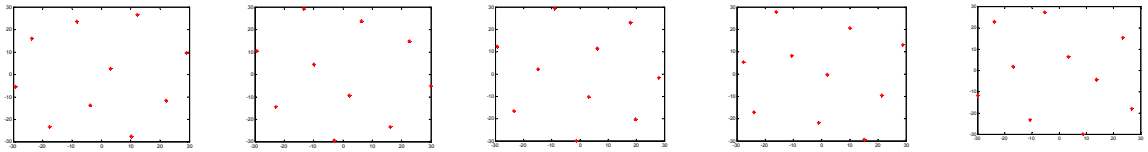
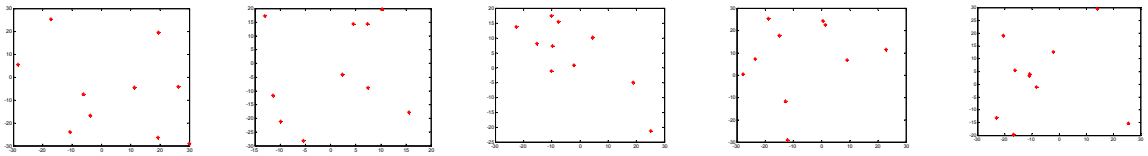


Figure 5: Comparing four designs based on comparing RRMSE, m=16 (m: sample size), p-v = 0.4634

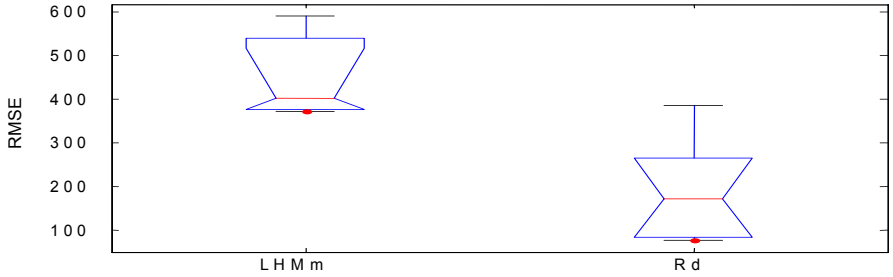
1) LHMm



2) Random sampling (Rd)



3) ANOVA Plot and p-value



P-value = 0.00624.

Figure 6: Comparing LHMm and Rd based on comparing RMSE, m=10 (m: sample size), function= AC

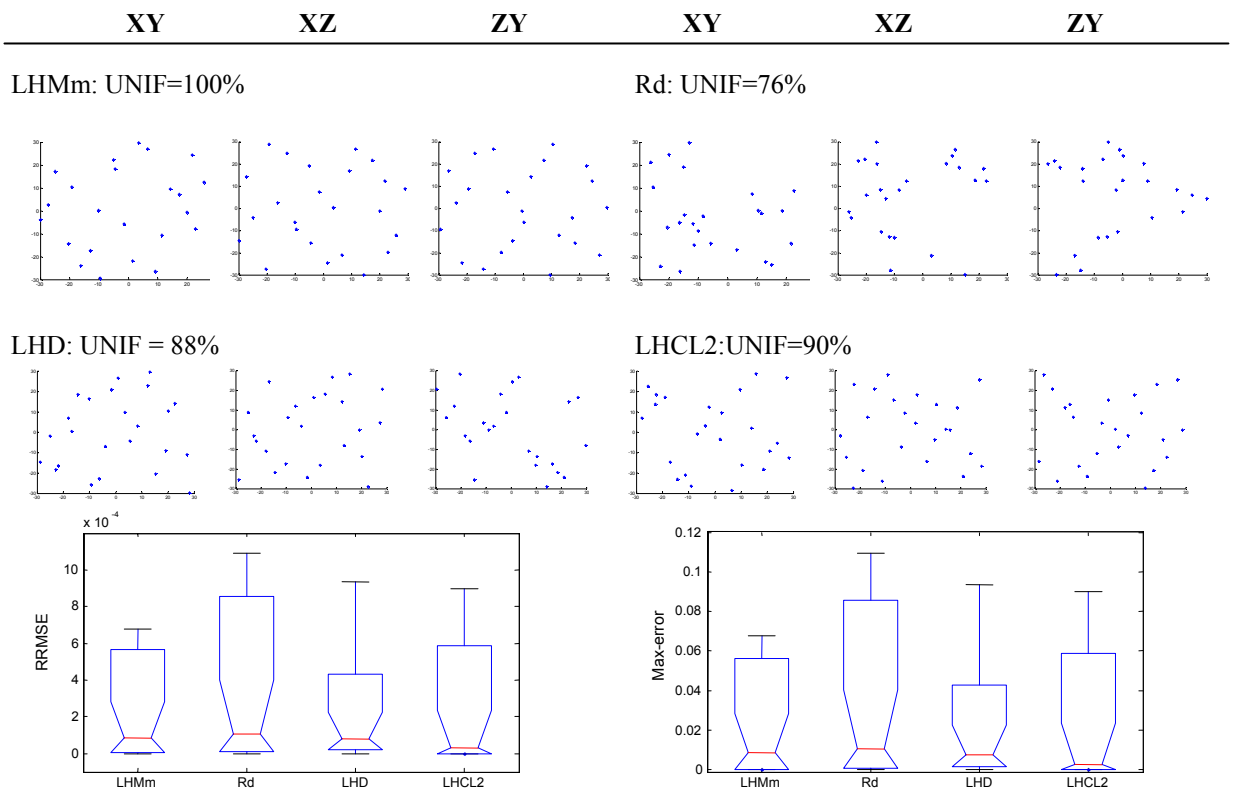


Figure 7: Comparing 4 designs by comparing RRMSE and Max-error, 3D, m =24, p-v = 0.5313

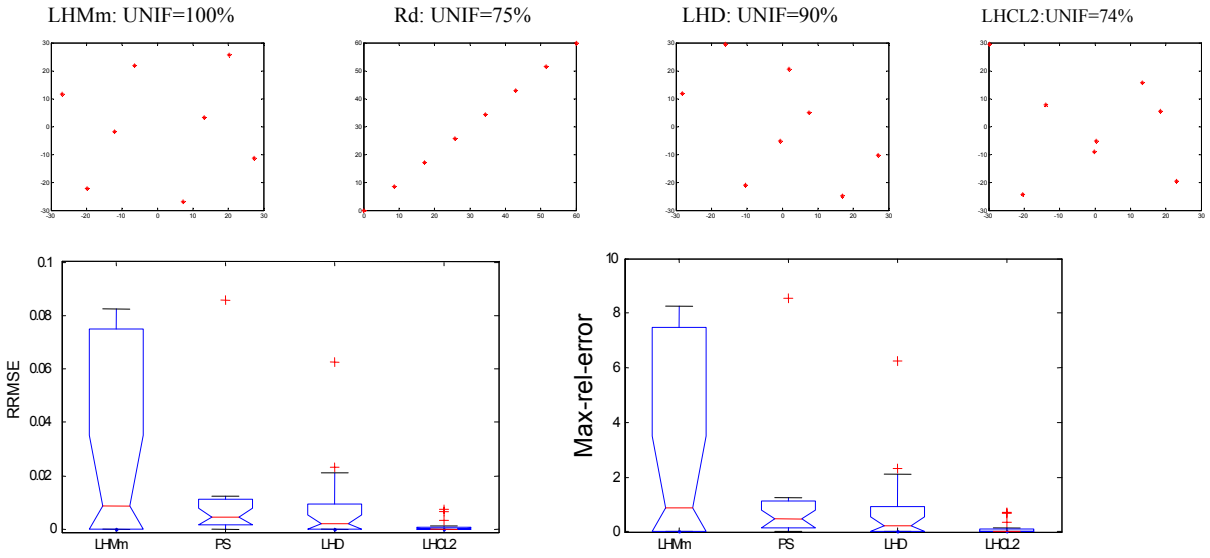


Figure 8: Comparing LHMm, PS, LHD, and LHCL2; 2D, m (sample size) =8, p-v = 0.2664

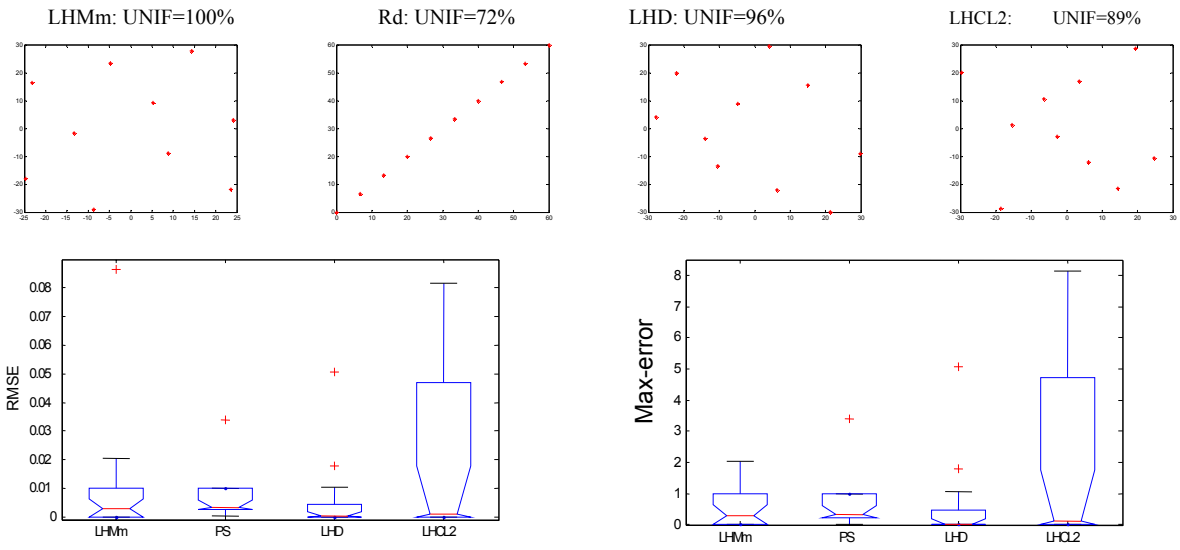


Figure 9: Comparing LHMm, PS, LHD, and LHCL2; 2D, m (sample size) = 10, p-v = 0.4887

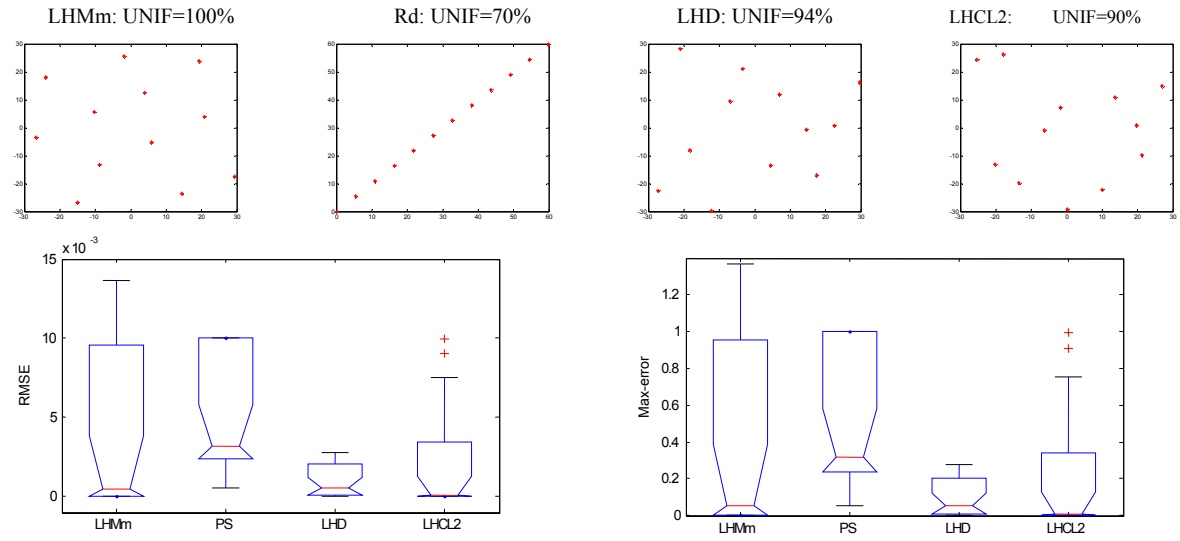


Figure 10: Comparing LHMm, PS, LHD, and LHCL2; 2D, m (sample size) = 12, p-v = 0.5653

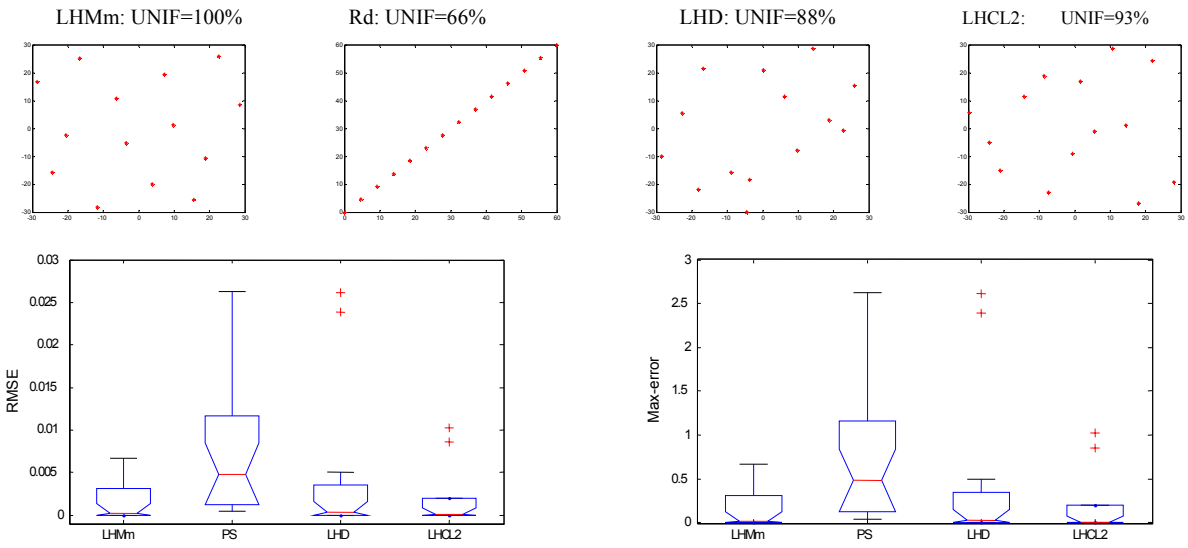


Figure 11: Comparing LHMm, PS, LHD, and LHCL2; 2D, m (sample size) = 14, p-v = 0.7135

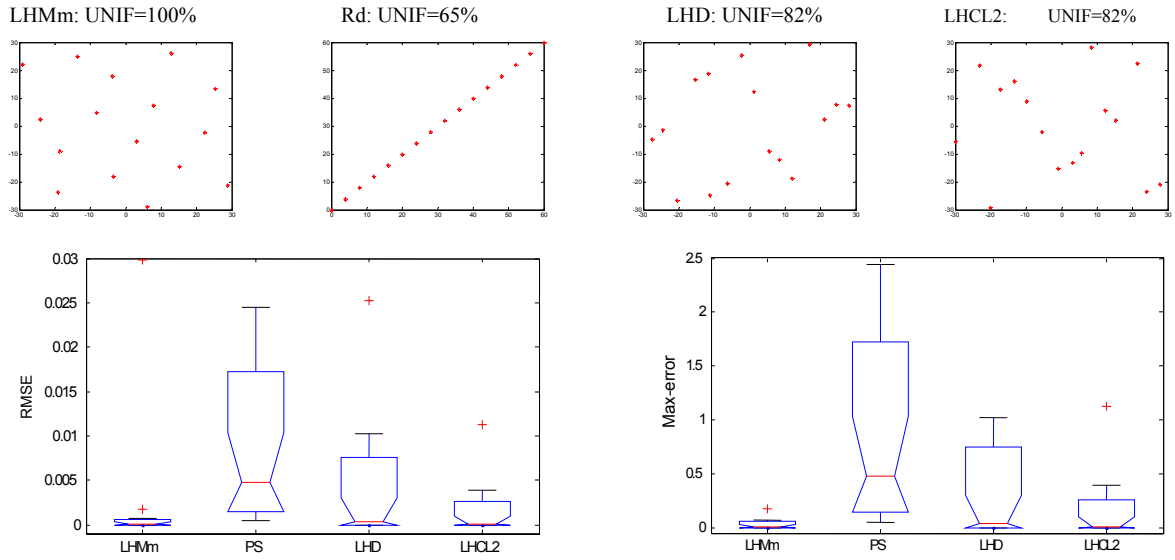


Figure 12: Comparing LHMm, PS, LHD, and LHCL2; 2D, m (sample size) = 14, p-v = 0.6474

Table 1: FIINV values for samplings in Figures 8-12 (m: sample size)

m	**LHMm**	**PS**	**LHD**	**LHCL2**
8	1.1738	0.8820	1.0551	0.8632
10	0.6743	0.4888	0.6485	0.6011
12	0.4348	0.3056	0.4113	0.3934
14	0.3133	0.2071	0.2761	0.2904
16	0.2298	0.1485	0.1878	0.1886

Table 2: SFIINV values for samplings in Figures 8-12 (m: sample size)

m	**LHMm**	**PS**	**LHD**	**LHCL2**
8	212.5	159.7	191.0	156.3
10	213.2	154.6	205.1	190.1
12	216.9	152.4	205.2	196.2
14	229.8	151.9	202.5	213.0
16	235.3	152.1	192.3	193.1

In Table 2, the SFIINV values by PS were much lower than most other values by other design types. When SFIINV is quite low, another LHD can be generated and checked, until a higher SFIINV is obtained. Based on our simulation, the likelihood of obtaining an extreme sampling is very small. We conducted thousands of LHD samplings and did not see a single extreme case that was even close to the PS case. It was much more likely to obtain much better designs than the extreme designs. Much more and deeper study is needed to find a very reliable measure for absolute uniformity check.

5 CONCLUSIONS AND DISCUSSION

In general, more uniformly distributed sampling did not lead to more accurate modeling in computer experiments in most of the 2D and 3D cases shown here, unless the designs included extremely nonuniform cases. This result seems likely to be true for higher dimensions as well. Thus, it may not be appropriate to use uniformity as the sole criterion to compare different experimental designs or as the only goal to be pursued in the search for better designs, although it is true that very nonuniform samplings should be avoided. Much research effort has been devoted to finding "better" designs with more uniformly distributed sampling. It seems that other goals besides uniformity need to be studied. What is more important than uniformity or "better designs," at least when prediction accuracy is important, is to assure that sample size is large enough, as was discussed by Liu[2] (2005). Sample size determination and the development of good sequential design methods may warrant additional attention.

Because of very limited resources, this study was limited to Kriging models, twenty test functions, four designs. Larger design domain may be necessary to make general conclusions. Comparison for uniformity for higher dimensions than 3D has not been done. The interaction between sample size and uniformity needs to be studied. It seems that when the sample size was relatively large, the error range was larger for less uniform samplings. More and deeper study is needed beyond this preliminary investigation. Nevertheless, the research has revealed that uniformity may not be as important for prediction as previously thought. Since sampling uniformity has been taken as a fundamental issue for computer experiments, the results of this research might be useful to users as well as researchers regarding their choice or development of experimental designs.

We close this paper with a comment from a private communication from William Notz that we received on an earlier version of the manuscript in which he provided additional observations and a thoughtful analysis.

"The results you have found agree with what we have observed. Our experience is that any design that is reasonably uniform seems to work well. Only very nonuniform designs (for example, designs that take most of their observations on the boundary) seem to perform poorly.

We believe the reason is that interesting features of response surfaces (locations of maxima, minima, regions where the response surface varies greatly) generated by simulations tend to be "uniformly" spread out over the design region. The variation in the location of these interesting features is larger than subtle variations in the uniformity of designs. Thus, only designs that take observations in a very limited portion of the design space perform poorly."

ACKNOWLEDGMENTS

Our thanks go to Greg Saxton, PE, the Chief Engineer at Gunderson Inc., for his support and encouragement over many years. We are grateful to Professor William Notz who read the earlier manuscript and provided the detailed comment, explanation, and more observations on sampling uniformity as well as sample size effect. We thank the referee for careful critique and suggestions. We also thank Multnomah County Library and Portland State University Library for obtaining many research papers for us.

APPENDIX: TEST FUNCTIONS REFERENCES

Note: x in $f(x)$ is a vector with coordinates $x_1,\ldots,x_n$; $x_i \in [-30,30]$ (unless otherwise specified).

1. Function 1 (AC): Ackley's path function

$$f(x) = -20 * \exp(-0.2 * \sqrt{(1/n)(\sum_{i=1}^{n} x_i^2)}$$

$$- \exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)) + 20 + e$$

2. Function 2 (AX): Axis parallel hyper-ellipsoid function

$$f(x) = \sum_{i=1}^{n} i x_i^2$$

3. Function 3 (DE): De Jong's function 1

$$f(x) = \sum_{i=1}^{n} x_i^2$$

4. Function 4 (RB): Rosenbrock's valley (De Jong's function 2)

$$f(x) = \left[\sum_{i=1}^{n} 100(x_{i+1} - x_i^2)\right]^2 + (1 - x_i)^2$$

5. Function 5 (RY): rotated hyper-ellipsoid function

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{i} x_j^2$$

6. Function 6 (MI): Michalewicz's function

$$f(x) = -\left(\sum_{i=1}^{n} \sin x_i\right) * \left[\frac{\sin(i x_i^2)}{\pi}\right]^{2m}$$

m=10; $0 \le x(i) \le$ pi (π).

7. Function 7 (BR): Branins's rcos function

$$f(x_1,x_2) = (x_2 - bx_1^2 + cx_1 - 6)^2 + 10(1-f)\cos x_1 + 10$$

b=5.1/(4·pi^2), c=5/pi, f=1/(8·pi);

$-5 \le x_1 \le 10$, $0 \le x_2 \le 15$.

8. Function 8 (GD): Goldstein-Price's function

$$f(x_1,x_2) = [1+(x_1+x_2+1)^2(19-14x_1+3x_1^2 - 14x_2+6x_1x_2+3x_2^2)]*[30+(2x_1-3x_2)^2(18 - 32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2)]$$

$-2 \le x(i) \le 2$, i=1:2.

9. Function 9 (SX): Six-hump camel back function

$$f(x_1,x_2) = (4-2.1x_1^2+\frac{x_1^4}{3})x_1^2 + x_1x_2 + (-4+4x_2^2)x_2^2$$

$-3 \le x_1 \le 3$, $-2 \le x_2 \le 2$.

10. Function 10 (PK): Peaks function

$$f(x_1,x_2) = 3(1-x_1)^2\exp(-x_1^2-(x_2+1)^2) - 10(\frac{x_1}{5}-x_1^3-x_2^5)\exp(-x_1^2-x_2^2) - \frac{1}{3}\exp(-(x_1+1)^2-x_2^2)$$

$-3 \le x_1 \le 3$, $-3 \le x_2 \le 3$.

11. Froth function

$f(x_1, x_2) = -13 + x_1 + ((5 - x_2)\, x_2 - 2)\, x_2 - 29 + x_1 + ((x_2 + 1)\, x_2 - 14)\, x_2$;

12. Helix2 function

$$f(x_1,x_2) = 10(\sqrt{x_1^2+x_2^2}-1)$$

13. Rose function: $f(x_1, x_2) = 10(x_2 - x_1^2)$

14. Sing2 function: $f(x_1, x_2) = \sqrt{5}\,(x_1 - x_2)$;

15. Sing3 function: $f(x_1, x_2) = (x_1 - 2x_2)^2$;

16. Sing4 function: $f(x_1, x_2) = \sqrt{10}\,((x_1 - x_2)^2)$;

17. Wood1 function: $f(x_1, x_2) = 10*(x_2 - x_1^2)$;

18. Wood3 function: $f(x_1, x_2) = \sqrt{90}\,(x_2 - x_1^2)$;

19. Wood5 function: $f(x_1, x_2) = \sqrt{10}\,(x_1 + x_2 - 2)$;

20. Wood6 function: $f(x_1, x_2) = (x_1 - x_2)/\sqrt{10}$

REFERENCES

Hickernell, F.J. (1998): A generalized discrepancy and quadrature error bound, *Mathematics of Computation*, 67, 299-322.

Koehler, J. R., Owen, A. B. 1996: *Computer experiments*. In Ghosh, S. and Rao, C. R., editors, *Handbook of Statistics*, 13, 261--308. Elsevier Science, New York.

Liu, L., 2004: Employing simulation and optimizer to optimize experimental design and structural topology, dissertation, Systems Science Ph.D. Program, Portland State University.

Liu, L., 2005: Could enough samples be more important than better designs for computer experiments? 38th Annual Simulation Symposium, Spring Simulation Multiconference, San Diego, April 2-8, 2005.

Morris, M. D. and Mitchell, T. J., 1995, "Exploratory Designs for Computer Experiments," Journal *of Statistical Planning and Inference*, 43, 381-402.

Nielsen, H.B., Lophaven, S.N., and Søndergaard, J.: DACE: A MATLAB KrigingToolbox <http://www.imm.dtu.dk/~hbn/dace>

Santner, T., Williams, B, and Notz, W. 2003: *Design and Analysis of Computer Experiments*, Springer-Verlag, New York.

AUTHOR BIOGRAPHIES

LONGJUN LIU, Ph.D., Senior Design Engineer. His main research interests include design optimization and computer experiments or metamodeling. He is a member of SCS, AIAA, ASME, etc. His email address is <longjunliu@comcast.net> and his web address is http://www.sysc.pdx.edu/lliu/pageLiu.htm

WAYNE WAKELAND, Ph.D., Associate Professor. His main research interests include modeling and simulation. His email address is <wakeland@pdx.edu> and his web address is <http://www.sysc.pdx.edu/faculty/Wakeland/index.html>