AIAA JOURNAL
Vol. 39, No. 10, October 2001

Metamodeling of Combined Discrete/Continuous Responses

Martin Meckesheimer,* Russell R. Barton,† and Timothy Simpson‡
Pennsylvania State University, University Park, Pennsylvania 16802
and
Frej Limayem§ and Bernard Yannou¶
Ecole Centrale Paris, 92295 Chatenay-Malabry, France

Metamodels are effective for providing fast-running surrogate approximations of product or system performance. Because these approximations are generally based on continuous functions, they can provide poor fits of discontinuous response functions. Many engineering models produce functions that are only piecewise continuous, due to changes in modes of behavior or other state variables. The use of a state-selecting metamodeling approach that provides an accurate approximation for piecewise continuous responses is investigated. The proposed approach is applied to a desk lamp performance model. Three types of metamodels, quadratic polynomials, spatial correlation (kriging) models, and radial basis functions, and five types of experimental designs, full factorial designs, D-best Latin hypercube designs, fractional Latin hypercubes, Hammersley sampling sequences, and uniform designs, are compared based on three error metrics computed over the design space. The state-selecting metamodeling approach outperforms a combined metamodeling approach in this example, and radial basis functions perform well for metamodel construction.

I. Introduction

MULTIDISCIPLINARY design has many applications in concurrent engineering, where product performance and manufacturing plans are designed simultaneously. The objective is to manage the design process more effectively to allow the designer to carry out rapid evaluations of design alternatives, analysis, and decision making in a multidisciplinary design environment.

With the aid of computers for simulation and analysis, discipline-specific product and process models can be formulated and used to analyze many complex engineering systems, as opposed to exercising time consuming and expensive experimentation on these physical systems. However, running complex computer models can also be expensive, especially during optimization, because a single evaluation of an alternative may take several minutes or even hours to complete.

Current research approximates the true input/output relationship of the disciplinary model with a moderate number of computer experiments and then uses the approximate relationship to make predictions at additional untried inputs. Inasmuch as these simple mathematical approximations to discipline-specific product and process models are models of an engineering model, we call them metamodels because they provide a model of the model.[1] Furthermore, metamodels facilitate integrating subsystem and disciplinary analyses because they are typically written on the same computer system and use the same language. This may not be the case for the original disciplinary analysis codes. Note that disciplinary computer simulation codes are deterministic in nature, in which case the analysis differs from experiments on physical systems in the sense that no random variations are observed, that is, for a particular set of input settings, two deterministic simulation runs will yield the exact same output. The implications of the deterministic nature of computer experiments on metamodel construction and assessment are discussed further by Simpson et al.[2]

Received 8 August 2000; revision received 1 April 2001; accepted for publication 1 April 2001.

*Graduate Research Assistant, Marcus Department of Industrial and Manufacturing Engineering, 310 Leonhard Building.
†Professor, Marcus Department of Industrial and Manufacturing Engineering, 310 Leonhard Building; rbarton@psu.edu.
‡Assistant Professor, Marcus Department of Industrial and Manufacturing Engineering, 310 Leonhard Building. Member AIAA.
§Graduate Research Assistant, Laboratoire Productique Logistique, Grande Voie des Vignes.
¶Professor, Laboratoire Productique Logistique, Grande Voie des Vignes.

With a metamodel-based design approach researchers hope to simplify optimization and/or examination of the system performance over the design space. Moreover, because metamodels generally allow rapid calculation, the designer obtains a tool for conducting real-time exploration of the design space in the preliminary stages of design. For example, metamodels have been successfully used for efficient Pareto frontier exploration (see Ref. 3) and have found many applications in structural[4] and multidisciplinary optimization.[5] The mathematical approximation can be written as $y = f(\boldsymbol{x}) \cong \phi(\boldsymbol{x})$, where $f(\boldsymbol{x})$ represents the original disciplinary model function and $\hat{y} = \phi(\boldsymbol{x})$ is the metamodel approximation to y. Whereas the use of metamodels allows a faster analysis than the original, complex engineering models permits, the metamodel approximation introduces a new element of uncertainty. As a result, efficient methods to assess metamodel fit at the system and subsystem level are required. Validation strategies for metamodel assessment can involve the use of additional data[6,7] or might be based on resampling strategies.[8]

Metamodeling is a process involving the choice of an experimental design, a metamodel type and its functional form for fitting, and a validation strategy to assess the metamodel fit. With these definitions in mind, in the next section we briefly discuss the metamodel and experiment design types that are utilized in this study. The objective in this study is to analyze different strategies for modeling combined discrete/continuous response problems. In Sec. III, we present a generic description of the problem, followed by some possible solution approaches. A desk lamp example is introduced in Sec. IV. Evaluation measures are discussed in Sec. V, and the results are presented in Sec. VI. A brief summary concludes the study in Sec. VII.

II. Metamodeling

A. Response Surface Methodology

Response surface methodology[9] has been used effectively in a variety of applications,[2,4,5] to produce a metamodel with a low-order polynomial in a relatively small region of the factor space. Typically, first- and second-order polynomial models of the form

$$\phi(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i \boldsymbol{x}_i$$

$$\phi(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i \boldsymbol{x}_i + \sum_{i=1}^{k} \beta_{ii} \boldsymbol{x}_i^2 + \sum_{i} \sum_{j>i} \beta_{ij} \boldsymbol{x}_i \boldsymbol{x}_j \qquad (1)$$

are fitted to the system response. The β parameters of the polynomials are calculated using least-squares regression to fit the response surface approximations to empirical data or data generated from simulation or analysis routines; these approximations can then be used for prediction.

Note that a second-order response surface is not intended to fit well over the entire region of operability, but only in a relatively narrow region of interest that has been located by prior experimentation. Consequently, many researchers advocate the use of a sequential response surface modeling approach using move limits[10] or a trust region approach.[11] For instance, the concurrent subspace optimization procedure uses data generated during concurrent subspace optimization to develop response surface approximations of the design space, which form the basis of the subspace coordination procedure.[12–14] Similarly, the hierarchical and interactive decision refinement methodology uses statistical regression and other metamodeling techniques to decompose recursively the design space into subregions and to fit each region with a separate model during design space refinement.[15] An advantage of using low-order polynomial metamodels is that they involve relatively few parameters and they permit gaining insight into the model behavior and identifying significant model parameters. Although approximation functions based on low-order polynomials are the most widely used, other regional or global approximating functions, such as radial basis functions, spatial correlation models, splines, or neural networks have also been investigated.[16,17] Simpson et al.[18] compare response surface and kriging models for multidisciplinary design optimization. Although these metamodels may provide better approximations to response functions of arbitrary shape than low-order polynomials, they may be more difficult to interpret, and sometimes involve a larger number of model coefficients, for example, radial basis functions, or more complex computations, for example, kriging. In addition to quadratic polynomial approximations, we also considered radial basis approximations and spatial correlation (kriging) models for the present study.

B. Radial Basis Functions

A heuristic approach led Hardy[19] to use linear combinations of a radially symmetric function based on Euclidean distance or similar metric. A simple radial basis function form is

$$\phi(\boldsymbol{x}) = \sum_i \beta_i \|\boldsymbol{x} - \boldsymbol{x}^i\| \tag{2}$$

where $\|\cdot\|$ represents the Euclidean norm and the sum is over an observed set of system responses, $\{[\boldsymbol{x}^i, f(\boldsymbol{x}^i)]\}$, $i = 1, \ldots, n$. Replacing $\phi(\boldsymbol{x})$ with $f(\boldsymbol{x}^i)$ and solving the resulting linear system yields the β_i coefficients. As commonly applied, the method is an interpolating approximation. Radial basis function approximations have produced good fits to arbitrary contours of both deterministic and stochastic response functions.[20]

C. Spatial Correlation (Kriging) Models

Spatial correlation metamodels are a class of approximation techniques that show good promise for building accurate global approximations of a design space. Spatial correlation metamodeling is also known as kriging[21] or (design and analysis of computer experiments modeling, named after the inaugural article by Sacks et al.[22] In a spatial correlation metamodel, the design variables are assumed to be correlated as a function of distance during prediction, hence the name spatial correlation metamodel. These metamodels are extremely flexible because the metamodel can either honor the data, providing an exact interpolation of the data, or smooth the data, providing an inexact interpolation, depending on the choice of the correlation function.[21]

A spatial correlation metamodel is a combination of a polynomial model plus departures of the form

$$y(\boldsymbol{x}) = f(\boldsymbol{x}) + Z(\boldsymbol{x}) \tag{3}$$

where $y(\boldsymbol{x})$ is the unknown function of interest, $f(\boldsymbol{x})$ is a known polynomial function of $\boldsymbol{x}$, and $Z(\boldsymbol{x})$ is the realization of a normally distributed Gaussian random process with mean zero, variance σ^2, and nonzero covariance. The $f(\boldsymbol{x})$ term in Eq. (3) is similar to the polynomial model in a response surface, providing a global model of the design space. In many cases $f(\boldsymbol{x})$ is simply taken to be a constant term β (see Refs. 18, 22, and 23).

Whereas $f(\boldsymbol{x})$ globally approximates the design space, $Z(\boldsymbol{x})$ creates localized deviations so that the kriging model interpolates the n_s sampled data points. The covariance matrix of $Z(\boldsymbol{x})$ is given by

$$\text{cov}[Z(\boldsymbol{x}^i), Z(\boldsymbol{x}^j)] = \sigma^2 \boldsymbol{R} \qquad (\boldsymbol{R} = [R(\boldsymbol{x}^i, \boldsymbol{x}^j)]) \tag{4}$$

In Eq. (4), $\boldsymbol{R}$ is the correlation matrix, and $R(\boldsymbol{x}^i, \boldsymbol{x}^j)$ is the correlation function between any two of the n_s sampled data points $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$. $\boldsymbol{R}$ is a $(n_s \times n_s)$ symmetric matrix with ones along the diagonal. The correlation function $R(\boldsymbol{x}^i, \boldsymbol{x}^j)$ is specified by the user; example correlation functions may be found in Refs. 22–24. In this work we employ a Gaussian correlation function of the form

$$\boldsymbol{R}(\boldsymbol{x}^i, \boldsymbol{x}^j) = \exp\left[-\sum_{k=1}^{n_{dv}} \theta_k \left|x_k^i - x_k^j\right|^2\right] \tag{5}$$

where n_{dv} is the number of design variables, θ_k are the unknown correlation parameters used to fit the model, and the x_k^i and x_k^j are the kth components of sample points $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$, respectively.

Predicted estimates $\hat{\boldsymbol{y}}$ of the response at untried values of $\boldsymbol{x}$ are given by

$$\hat{\boldsymbol{y}} = \hat{\beta} + \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{f}\hat{\beta}) \tag{6}$$

where $\boldsymbol{y}$ is the column vector of length n_s that contains the values of the response at each sample point and $\boldsymbol{f}$ is a column vector of length n_s that is filled with ones when $f(\boldsymbol{x})$ is taken as a constant as is the case in this paper. In Eq. (6), $\boldsymbol{r}^T(\boldsymbol{x})$ is the correlation vector of length n_s between an untried $\boldsymbol{x}$ and the sample points $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^{ns}\}$ and is given by

$$\boldsymbol{r}^T(\boldsymbol{x}) = [R(\boldsymbol{x}, \boldsymbol{x}^1), R(\boldsymbol{x}, \boldsymbol{x}^2), \ldots, R(\boldsymbol{x}, \boldsymbol{x}^{ns})] \tag{7}$$

Finally, in Eq. (6), $\hat{\beta}$ is estimated as

$$\hat{\beta} = (\boldsymbol{f}^T\boldsymbol{R}^{-1}\boldsymbol{f})^{-1}\boldsymbol{f}^T\boldsymbol{R}^{-1}\boldsymbol{y} \tag{8}$$

The estimate of the variance $\hat{\sigma}^2$ of the sample data, denoted as $\boldsymbol{y}$, from the underlying global model (not the variance in the observed data itself) is given by

$$\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{f}\hat{\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{f}\hat{\beta})/\boldsymbol{n}_s \tag{9}$$

where $\boldsymbol{y}$ is the vector of y values and $f(\boldsymbol{x})$ is assumed to be the constant $\hat{\beta}$. The maximum likelihood estimates, that is, best guesses, for the θ_k in Eq. (5) used to fit the model are found by maximizing the expression

$$\max_{\theta > 0, \theta \in \Re^n} -[\boldsymbol{n}_s \ln(\hat{\sigma}^2) + \ln|\boldsymbol{R}|]/2 \tag{10}$$

where $\hat{\sigma}^2$ and $|\boldsymbol{R}|$ are both functions of θ. Although any values for θ_k create an interpolative model, the best kriging model is found by solving the k-dimensional unconstrained nonlinear optimization problem given by Eq. (10). We use a simulated annealing algorithm[25] to perform this optimization.

D. Design of Experiments

The design of experiments for metamodel fitting is critical to the success of the method. If the experimental runs of the model codes are not carefully chosen, the fitted metamodel approximation can be poor. Experimental design issues for metamodels are discussed in Refs. 3 and 17. In this study, we use five different strategies to select points in the design space for fitting each type of metamodel; the designs are described briefly as follows.

1. Full Factorial Designs

The most commonly used experimental design strategies are factorial designs[26] which comprise full-factorial and fractional-factorial designs. For factorial designs, each dimension of the design space is covered by a series of (typically) uniformly spaced values, and their Cartesian product provides a map of the system response. Because the number of points required in factorial designs becomes prohibitively large as the number of factors in the model increases, fractional factorials are often used as an efficient alternative to full-factorial designs; however, the effectiveness of these designs depends largely on the nature of the response.

2. Latin Hypercube Designs

An alternative experimental design strategy is a Latin hypercube, which was the first type of design proposed specifically for computer experiments.[27] Latin hypercubes offer flexible sample sizes while distributing points randomly over the design factor space; it has been shown that these designs can have relatively small variance when measuring output variance.[22]

Because Latin hypercube design points are selected at random, it is possible to generate poor, for example, diagonal, designs. To obtain a good Latin hypercube design, we create a set of randomly generated candidate designs. The best design is then selected based on the D-optimality criterion in which the volume of the confidence ellipsoid for the true value β about the observed random vector $\hat{\beta}$ is minimized; this is equivalent to maximizing the determinant of the matrix $\boldsymbol{X}'\boldsymbol{X}$. Because we select the design from a set of candidate designs (rather than from all possible Latin hypercube designs of that size), we refer to it as a D-best Latin hypercube design, rather than a D-optimal Latin hypercube design.

3. Full Factorial Latin Hypercube Designs

We also consider the use of a slightly modified version of factorial hypercube designs, a hybrid design strategy that combines the use of fractional factorial designs with Latin hypercube sampling.[28] In this strategy, the design space is divided into p^p hypercubes, with $p = k - m$, where k is the number of design variables and m is the number of fractionation of the design. In each hypercube, n points are generated using Latin hypercube sampling.

4. Hammersley Sampling Sequence

Latin hypercube techniques are designed for uniformity along a single dimension where subsequent columns are randomly paired for placement on a k-dimensional cube. Hammersley sequence sampling (HSS) provides a low-discrepancy experimental design for placing n points in a k-dimensional hypercube,[29] providing better uniformity properties over the k-dimensional space than a Latin hypercube. Low discrepancy implies a uniform distribution of points in space. One measure of discrepancy is the rectangular star discrepancy, defined in Eq. (1.2.3) in Ref. 30. It computes a numerical measure of the difference between the actual placement of the design points and a perfectly uniform dispersion of design points across a rectangular design region. The reader is referred to Ref. 29 for a formal definition of Hammersley points and an explicit procedure for generating them.

5. Uniform Designs

Uniform designs are a class of designs based solely on number–theoretic methods in applied statistics.[30] A uniform design seeks to uniformly scatter n points in a k-dimensional design space to minimize their discrepancy. Further details on uniform designs and their construction may be found in Ref. 31.

E. Combined Discrete/Continuous Responses

Often, engineering applications are characterized by combined discrete/continuous responses. For example, in chemical processes a discontinuity may correspond to the beginning of a reaction. Similarly, the phase change of elastic–plastic deformation in materials represents a discrete change in the nature of the response of a mechanical part or system, although the response is continuous. Identifying the region in which the discontinuity occurs is not necessarily trivial because it may be defined by an implicit parameter relationship in the engineering model.

In this paper, we consider a model of light intensity vs position, used for the design of a desk lamp. This model has a combined discrete/continuous nature, because the position at which a light ray reaches the table surface can change discretely depending on whether the ray reaches the table directly or is deflected by the lamp reflector. For example, the table position of a ray that reflects near the outside edge of the reflector will change discretely when the design parameter corresponding to the length of the reflector is shortened a small amount. The discrete change in ray position corresponds to a change in the discrete state variable $s \in S = \{\text{direct, reflected}\}$.

III. Generic Problem Description

In this section, an artificial example with an imaginary system is used to describe the problem and identify three distinct solutions. Consider the system composed of combined discrete/continuous responses shown in Fig. 1. The imaginary system has a set of continuous calculations that produce the smooth input/output relationships for the original model; these are shown in the left-hand plots of Fig. 1. After the logic calculations, the overall response function is no longer continuous; this is illustrated in the right-hand plot of Fig. 1.

In general, such a system can be represented with a model consisting of S different states, where in each state a different continuous function $f_s(\boldsymbol{x})$, $s \in S$, establishes the map between the input and output parameters. In other words, the design space that is being explored consists of a finite number of design subspaces, which are accessed after activation of a logical switch or discriminant function that indicates which design space is being explored. The challenge is to find a strategy to metamodel the full design space accurately while modeling the continuous and logical components of the system in an efficient manner.

Because the approximation functions that are typically used for metamodeling are continuous, in fact often differentiable, their capability to approximate combined discrete/continuous responses is limited. In the following sections, we discuss three possible solution approaches to problems with discrete/continuous responses.

A. Combined Metamodel Strategy

The combined metamodel strategy attempts to build a single approximation to the curve shown in the right-hand plot of Fig. 1. An advantage of this strategy is that no separate metamodels need to be fit for each state or design subspace. Furthermore, the original system is treated as a black box and need not be modified to fit the metamodel. However, because of the discontinuity in the curve, the metamodel can never be completely successful using continuous approximating functions. The qualitative nature of the resulting approximation is shown in the plot for the combined metamodel strategy in Fig. 2. The dashed line represents the true response, whereas the solid line corresponds to the estimated response based on a single continuous approximation.

B. Metamodel Plus Original Logic

It is less demanding to fit a metamodel to the continuous, differentiable curves in the left-hand plots for the original system in Fig. 1. This approximation is shown in the left-hand plots of

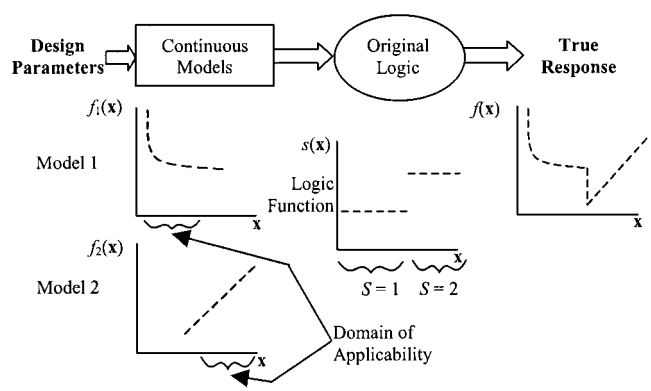

Fig. 1 Original system.

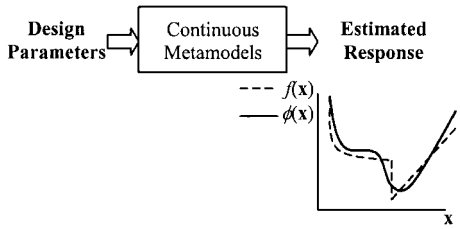

Fig. 2 Combined metamodel.

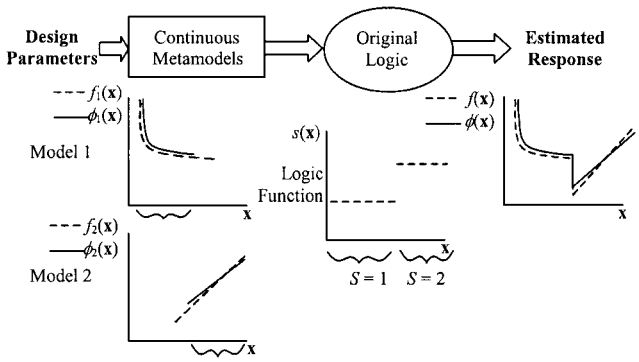

Fig. 3 Metamodels plus original logic.

the metamodel-plus-logic strategy in Fig. 3. If the logic calculations in the original model can be separated from the continuous calculations, a metamodel can be constructed for the continuous computations and interfaced with the original code used for logical calculations. Because the continuous (floating-point) calculations are often the computationally expensive ones, the computational advantage of a metamodel approximation can be retained in large part, in many cases.

Because the original model's logic code is then applied to the metamodel output, the resulting overall metamodel-based output, the right-hand plot in Fig. 3, looks very similar to the right-hand plot in Fig. 1. Although this seems to be a better option than the combined metamodel strategy, in many cases it will not be practical to modify the original system model, extract the logic component, and interface the logic unit with the metamodels. To clarify this, one must distinguish between cases where the original model logic is defined explicitly and implicitly. Venter et al.[32] discuss an application of an isotropic plate in which the original model logic is given by a simple geometric criterion. Thus, the logic component is known explicitly and can be extracted. State-selecting metamodels also have potential applications in variable complexity response surface modeling for structural design of aircraft wings,[33,34] composite frame design,[35] and computational fluid dynamics problems involving both laminar and turbulent flow regimes.[36] In Sec. IV.A, we consider the case of a desk lamp design where the original model logic is implicit and cannot be easily extracted.

C. State-Selecting Metamodel

The third modeling solution is to use a state-selecting metamodel as shown in Fig. 4, where continuous metamodels for each submodel are used with a logic metamodel to approximate the discontinuous function. In this strategy, an approximation that takes into account the different system states is constructed without modifying the original model. However, the success of this strategy relies heavily on the capability of the logic metamodel to discriminate the different states of the system based on the values of the design parameters. As illustrated in the right-hand plot of Fig. 4, a shift between the metamodel and the true function might occur at the point of discontinuity. This is not the case in the right-hand plot of Fig. 3 because it uses the original logic. This approach is shown with more detail in Fig. 5 and 6, where the metamodel calibration and prediction stages have been separated.

In the calibration stage (Fig. 5), the original system model to be approximated is evaluated at points $\boldsymbol{x}$ prescribed by the design

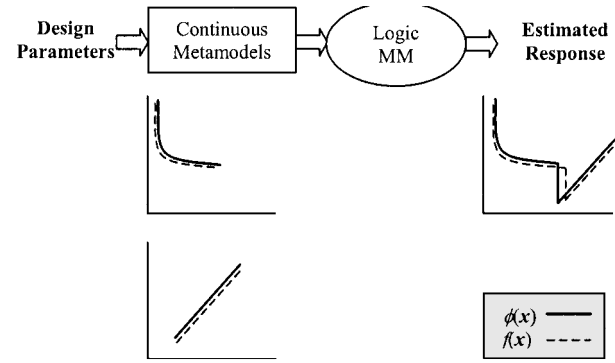

Fig. 4 State-selecting metamodel.

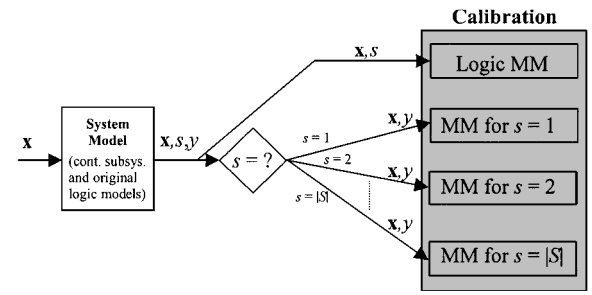

Fig. 5 Fitting a state-selecting metamodel.

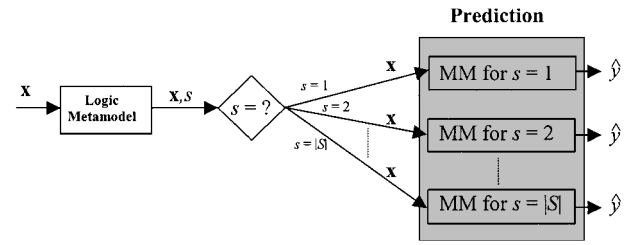

Fig. 6 Predicting with a state-selecting metamodel.

matrix $\boldsymbol{X}_D$, to obtain a response matrix $\boldsymbol{Y}_D$. This information is used to fit a logic metamodel and fit $|S|$ continuous metamodels, one for each of the system states.

Once the logic metamodel $\hat{s} = \phi_L(\boldsymbol{x})$ has been calibrated and the continuous-response metamodels for each state have been fit, they can be combined and used for prediction purposes at a new set of points $\boldsymbol{x}$, prescribed by the prediction matrix, $\boldsymbol{X}_P$ as shown in Fig. 6.

D. Logic Metamodels

At the core of a logic metamodel lies essentially a pattern recognition problem, which is also referred to as a classification or grouping problem, depending on the area of application. In our case, an observation is an input vector whose attributes are the parameter settings specified by an experimental design, and a population is equivalent to a state of the system. Given a sample set of observations from a population, the objective is to define a rule that allows us to decide to which population, that is, state s, a new observation, that is, $\boldsymbol{x}$, belongs.

Different methods can be used for classifying the states of a system. Discriminant analysis is a classical statistics method first introduced by Fisher.[37] It has been used extensively for identifying relationships between qualitative criterion variables and quantitative predictor variables. In other words, discriminant analysis is a procedure for identifying boundaries between the states of a system; a linear discriminant function uses a weighted combination of the predictor variable values to classify an object into one of the criterion groups. An extension of this method is quadratic discriminant analysis.

An alternative approach for discriminating between different states has emerged from the field of mathematical programming

(MP). One of the advantages of the MP approach over traditional discriminant analysis is that the normality and equality of covariance matrices assumptions are no longer required. The objective function minimizes the number of mis-classified observations when using the metamodel for prediction purposes, that is, the measure of performance is the minimization of the weighted sum of the misclassifications. Ignizio and Cavalier[38] describe the linear programming formulation for a two-group classification problem in detail. The complexity of the problem increases when classifying observations into more than two groups.

Artificial neural networks[39] (ANNs) are an attractive alternative classification tool. Based on the functioning of the human brain, ANNs are models that consist of a number of computational units, or neurons, which are organized into layers and interconnected through modifiable weights. During a training phase, a representative training data set is repeatedly presented to the ANN, which adjusts the weights by means of a training rule until the sum of squared errors between the network output (for classification: predicted state) and the desired output (actual state) is minimized. Once this has been achieved, the ANN may be used for prediction purposes. However, important concerns that affect the model prediction capabilities are how to avoid overfitting and coming up with an appropriate network structure. De Veaux et al.,[40] discuss the prediction properties of ANNs for response functions. We are unaware of any work identifying uncertainties in the classification boundary for ANNs used for discriminant analysis.

Other approaches such as logistic discriminant analysis, methods based on nearest neighbor algorithms, and likelihood-based nonparametric kernel smoothing exist. In this paper, we have chosen to use ANNs based on the success of preliminary results that are documented in Ref. 41; however, future work may involve the use of other methods, to evaluate their performances.

IV. Desk Lamp Design Example

In this section we present a description of the desk lamp design model that is used to evaluate the solution strategies proposed in the preceding section. The code allows the user to study many different aspects of the metamodeling process (see Ref. 42 for a more complete description). The design objective is to maximize the light quality on a particular area of a desk, such as on a sheet of paper or a book. The lamp and some of the key design and kinematics variables are shown in Fig. 7.

The lamp has two extendable arms and is described by the bulb characteristics and a kinematics model; external kinematic parameters specify the adjustment of the lamp in terms of rod angle $\boldsymbol{r}$, rod length $\boldsymbol{L}_2$, rod twist angle $\boldsymbol{t}$, vertical rotation $\boldsymbol{v}$, and bulb rotation $\boldsymbol{h}$.

The desk lamp model can be studied as a full system but also as a group of subsystems that presents different modeling challenges. In this study, the focus is on creating a metamodel for the light intensity calculation model because it is characterized by both discrete and continuous responses. Figure 8 shows the structure of the desk lamp design model and the location of the light intensity calculation model considered in this paper; additional details may be found in Ref. 42. Because the light intensity module is part of the optimization loop within the global model, a high-quality approximation is desired to minimize error propagation during optimization. The light intensity module computes the light ray position coordinates as well as the angle of incidence for each light ray that is emitted from the bulb filament. The model also generates a discrete response (state value), indicating whether a particular ray has been direct ($s = 1$), reflected ($s = 2$), or does not reached the table (whether direct or reflected).

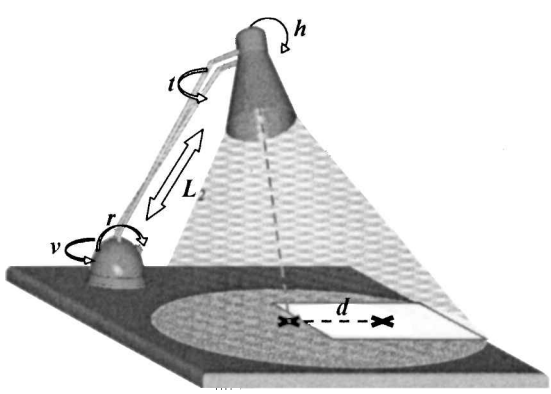

Fig. 7 Desk lamp model.

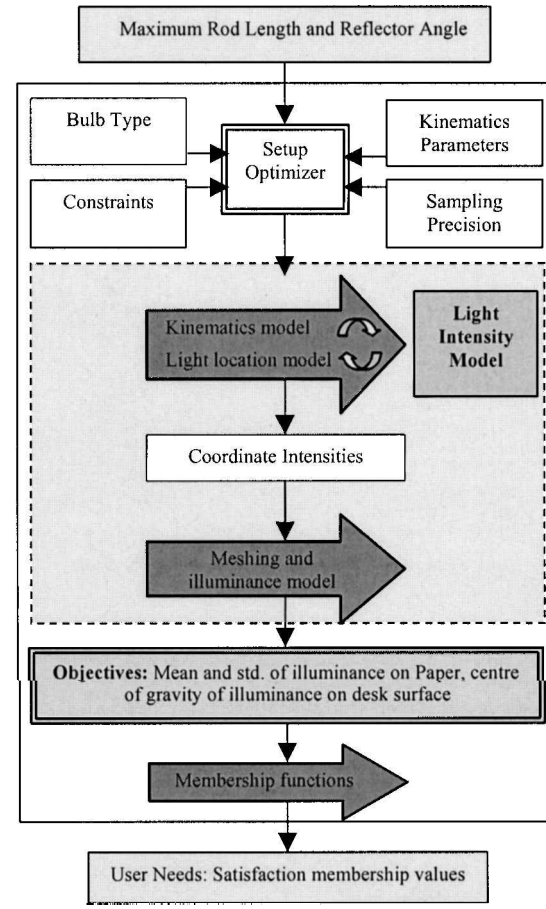

Fig. 8 Map of function requirements for analysis.

A. Logic Modeling Issues

The light intensity calculation model presents interesting issues for modeling combined discrete/continuous responses; a continuous function provides the origin and angle for each light ray, based on lamp geometry and kinematics parameters. As an example, imagine a light ray being emitted from the bulb filament at a particular angle; varying a design parameter such as reflector length will affect the trajectory of the light ray. Similarly, for a particular position of the lamp head, some emitted light rays may never reach the desk surface. To simplify the problem, the lamp has been positioned so that all emitted rays reach the table. Therefore, two distinct states are identified in this study: 1) $s = 1$, contact with the desk surface directly (without reflection from the lamp reflector); and 2) $s = 2$, contact with the desk surface after one reflection from the lamp reflector.

B. Combined Discrete/Continuous Response Modeling for the Desk Lamp Problem

The light intensity subsystem of the desk lamp model has a structure similar to the structure of the artificial example described in Sec. III; however, the main difference is that there are several inputs, such as lamp geometry and initial angle of light ray, and several responses (horizontal position, vertical position, angle). A piecewise continuous response, corresponding to the location and angle of incidence of the light ray on the surface of the desk is produced. In addition, the state response, either $s = 1$ or 2, is reported.

For the desk lamp example, only two of the three metamodeling strategies discussed in Sec. III could be used. These approaches are explained in more detail as follows.

1. Combined Metamodel Strategy

In this approach, we fit a metamodel to the light intensity module, using it as a black box, ignoring the presence of different states that specify whether light rays are reflected by the reflector or reach the desk surface directly.

2. Metamodel Plus Original Logic

In this approach, the combined discrete/continuous response vector from the original model is used to fit a metamodel. When using the metamodel for prediction, the predicted responses from the metamodel pass through the original model logic to determine to which state they belong. Although a promising strategy, it was not possible to extract the original logic from the light intensity model. The logic calculations are distributed throughout the code, making the elements extremely difficult to extract. The state identification is based on intermediate calculations and would have required building metamodels for a large number of intermediate variables.

3. State-Selecting Metamodel

In the state-selecting metamodel approach, the original lamp model is used to build three metamodels. A neural network logic metamodel is used. Several approaches are tested for constructing the continuous-response metamodels for each state (see Sec. V.A). For each prediction run, the logic metamodel produces a predicted state value and activates the corresponding continuous-response metamodel.

Calibrating the logic metamodel classifier is a critical part of this modeling strategy because it can be sensitive to the experimental design. Furthermore, the choice of the error tolerance for neural network training is also quite important. By error we mean the percentage of incorrectly classified states. A very small error tolerance may result in a longer model calibration; however, relaxing the error tolerance may lead to misclassification during prediction. Incorrectly predicted states generally result in large prediction errors. In the experiments described next, the error tolerance for the logic metamodel unit is set to 0, that is, during calibration, the logic metamodel is required to classify all experiment design points into their correct state. This does not guarantee, however, that it will classify all design points used in response prediction correctly.

V. Evaluation

A. Experiment Factors

A computational study using the desk lamp model is conducted to compare several strategies for metamodeling the desk lamp model. The factors considered are as follows: 1) the metamodeling approach, that is, combined and state selecting; 2) the continuous-response metamodel type, that is, quadratic polynomial (QP), kriging (KRG), and radial basis function (RBF); 3) the fitting experiment design, that is, approximate D-optimal latin hypercube (D-best), full factorial designs (FFD), factorial hypercube (FFLH), an HSS, and a random design with points selected from a multivariate uniform distribution (UNI); and 4) the number of fitting runs, that is, 25, 100, 225, and 400 (generated from factorial designs of size 5×5, 10×10, 15×15, and 20×20).

B. Performance Measures

The performance measures relate to the ability of the metamodels to reproduce the behavior of the lamp model over a range of values of the lamp design parameters. We have simplified the lamp design problem to vary only one internal kinematics variable, the ray angle a that specifies the angle at which a light ray is emitted from the bulb filament, and one controllable design variable, the length of the lamp reflector L_6. The lower and upper bounds for these variables are defined in Table 1.

To calculate our metamodel performance measures, the original lamp model is compared with the continuous metamodel and the state-selecting metamodel. We consider three measures of metamodel performance, computed over the lamp design parameter space given in Table 1. The calculation of average errors and maximum errors are approximated by calculating the errors on a 31×31 grid in the design space, yielding $\boldsymbol{n} = 961$ error values. The resulting measures are mean squared error (MSE)

$$\text{MSE} = \frac{1}{\boldsymbol{n}} \sum_{n=1}^{n} (\hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i)^2$$

mean absolute error (MAE)

$$\text{MAE} = \frac{1}{\boldsymbol{n}} \sum_{n=1}^{n} |\hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i|$$

and maximum absolute error (MAX)

$$\text{MAX} = \max_i |\hat{\boldsymbol{Y}}_i - \boldsymbol{Y}_i|$$

At each experimental run, the coordinates and the angle of incidence of a light ray are computed. To simplify presentation of the results, we only compare how well the solution strategies predict the true horizontal position of the light ray on the desk.

VI. Results

The performance assessment of the combined and state-selecting metamodeling approaches is summarized in Figs. 9–11. The graphs are response-scaled design plots,[42] where the circles on the cubes are scaled according to the size of the error with smaller circles indicating better performance. Notice also that gray shading is used to indicate condition numbers in the kriging metamodels, which are suspect. As the condition number gets smaller, the correlation matrix in the kriging metamodel becomes near singular, resulting in possibly large roundoff error during matrix inversion. Hence, for kriging models with condition numbers less than 10^{-14}, it is difficult to determine if the error results from inherently poor fit of the metamodel type or a degrading of the fit due to round-off error.

Table 1 Model parameter bounds

Model parameter	Low	High
Ray angle a, rad	0	1.4
Reflector length L_6, mm	60	100

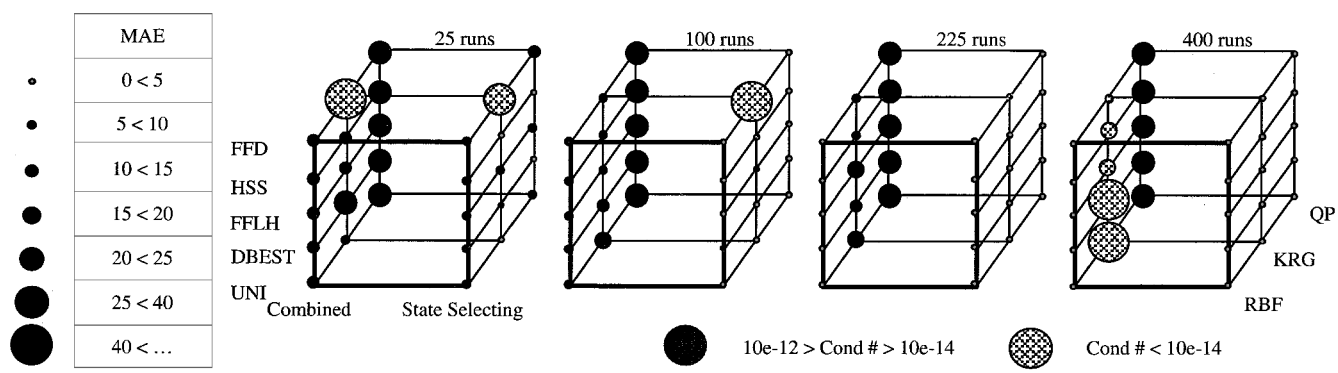

Fig. 9 Response-scaled design plot for MAE.

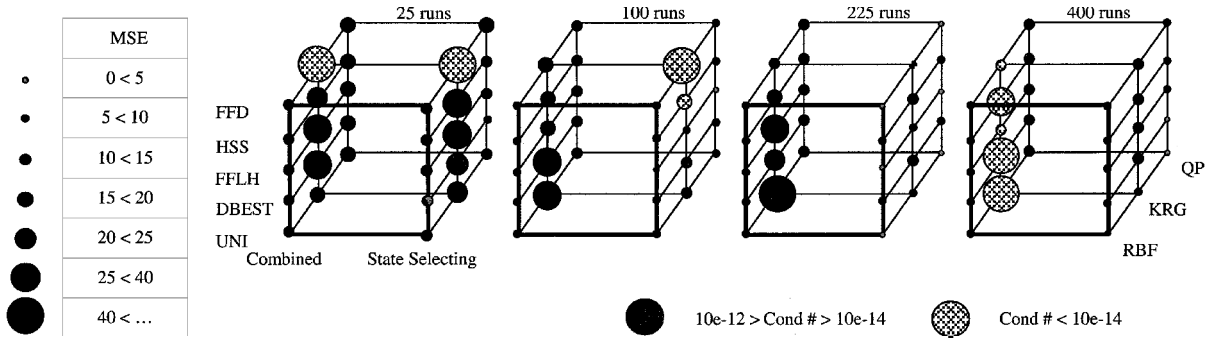

Fig. 10 Response-scaled design plot for MSE.

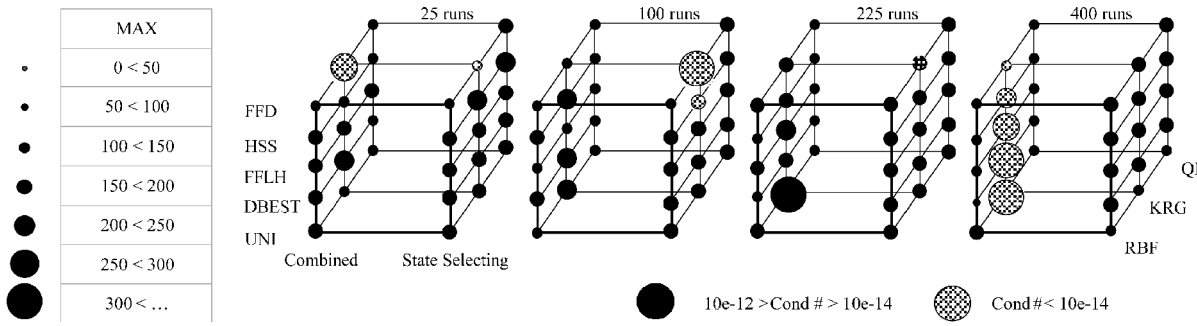

Fig. 11 Response-scaled design plot for MAX.

The three measures of performance, that is, MAE, MSE, and MAX, are shown in Figs. 9–11, respectively.

The plots for MSE and MAE show similar results: The state-selecting metamodels (on the right side of each cube) show generally lower errors than the combined metamodels. There is little difference vertically on each cube, indicating that, for this example, the choice of fitting experiment design has little impact on the accuracy of the metamodel. Moving from the leftmost cube to the rightmost cube corresponds to increasing the fitting experiment design from 25 to 400 fitting runs. There is little improvement in performance beyond a fitting experiment of 100 runs.

The RBF metamodels (front face of each cube) show small error for both combined and state-selecting strategies (left and right sides of each cube), for all but the leftmost cubes (only 25 fitting runs).

Finally, the Fig. 11 shows no improvement in the maximum error for the state-selecting metamodels when compared with the combined metamodels, and no improvement as the fitting experiment sample size is increased. This is because the maximum error occurs at points near the state transition boundary: A slight misclassification of the transition results in a large error for the state-selecting metamodel. Because the combined metamodel has a smooth transition, the size of this error is reduced by approximately one-half, although the abrupt feature in the true response has been lost. This phenomenon can be seen by visually comparing the maximum error in Fig. 2 with the maximum error in Fig. 4. Generally, we think that the approximation in Fig. 4 is superior to that in Fig. 2; however, the maximum error in Fig. 4 is actually larger.

For the state-selecting approach, we observe that using second-order polynomials as opposed to radial basis functions results in a slightly lower MSE. However, when the number of design points for fitting is increased, both MSE and MAE indicate that radial basis functions are better suited than second-order polynomials for this example.

In general, relatively high maximum errors occur in all methods; these errors occur in regions near the discontinuity. Furthermore, one expects the state-selecting metamodeling approach to incur in high maximum errors when the logic metamodel unit incorrectly classifies a design state. Unfortunately, even well-calibrated logic metamodel units will incur state prediction errors occasionally.

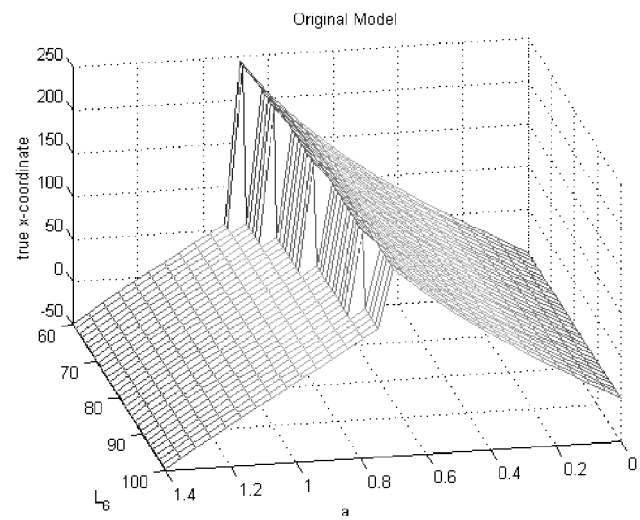

Fig. 12 True response surface.

Plots of the original response, and the estimated responses for the combined and state-selecting metamodels using polynomial regression, radial basis function, and spatial correlation metamodels are shown in Figs. 12–15, respectively. The metamodels in Figs. 12–15 are fitted using the 25-run uniform design, and the predictions are over a 31×31 point grid. In the true response in Fig. 12, it is easy to see the discontinuity that results when the ray is reflected vs direct. Notice also that the smooth fit of the combined radial basis function metamodel (Fig. 14, left) is comparable to Fig. 2. The metamodel has difficulties approximating the original model at the discontinuity.

The error plots for all three metamodels fitted for the 25-point uniform design are shown in Figs. 16 (combined) and 17 (state selecting). We observe how primarily small errors occur over a wide range along the discontinuity for the combined metamodel approach, whereas smaller regions characterize the state-selecting metamodeling approach. This again reflects the earlier discussion regarding the errors shown in Figs. 2 and 4.

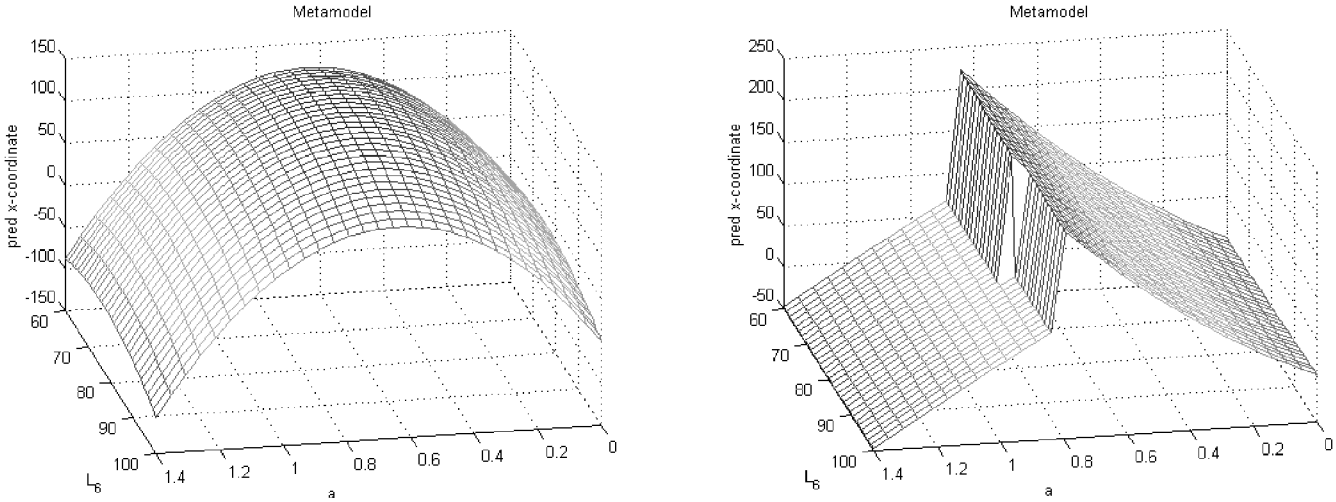

Fig. 13 Combined (left) and state-selecting (right) quadratic polynomial metamodels.

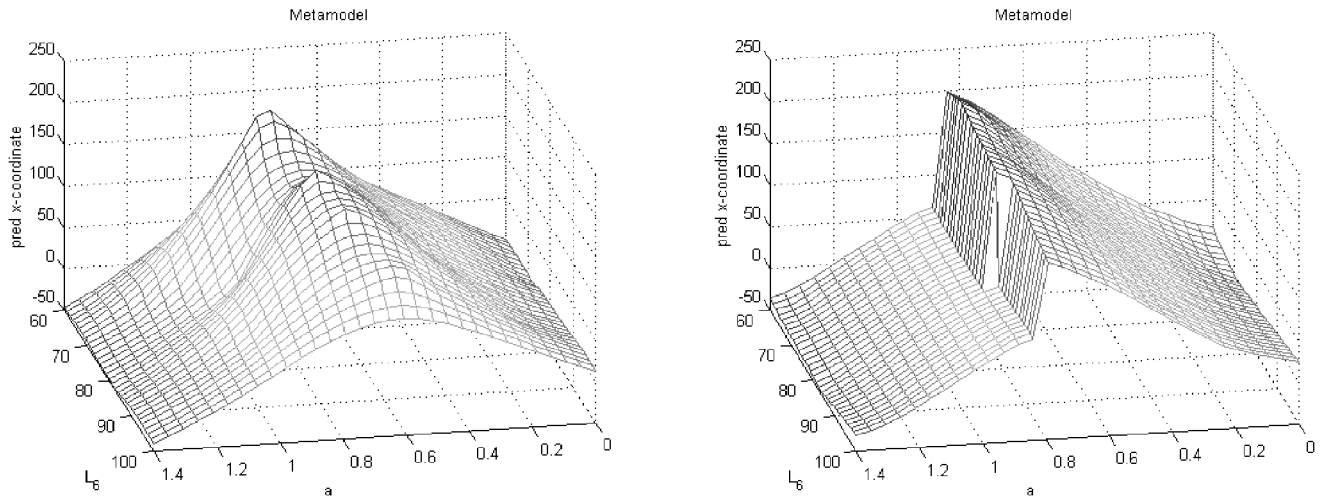

Fig. 14 Combined (left) and state-selecting (right) radial basis function metamodels.

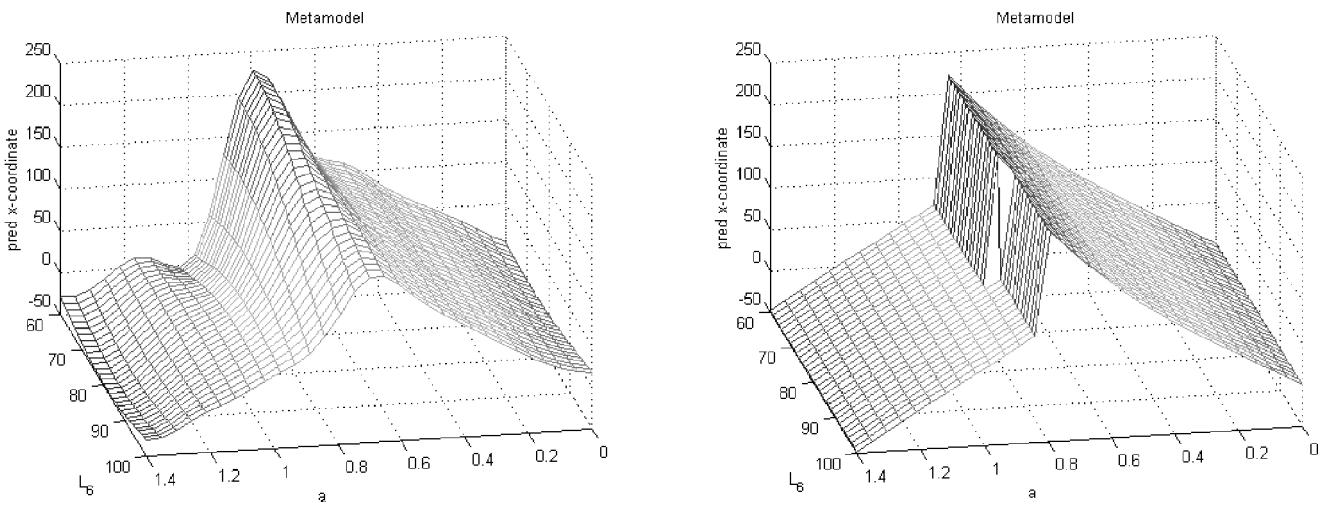

Fig. 15 Combined (left) and state-selecting (right) spatial correlation metamodels.

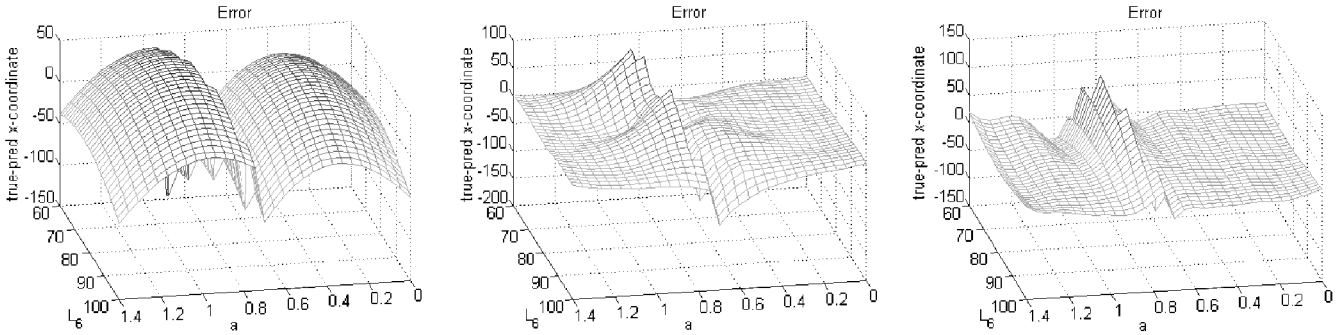

Fig. 16 Combined metamodel error plots: a) quadratic polynomial, b) radial basis function, and c) spatial correlation (kriging).

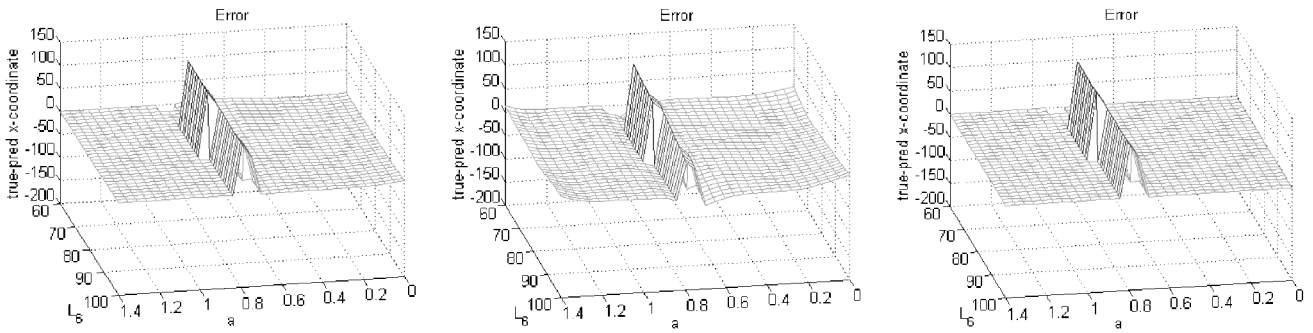

Fig. 17 State-selecting metamodel error plots: a) quadratic polynomial, b) radial basis function, and c) spatial correlation (kriging).

VII. Summary

In this paper, we have discussed the use of metamodels for approximating combined discrete/continuous responses. Three generic solution strategies are proposed, two of which are illustrated through a desk lamp design example. From our experiments, we conclude that the state-selecting metamodeling approach can outperform the combined metamodeling approach, and second-order polynomials with the state-selecting approach can perform well relative to the more flexible radial basis function and kriging models, especially when relatively few fitting runs are possible. Although the state-selecting approach is promising and preserves the discontinuous characteristic of the response function, misclassification by the logic metamodel can produce large errors near the discontinuity.

The work presented in this paper will be expanded in the future in several aspects. For instance, to fit different metamodels in a state-selecting metamodeling approach, sufficient design points need to be generated for each of the states. This raises questions such as how to generate additional points for different states, especially when these are not modeled explicitly by the original system.

Acknowledgments

This work is supported by the National Science Foundation (NSF) under NSF Grants DMI-9700040 and DMI-0084918 and by the Office of Naval Research (ONR) under ONR Grant N00014-00-G-0058. The cost of computer time is underwritten by the Laboratory for Intelligent Systems and Quality at Pennsylvania State University and the Laboratoire Productique Logistique at Ecole Centrale in Paris. We are indebted to the editor and anonymous referees, whose comments helped us improve the quality of this paper.

References

[1]Kleijnen, J. P. C., *Statistical Tools for Simulation Practitioners*, Marcel Dekker, New York, 1987, pp. 147–150.

[2]Simpson, T. W., Peplinski, J., Koch, P. N., and Allen, J. K., "On the Use of Statistics in Design and the Implications for Deterministic Computer Experiments," *Design Theory and Methodology—DTM'97*, American Society of Mechanical Engineers, 1997 (Paper DETC97/DTM-3881).

[3]Wilson, B., Cappelleri, D. J., Frecker, M. I., and Simpson, T. W., "Efficient Pareto Frontier Exploration Using Surrogate Approximations," *Optimization and Engineering* (to be published).

[4]Barthelemy, J.-F. M., and Haftka, R. T., "Approximation Concepts for Optimum Structural Design—A Review," *Structural Optimization*, Vol. 5, No. 1, 1993, pp. 129–144.

[5]Sobieszczanski-Sobieski, J., and Haftka, R. T., "Multidisciplinary Aerospace Design Optimization: Survey of Recent Developments," *Structural Optimization*, Vol. 14, No. 1, 1997, pp. 1–23.

[6]Yesilyurt, S., and Patera, A. T., "Surrogates for Numerical Simulations; Optimization of Eddy-Promoter Heat Exchangers," *Computer Methods in Applied Mechanics and Engineering*, Vol. 121, No. 1–4, 1995, pp. 231–257.

[7]Yesilyurt, S., Ghaddar, C. K., Cruz, M. E., and Patera, A. T., "Bayesian-Validated Surrogates for Noisy Computer Simulations: Application to Random Media," *SIAM Journal on Scientific and Statistical Computing*, Vol. 17, No. 4, 1996, pp. 973–992.

[8]Laslett, G. M., "Kriging and Splines: An Empirical Comparison of Their Predictive Performance in Some Applications," *Journal of the American Statistical Association*, Vol. 89, No. 426, 1994, pp. 391–400.

[9]Myers, R. H., and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley, New York, 1995, pp. 1–15.

[10]Toropov, V., van Keulen, F., Markine, V., and de Doer, H., "Refinements in the Multi-Point Approximation Method to Reduce the Effects of Noisy Structural Responses," *6th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Vol. 2, AIAA, Reston, VA, 1996, pp. 941–951.

[11]Rodriguez, J. F., Renaud, J. E., and Watson, L. T., "Trust Region Augmented Lagrangian Methods for Sequential Response Surface Approximation and Optimization," *Advances in Design Automation*, edited by D. Dutta, American Society of Mechanical Engineers, Fairfield, NJ, 1997.

[12]Renaud, J. E., and Gabriele, G. A., "Approximation in Nonhierarchic System Optimization," *AIAA Journal*, Vol. 32, No. 1, 1994, pp. 198–205.

[13]Renaud, J. E., and Gabriele, G. A., "Sequential Global Approximation in Non-Hierarchic System Decomposition and Optimization," *Advances in Design Automation—Design Automation and Design Optimization*, edited by G. Gabriele, Vol. 32-1, American Society of Mechanical Engineers, Fairfield, NJ, 1991, pp. 191–200.

[14]Wujek, B. A., Renaud, J. E., Batill, S. M., and Brockman, J. B., "Concurrent Subspace Optimization Using Design Variable Sharing in a Distributed Computing Environment," *Advances in Design Automation*, edited by S. Azarm, D. Dutta, H. Eschenauer, B. Gilmore, M. McCarthy, and M. Yoshimura, Vol. 82, American Society of Mechanical Engineers, Fairfield, NJ, 1995, pp. 181–188.

[15]Reddy, S. Y., "HIDER: A Methodology for Early-Stage Exploration of Design Space," *Advances in Design Automation*, edited by D. Dutta,

American Society of Mechanical Engineers, Fairfield, NJ, 1996 (Paper 96-DETC/DAC-1089).

[16]Barton, R. R., "Metamodels for Simulation Input-Output Relations," *Proceedings of the 1992 Winter Simulation Conference*, edited by J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, Inst. of Electrical and Electronics Engineers, New York, 1992, pp. 289–299.

[17]Barton, R. R., "Metamodeling: A State of the Art Review," *Proceedings of the 1994 Winter Simulation Conference*, edited by J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, Inst. of Electrical and Electronics Engineers, New York, 1994, pp. 237–244.

[18]Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F., "Comparison of Response Surface and Kriging Models for Multidisciplinary Design Optimization," *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Vol. 1, AIAA, Reston, VA, 1998, pp. 381–391.

[19]Hardy, R. L., "Multiquadratic Equations of Topography and Other Irregular Surfaces," *Journal of Geophysical Research*, Vol. 76, No. 8, 1971, pp. 1905–1915.

[20]Powell, M. J. D., "Radial Basis Functions for Multivariable Interpolation: A Review," *IMA Conference on Algorithms for the Approximation of Functions and Data*, edited by J. C. Mason and M. G. Cox, Oxford Univ. Press, London, 1987, pp. 143–167.

[21]Cressie, N. A. C., *Statistics for Spatial Data*, rev., Wiley, New York, 1993, p. 129.

[22]Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, No. 4, 1989, pp. 409–435.

[23]Koehler, J. R., and Owen, A. B., "Computer Experiments," *Handbook of Statistics*, edited by S. Ghosh and C. R. Rao, Elsevier Science, New York, 1996, pp. 261–308.

[24]Journel, A. G., and Huijbregts, C. J., *Mining Geostatistics*, Academic Press, New York, 1978, pp. 161–171.

[25]Goffe, W. L., Ferrier, G. D., and Rogers, J., "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, Vol. 60, No. 1–2, 1994, pp. 65–100 (source code available online at URL:http://netlib2.cs.utk.edu/opt/simann.f).

[26]Montgomery, D. C., *Design and Analysis of Experiments*, 5th ed., Wiley, New York, 2001, Chap. 5.

[27]McKay, M. D., Beckman, R. J., and Conover, W. J., "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, Vol. 21, No. 2, 1979, pp. 239–245.

[28]Salagame, R. R., and Barton, R. R., "Factorial Hypercube Designs for Spatial Correlation Regression," *Journal of Applied Statistics*, Vol. 24, No. 4, 1997, pp. 453–473.

[29]Kalagnanam, J. R., and Diwekar, U. M., "An Efficient Sampling Technique for Off-Line Quality Control," *Technometrics*, Vol. 39, No. 3, 1997, pp. 308–319.

[30]Fang, K.-T., and Wang, Y., *Number-Theoretic Methods in Statistics*, Chapman and Hall, New York, 1994, Chap. 1.

[31]Fang, K.-T., Lin, D. K. J., Winker, P., and Zhang, Y., "Uniform Design: Theory and Application," *Technometrics*, Vol. 42, No. 3, 2000, pp. 237–248.

[32]Venter, G., Haftka, R. T., and Starnes, J. H., Jr., "Construction of Response Surface Approximations for Design Optimization," *AIAA Journal*, Vol. 36, No. 12, 1998, pp. 2242–2249.

[33]Giunta, A. A., Balabanov, V., Kaufmann, M., Burgee, S., Grossman, B., Haftka, R. T., Mason, W. H., and Watson, L. T., "Variable-Complexity Response Surface Design of an HSCT Configuration," *Multidisciplinary Design Optimization: State of the Art*, edited by N. M. Alexandrov and M. Y. Hussaini, Society for Industrial and Applied Mathematics, Philadelphia, 1995, pp. 348–367.

[34]Kaufman, M., Balabanov, V., Burgee, S. L., Giunta, A. A., Grossman, B., Mason, W. H., and Watson, L. T., "Variable-Complexity Response Surface Approximations for Wing Structural Weight in HSCT Design," AIAA Paper 96-0089, Jan. 1996.

[35]Mason, B. H., Haftka, R. T., and Johnson, E. R., "Analysis and Design of Composite Channel Frames," *5th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Vol. 2, AIAA, Washington, DC, 1994, pp. 1023–1040.

[36]Narducci, R., Grossman, B., Valorani, M., Dadone, A., and Haftka, R. T., "Optimization Methods for Non-Smooth or Noisy Objective Functions in Fluid Design Problems," *12th AIAA Computational Fluid Dynamics Conference*, Vol. 1, AIAA, Washington, DC, 1995, pp. 21–32.

[37]Fisher, R. A., "The Use of Multiple Measurements in Taxonomy Problems," *Annals of Eugenics*, Vol. 7, 1936, pp. 179–188.

[38]Ignizio, J. P., and Cavalier, T. M., *Linear Programming*, Prentice–Hall, Englewood Cliffs, NJ, 1994, pp. 274–312.

[39]Haykin, S., *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, 1994, Chap. 1.

[40]DeVeaux, R. D., Schumi, J., Schweinsberg, J., and Ungar, L. H., "Prediction Intervals for Neural Networks via Nonlinear Regression," *Technometrics*, Vol. 40, No. 4, 1998, pp. 273–282.

[41]Meckesheimer, M., Barton, R. R., Limayem, F., and Yannou, B., "Metamodeling of Combined Discrete/Continuous Responses," *ASME 2000 Design Engineering Technical Conferences—Design Theory and Methodology Conference (DTM'00)*, edited by J. K. Allen, American Society of Mechanical Engineers, 2000 (Paper DETC2000/DTM-14573).

[42]Barton, R. R., Limayem, F., Meckesheimer, M., and Yannou, B., "Using Metamodels for Modeling the Propagation of Design Uncertainties," *5th International Conference on Concurrent Engineering (ICE'99)*, edited by N. Wogum, K.-D. Thoben, and K. S. Pawar, Centre for Concurrent Enterprising, The Hague, The Netherlands, 1999, pp. 521–528.

E. Livne
Associate Editor