

## DETC99/DAC-8622

### FITTING FUNCTIONS TO DATA IN HIGH DIMENSIONAL DESIGN SPACE

Xiaoou Wang

Yingying Liu

Erik K. Antonsson, Ph.D., P.E.\*

Engineering Design Research Laboratory

California Institute of Technology

1200 E. California Blvd.

Pasadena, California 91125

#### ABSTRACT

One approach to rapidly exploring large design spaces is to evaluate the performance of a small number of representative points in the space, then based on those points, construct an approximation to the response over a region of interest. Linear, piecewise linear, quadratic and multivariate adaptive regression splines (MARS) models are fit to an example 5-dimensional data set representative of information available in preliminary engineering design. When the number of data points representing a high-dimensional response is small, all of the approximation models appear to perform nearly equally.

#### KEYWORDS:

Data Fitting; Response Surfaces; Piecewise Linear Approximations; Multivariate Adaptive Regression Splines; Rapid Exploration of Large Design Spaces.

#### 1 INTRODUCTION

An important task in engineering design analysis consists of running complex computer analysis codes, *i.e.*, supplying a vector of design variables (inputs)  $\vec{d}$  and computing a vector of responses (outputs)  $\vec{p}$ . Despite advances in computing power, the cost of running many analysis codes remains significant, *e.g.*, one evaluation of a finite-element model can take minutes to hours. When the number of dimensions of the design variable space is large, any systematic exploration of this space becomes pro-

hibitively expensive.

*Design and Analysis of Experiments* (Montgomery, 1991) is a well developed method for choosing points to evaluate in a high dimensional space. However, because computer models are deterministic, and thus lack random errors, computer analyses are different from physical experiments, calling for distinct techniques (Jerome Sacks and Wym, 1989; Simpson, Peplinski, Koch, and Allen, 1997). In prior work (Law and Antonsson, 1995, 1996) we implemented Design of Experiment methods to explore design spaces with many dimensions. The goal of this prior work was to provide an approximate exploration of a large design space with the minimum number of evaluations. This parsimonious exploration of a large space is particularly valuable in preliminary design, when much information related to the design is not yet precise.

In the work reported here, the prior Design of Experiments research mentioned above is expanded to assess the performance of several different techniques for approximating design performance over a large space based on a limited number of points (experiments). The efficiency and accuracy of different methods for approximating the output of a computer model of design performance are compared. The particular example used here is the bending stiffness computed from a finite-element model of a passenger automobile chassis. The approximation techniques that are compared are a traditional linear regression model (linear (LM) and piecewise linear (PLM)), a second-order quadratic regression model (QDM), a higher-order polynomial regression model (HM), and a high-order nonlinear regression model

---

\*corresponding author (erik@design.caltech.edu)

(Multivariate Adaptive Regression Splines (MARS)) (Friedman, 1991; Rai, 1998). These approximations provide a computationally inexpensive way to examine the relationship between  $\vec{d}$  and  $\vec{p}$  in a large design space.

## 2 MODEL DESCRIPTION AND DESIGN OF EXPERIMENTS

As mentioned above, several types of polynomial regression models are used in this paper. Regression analysis uses a collection of data points to construct a function  $\hat{y} = \hat{f}(x_1, \dots, x_n)$  that approximates  $y = f(x_1, \dots, x_n)$  over a region of interest. These models span a range of polynomial order as well as a range of inter-variable interactions. The linear model includes one-way interactions, the quadratic model includes restricted interactions, and the nonlinear MARS model includes unrestricted interactions.

### 2.1 Linear Model (LM)

It is called *Linear Model* just for convenience. It is actually only linear in the main direction of each independent variable.

The linear model of  $n$  independent variables with up to  $m$  order interactions is:

$$\hat{y} = \hat{f}(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i x_i + \sum_{i_1=1}^n \sum_{i_2=i_1+1}^n a_{i_1 i_2} x_{i_1} x_{i_2} + \dots + \sum_{i_1=1}^n \dots \sum_{i_m=i_{m-1}+1}^n a_{i_1 \dots i_m} x_{i_1} \dots x_{i_m} \quad (1)$$

If  $m < n$ , then a fractional factorial design will be used and its resolution is determined by design of experiments. If  $m = n$ , then a full factorial design will be used.

### 2.2 Piecewise Linear Model (PLM)

In piecewise linear model, the design variable space is divided into many rectangular sub-spaces and each sub-space has its own linear model, as above.

### 2.3 Quadratic Model (QDM)

The quadratic model of  $n$  independent variables with up to  $m$  order interactions adds quadratic terms to the linear model above:

$$\hat{y} = \hat{f}(x_1, \dots, x_n) = \hat{y}_{\text{linear}}(x_1, \dots, x_n) + \sum_{i=1}^n a_{ii} x_i^2 \quad (2)$$

The central composite design is used to decide the coefficients of the quadratic terms. It is simply the factorial design plus the central points of each face.

### 2.4 Higher-Order Polynomial Regression Model (HM)

The higher-order model of  $n$  independent variables, with higher-order interactions determined by the values of  $n$  and  $m$ , is:

$$\hat{y} = \hat{f}(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i x_i + \sum_{i_1=1}^n \sum_{i_2=1}^n a_{i_1 i_2} x_{i_1} x_{i_2} + \dots + \sum_{i_1=1}^n \dots \sum_{i_m=1}^n a_{i_1 \dots i_m} x_{i_1} \dots x_{i_m} \quad (3)$$

This is equivalent to a product of several linear regression polynomials. If the number of data points is less than the number of terms in the polynomial, only the coefficients of an equal number of lower-order terms are non-zero.

### 2.5 Nonlinear Regression MARS Model

MARS (Friedman, 1991; Rai, 1998) fits high dimensional data to an expansion in multivariate spline basis functions. The number of basis functions, the product degree, and the knot locations are automatically determined by, and adaptive to, the data. The model produces strictly continuous approximation with continuous derivatives, and identifies the contributions from additive terms and multivariable interactions. The method is attractive due to its low computational cost.

“The approximation takes the form of an expansion in multivariate spline basis functions:

$$\hat{y} = \hat{f}(x_1, \dots, x_n) = a_0 + \sum_{m=1}^M a_m B_m(x_1, \dots, x_n) \quad (4)$$

with:

$$B_0(x_1, \dots, x_n) = 1, \quad (5)$$

$$B_m(x_1, \dots, x_n) = \prod_{k=1}^{K_m} b_{km}(x_{v(k,m)} | t_{km}). \quad (6)$$

The  $\{a_m\}_0^M$  are the coefficients of the expansion, obtained by minimizing a GCV (generalized cross-variance) criterion that is the average squared residual of the fit to the data (numerator) times a penalty (inverse denominator) to account for the increased variance associated with increasing model complexity (number of basis functions  $M$ ). Each multivariate spline basis function  $B_m$  is the product of univariate spline basis functions  $b_{km}$  which is either order 1 or cubic, depending on

the degree-of-continuity of the approximation, each of a single input variable  $x_{v(k,m)}$ , and characterized by a knot at  $t_{km}$ . The multivariate spline basis functions  $B_m$  are adaptive in that the number of factors  $K_m$ , the variable set  $V(m) = \{v(k,m)\}_1^{K_m}$ , and the parameter set  $t_{km}$  are all determined by the data.” (Friedman, 1988, Page 17)

For further details see Friedman (1991) (Friedman, 1991).

In our modeling of 5-dimensional input variables on an engineering workstation, results can be obtained essentially immediately when 32 observations are used and a maximum of 30 basis functions are allowed. Even when 3125 data points are used and a maximum of 40 basis functions are allowed, the regression computations take less than 80 seconds.

### 3 PROBLEM DESCRIPTION

The actual function to be fitted in the example presented here is the bending stiffness of a Volkswagen passenger automobile chassis (shown in Figure 1) computed from a finite-element model (shown in Figure 2) in a design space of five variables ( $n = 5$ ):

- $x_1 =$  A Pillar Thickness [mm] (0.7–1.1)
- $x_2 =$  B Pillar Thickness [mm] (1.1–1.5)
- $x_3 =$  Floor Rail Thickness [mm] (0.8–1.2)
- $x_4 =$  Floor Thickness [mm] (1.0–1.4)
- $x_5 =$  B Pillar Location [mm] (-50–150)

The bending stiffness has unit of pounds force per inch (lbf/in), and varies from 2740.9 to 3364.9, and the range is 624.0.

Although  $n = 5$  is not a high-dimensional design space, all the models we build here can be easily extended to higher dimensional data without any modification. We use this  $n = 5$  VW automobile design problem here just for simplicity and existent meaningful data.

Table 1 lists the models that have been built to fit the actual response function. In all cases,  $n=5$ . For the LM model,  $m = 2$  was used for 17 points, and  $m = 5$  was used for 32 points. For the QDM model,  $m = 2$  for 14 and 19 points. For the HM model,  $m = 4$  was used for 34 and 106 points, and  $m = 10$  for 243 points. For each model,  $m$  and the number of evaluations are decided by the design of experiment.

### 4 RESULTS

Figure 3 shows a projection of the 5-dimensional bending stiffness data surface onto 2-dimensions (*Floor thickness* and *B pillar location*). This illustrative projection was created by holding each of the 3-dimensions not shown (*A pillar thickness*, *B pillar thickness* and *Floor rail thickness*) at a constant value. The approximations computed here fit all 5 dimensions of the

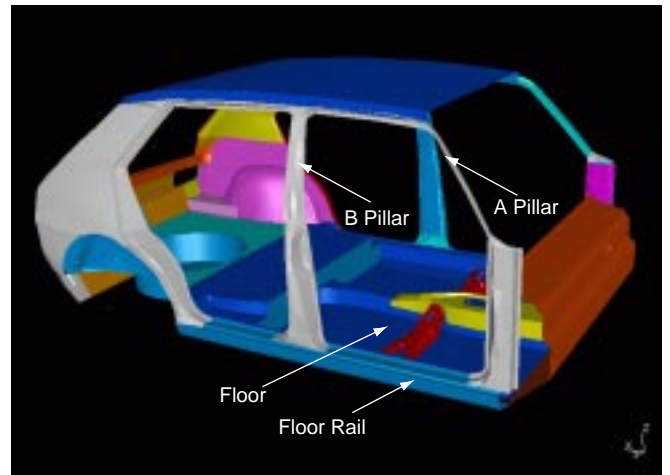


Figure 1. GEOMETRIC MODEL OF BODY-IN-WHITE IN SDRC I-DEAS.

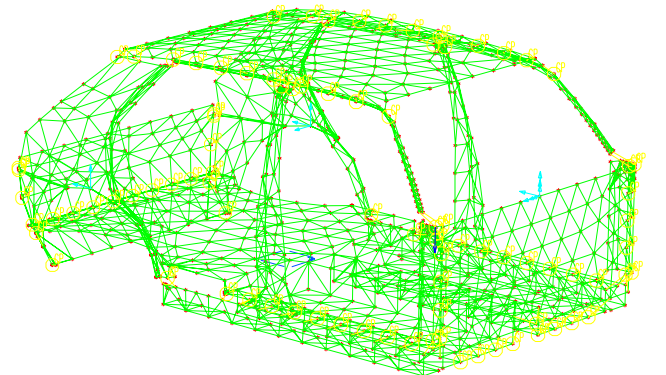


Figure 2. FINITE ELEMENT MODEL OF BODY-IN-WHITE.

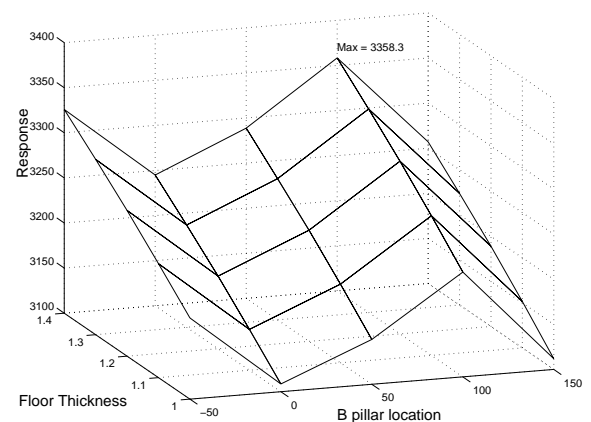


Figure 3. BENDING STIFFNESS DATA PROJECTEC ONTO 2-DIMENSIONS.

Table 1. Data Fitting Models.

Model	Design of Experiments	No. of Evaluations for a Single Model	No. of Models
QDM, MARS	Resolution III + half central pts	14	1
LM, MARS	Resolution V + all central pts	17	1
QDM, MARS	Resolution III	19	1
LM, MARS	Full	32	1
HM		34	1
HM		106	1
PLM	Full	243	32
HM		243	1
MARS	Full	243	1
PLM	Full	3125	1024
MARS	Full	3125	1

design space, however, to enable graphical comparison, the figures only show variations in bending stiffness as a function of 2 dimensions.

The bending stiffness response in the *B pillar location* direction has the highest nonlinearity of the 5 directions. The nonlinearity in that direction is reflected, to some degree, in the errors of most regression models. Polynomial models and the MARS model were fitted to the data selected by each design of experiment listed in Table 1. Figures 4 through 6 show the error surfaces of the polynomial and MARS models with 19, 243 and 3125 evaluations. The errors are computed by systematically computing the bending stiffness at 5 equally spaced points in each of the 5 variables, thus producing a 5-dimensional set of 3125 data points. The error is the difference between each approximation model and the 3125 computed data points. As before, these figures are plotted by projecting the error onto two directions (*Floor thickness* and *B pillar location*) of the design space while fixing the values in each of the other three directions.

Table 2 shows some numerical results.

Figures 7 through 9 show error statistics from the regression models. In each case, the mean square error is computed as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^m (error_i)^2. \quad (7)$$

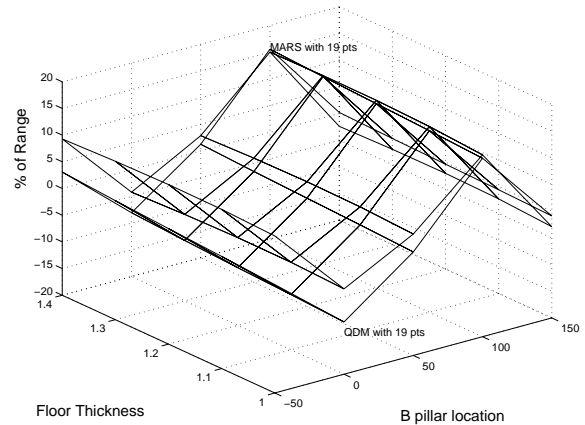


Figure 4. THE ERRORS OF POLYNOMIAL AND MARS MODELS WITH 19 EVALUATIONS.

Figure 7 shows the mean square error computed from the difference between the computed bending stress and the regression models at the points used to build the regression model. This illustrates how close each regression model is to the known data. Figure 8 shows the mean square error computed at the balance of the 5-dimensional set of 3125 data points not used to build the regression model. Figure 9 shows the maximum error at all 3125 data points. Figures 8 and 9 illustrate how well each regression

Table 2. Selected Numerical Results

Model	No. Pts. Evaluated	Max Error (%)	Mean Squared Error (%) at	
			Evaluated Pts.	Unevaluated Pts.
LM	14	15.23	0.65	6.12
QDM	14	15.48	0.05	6.07
MARS	14	15.55	4.19	5.89
LM	17	18.41	1.32	7.77
MARS	17	-22.40	2.30	7.77
LM	19	15.35	0.83	6.14
QDM	19	15.56	0.14	6.07
MARS	19	15.51	3.86	5.85
LM	32	16.09	0.00	6.23
HM	34	15.00	0.02	6.13
HM	106	14.93	0.00	6.18
LM	243	15.82	1.05	6.40
QDM	243	14.76	0.18	6.32
HM	243	14.93	0.00	6.33
PLM	243	-8.23	0.00	1.17
MARS	243	14.89	0.04	6.33
LM	3125	12.50	5.49	N/A
QDM	3125	11.27	5.41	N/A
PLM	3125	0.00	0.00	N/A
MARS	3125	1.06	0.24	N/A

model approximates the data in areas of the design space away from the points used to build the regression model.

Finally, the error at the point where the bending stiffness itself is a maximum is shown in Table 3 and in Figure 10 for all models.

## 5 DISCUSSION

LM, QDM, HM and PLM are all polynomial models, and MARS method also uses cubic polynomial as the spline. We choose the polynomial just because its simplicity, both in expression and computation of the model.

Polynomial models can produce the smallest Mean Square Error (MSE) at all evaluated points because they are generated by the method of the least MSE. The MSE of all unevaluated

points is a more important indicator of the quality of an approximation because it indicates how well the model fits the actual function at the unknown (for this model) points. It is used as an indication of the accuracy of each model. For MARS models, the accuracy is similar to the other models when the number of the points used to build the model is less than or equal to 243. The MSE for all unevaluated points is, of course, not meaningful when all 3125 points are used to generate the model (since there are no “unevaluated” points). Comparing the maximal error of each MARS model, the one with 3125 evaluations is much better than the one with 243 evaluations.

In the same way, similar conclusions can be drawn for polynomial models. The only difference is that the piecewise polynomial model with 243 evaluations is much better than the poly-

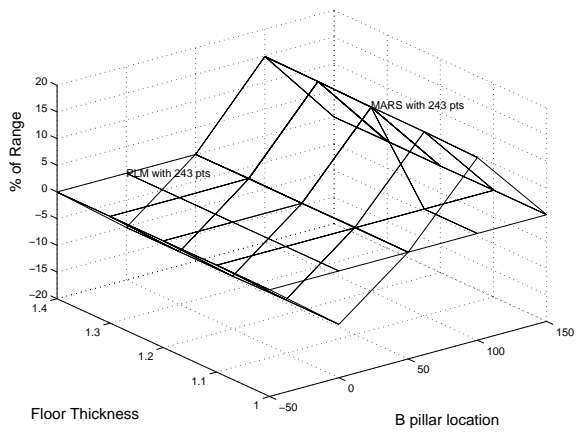


Figure 5. THE ERRORS OF POLYNOMIAL AND MARS MODELS WITH 243 EVALUATIONS.

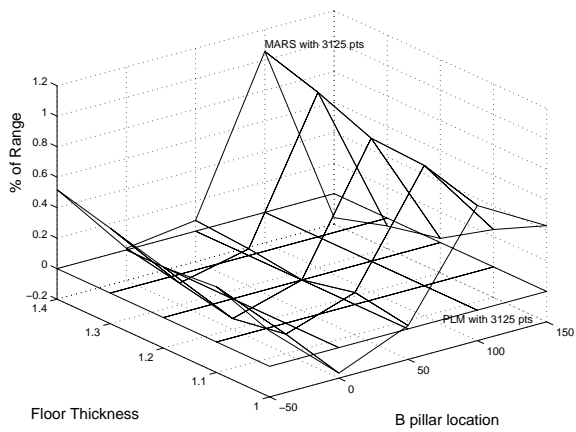


Figure 6. THE ERRORS OF POLYNOMIAL AND MARS MODELS WITH 3125 EVALUATIONS.

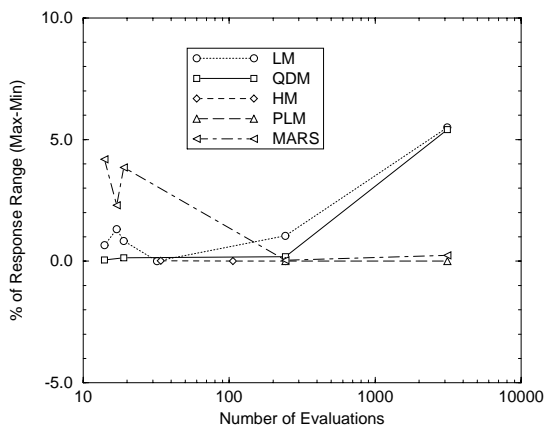


Figure 7. THE MEAN SQUARE ERROR AT ALL EVALUATED POINTS.

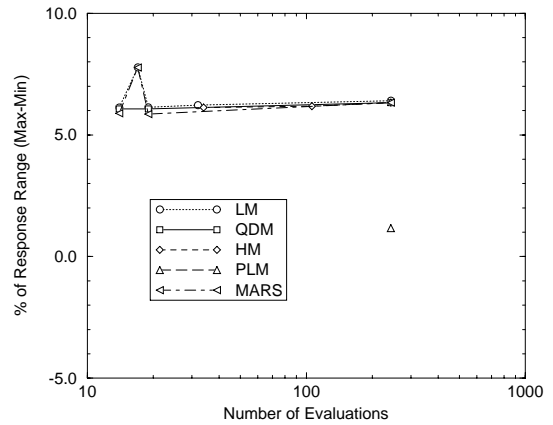


Figure 8. THE MEAN SQUARE ERROR AT ALL UNEVALUATED POINTS.

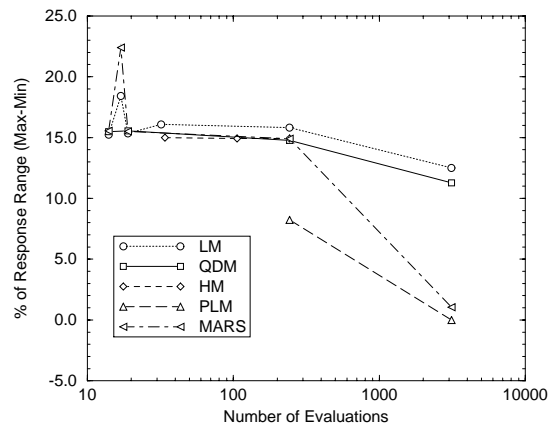


Figure 9. THE MAXIMUM ERROR OF ALL DATA POINTS.

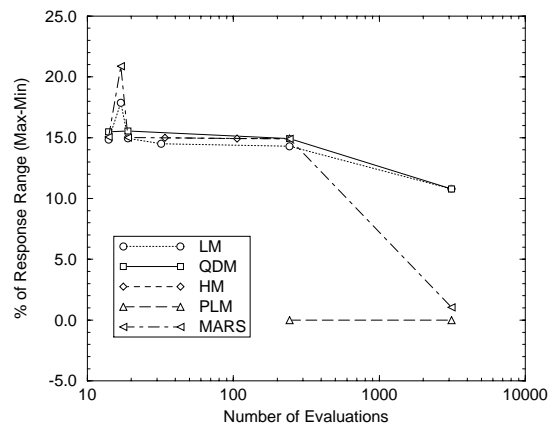


Figure 10. THE ERRORS AT MAXIMUM RESPONSE.

Table 3. Error at Maximum Response.

Model	No. of Points Evaluated	Error at Max Response (%)
LM	14	14.83
QDM	14	15.48
MARS	14	15.06
LM	17	17.88
MARS	17	20.88
LM	19	14.96
QDM	19	15.56
MARS	19	15.02
LM	32	14.51
HM	34	15.00
HM	106	14.93
LM	243	14.30
QDM	243	14.93
HM	243	14.93
PLM	243	0.00
MARS	243	14.89
LM	3125	10.78
QDM	3125	10.78
PLM	3125	0.00
MARS	3125	1.05

mial model with 32 evaluations. When the number of evaluations is small ( $\leq 32$ ), there is not a significant difference between the polynomial models and the MARS model. However, when the number of evaluations is increased and the piecewise linear model is used, the polynomial model is better than the MARS model, especially with 243 evaluations.

The single linear model with 3125 evaluations is also generated (listed in tables but not plotted in figures). The accuracy is almost the same as any other single polynomial model.

## 6 CONCLUSION

The accuracy of the approximation is decided by the choice of regression model, and the number of data points used. When

very little information is available, the result is almost the same no matter which model is used. When a large number of evaluations are available (*e.g.*,  $> 5n$  where  $n$  is the number of design space dimensions), using a more complex regression model can generate a better approximation. In the example shown here, when the number of evaluations is 19, the mean square error for all unevaluated points is almost the same for each model. But when the number of evaluations is 243, the mean square error for all unevaluated points of the piecewise linear model is the smallest. When enough evaluations are available to fit a higher order (larger than 3) model, the piecewise linear model can get almost the same result with less oscillation. When little is known about the actual shape of the response (typical of preliminary engineering design) (linear, nonlinear or highly nonlinear) and the evaluated data points are sparse in the design space (typical of preliminary engineering design), all of the regression methods examined here appear to be nearly equally good (at least for problems similar to the the non-linear, non-monotonic 5-dimensional example problem shown here).

## ACKNOWLEDGEMENTS

This material is based upon work supported, in part, by the National Science Foundation under NSF Grant Number DMI-9523232. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsors.

## REFERENCES

- Friedman, J. H., 1988, "Fitting Functions to Noisy Data in High Dimensions," In *Computing Science and Statistics*, pp. 13–43. American Statistical Association Proceedings of the 20th Symposium on the Interface.
- Friedman, J. H., 1991, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, Vol. 19, (No. 1), pp. 1–141.
- Jerome Sacks, William J. Welch, T. J. M., and Wym, H. P., 1989, "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, (No. 4), pp. 409–435.
- Law, W. S., and Antonsson, E. K., 1995, "Optimization Methods for Calculating Design Imprecision," In *Advances in Design Automation - 1995*, Vol. 1, pp. 471–476. ASME.
- Law, W. S., and Antonsson, E. K., 1996, "Multi-Dimensional Mapping of Design Imprecision," In *8th International Conference on Design Theory and Methodology*. ASME.
- Montgomery, D. C., 1991, *Design and Analysis of Experiments* Wiley, New York.
- Rai, S., 1998, "Optimal Experimental Setpoint Determination in Systems with Multiple Output Responses," In *Proceedings of DETC98: 1998 ASME Design Engineering Technical Conference*. ASME Paper Number DETC98/DTM-5678.

Simpson, T. W., Peplinski, J., Koch, P. N., and Allen, J. K., 1997, "On the Use of Statistics in Design and the Implications for Deterministic Computer Experiments," In *9th International Conference on Design Theory and Methodology*. ASME.