

# A Response-Modeling Alternative to Surrogate Models for Support in Computational Analyses

Brian Rutherford  
Sandia National Laboratories  
Department 12323, MS0829  
Albuquerque NM 87185  
Tel: (1-505-844-3120)  
Fax: (1-505-844-9037)  
e-mail: bmruthe@sandia.gov

## Abstract

Surrogate models can perform a number of functions in support of a computational analysis. Through interpolation and extrapolation, these models can be used to address complex problems involving experimental design, analysis and prediction. Often, the objectives in a computational analysis involve the characterization of system performance based on some function of the response. We refer to this function as the performance function  $g(\mathbf{r}(\mathbf{x}))$  and consider applications where  $g$  is a scalar and  $\mathbf{r}(\mathbf{x})$  is a  $q$ -dimensional computer response depending on the  $p$ -dimensional inputs  $\mathbf{x}$ . The inputs  $\mathbf{x}$  may or may not be modeled probabilistically (with distribution  $F(\mathbf{x})$ ). The performance measure is computed as:

$$pm(\mathbf{r}) = g(\mathbf{r}(\mathbf{x})) \quad (\text{simple performance prediction problem});$$

$$pm(\mathbf{r}) = \int g(\mathbf{r}(\mathbf{x}))dF(\mathbf{x}) \quad (\text{average performance prediction problem});$$

$$pm(\mathbf{r}) = \min_{\mathbf{x}} g(\mathbf{r}(\mathbf{x})) \quad (\text{worst-case performance prediction problem});$$

$$pm(\mathbf{r}) = \{\mathbf{x}^* : g(\mathbf{r}(\mathbf{x}^*)) = \min_{\mathbf{x}} g(\mathbf{r}(\mathbf{x}))\} \quad (\text{engineering design or optimization problem}); \text{ or}$$

$$pm(\mathbf{r}) = \int I(g(\mathbf{r}(\mathbf{x})) \in R^*)dF(\mathbf{x}) \quad (\text{reliability prediction problem})$$

where,  $R^*$  is some subset of the response space and  $I$  is an indicator function taking on the value 1 when the enclosed expression is true. In general, we desire both a prediction for the performance measure and an estimate of prediction uncertainty.

Most surrogate modeling approaches are based on some kind of smoothing. Often, the surrogate models are not, themselves, members of the class of functions that are assumed to represent the actual computer response. Consequently, their use for some of the performance measures above might not be appropriate. Furthermore, the uncertainty associated with the surrogate model is typically specified in a point-wise fashion

depending on  $\mathbf{x}$ . This has the consequence of restricting uncertainty estimates to fairly simple performance measures. Together these drawbacks limit the utility of surrogate modeling in support of computational analysis.

One alternative is to construct an approximation to a probability measure  $G(\mathbf{r})$  for the computer response based on the available data. This approach will permit estimation in the general setting:

$$\text{Prob}(E | d) = \int_r I(E(\mathbf{r}))G(\mathbf{r})$$

where  $E(\mathbf{r})$  is an arbitrary event based on the computer response and  $d$  is the available data. Furthermore, one can use  $G(\mathbf{r})$  to calculate an induced distribution on the performance measure. For prediction problems where the performance measure is a scalar, the performance measure distribution  $F_{pm}(z)$  is determined by varying  $\mathbf{r}$  according to  $G(\mathbf{r})$ :

$$F_{pm}(z) = \int_r I(z \geq pm(\mathbf{r}))dG(\mathbf{r}).$$

For vector-valued or other, more involved, performance measures, alternative characterizations are possible.

The “response-modeling” approach provides an approximate probability measure using a discrete ensemble of “realizations”. A similar approach has been used in the geosciences to characterize two and three-dimensional regions based on limited spatial data (see Deutsch and Journel (1998), Chapter 5 and there references). Here, we make a deliberate effort to construct realizations that “span” the response space in the same sense as a Latin hypercube sample (McKay, Beckman, and Conover (1979)) produces a stratified sample for a random variable or vector. This is an alternative to relying on a random set of realizations as were used in the geoscience applications. The realizations are generated using a series of assumptions concerning the form of the response and appropriate methods of construction. Listed below is a brief summary of the process followed.

A basic model of the form:

$$r_i(\mathbf{x}) = P_i(\mathbf{x}) + \varepsilon_i(\mathbf{x}) \quad \text{is assumed for the } i\text{th realization.}$$

Here,  $P_i(\mathbf{x})$  is a polynomial in  $\mathbf{x}$  and  $\varepsilon_i(\mathbf{x})$  is a random function over  $\mathbf{x}$  defined using a stationary spatial covariance function. Steps (1) through (3) below describe construction of the polynomial component; steps (4), (5) and (6) describe construction of the random function term; step (7) completes the process.

- 1) Evaluate main effects, quadratic terms, and interactions, where possible, using the initial data, settling on an appropriate polynomial model.
- 2) Estimate the regression coefficients and their covariance structure.

- 3) Generate  $k$  sets (one set for each realization) of these coefficients (assuming a multivariate normal for their joint distribution) using a Latin Hypercube design with the appropriate correlation structure imposed on the sets of coefficients using rank correlation procedures described in Iman and Conover (1982).
- 4) The residuals to the fitted surface are transformed using the ‘Normal-scores transform’ as recommended in Deutsch and Journel (1998), see Patel (1982), page 217.
- 5) The transformed residuals are then used to estimate parameters of the exponential product spatial covariance function:

$$C(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^p C(|x_i - x'_i|) = \prod_{i=1}^p e^{-\phi_i |x_i - x'_i|} \quad \text{for any } \mathbf{x} \text{ and } \mathbf{x}'$$

where  $|\cdot|$  is the Euclidian norm and  $p$  is the input dimension. This covariance structure is fairly flexible and has been used with success in modeling computational data (see Sacks, Schiller, and Welch (1989)). More flexible spatial covariance functions are available and might be recommended for problems where more data are available. See Cressie (1991), Chapter 2 for a thorough discussion and for a more recent review, see O’Connell and Wolfinger (1997). A maximum likelihood estimate is used in most of our applications for estimating  $\phi$ .

- 6) The sequential-Gaussian conditional simulation procedure, Deutsch and Journel (1998) is used to generate the random function component. The algorithm generates a response surface over the grid in transformed space and then back-transforms the values according to a set of tables constructed during the transformation. The conditioning data are the transformed residuals to the polynomial surface. These are recomputed for each polynomial in the ensemble.
- 7) The back-transformed random function term is added to the polynomial to complete the realization.

Using the response models to approximate the performance measure, we compute the probability for the arbitrary event  $E$  using:

$$\text{Prob}(E | d) = \left( \frac{1}{k} \right) \sum_{i=1}^k I(E(r_i))$$

where  $r_i$  is the  $i$ th of  $k$  realizations.

The primary use of response models in engineering applications thus far has been for computer experimental design. The information contained in the response model concerning “likely” values at not-yet-evaluated input locations is used to predict the relative value of performing a simulation using those inputs. Briefly, the following algorithm is pursued:

- 1) A response-model approximation is made to a probability measure over the response space based on the initial data. The distribution of the performance measure is then computed using this approximation.

- 2) Candidate designs are selected through a random search optimization algorithm or through a grid of single points (if the process is pursued sequentially) and step (3) is applied to each.
- 3) Response values at the candidate design input locations are obtained (iteratively) according to the distribution in (1) above, and are combined with the initial data. The process in (1) is then repeated for each set of augmented data so constructed. The expected change in variability of the performance measure is calculated.
- 4) The candidate design that indicates the largest expected decrease in performance measure variability is selected and the actual computational simulation experiments are performed.

When the goal of the computer experiments is to reduce variability (uncertainty) in the performance measure distribution, the algorithm decomposes variability in the initial distribution into components representing the variability in the expected response and the expected variability remaining (see Parzen (1962)). The best design is selected as the design that minimizes this later quantity. Alternative optimization criteria are available for other problems like the engineering design problem where the solution is in terms of the input locations.

This presentation includes several examples of the use of response-modeling for experimental design, analysis and prediction. The examples illustrate the general applicability of this approach.

## **References**

- Cressie, N. (1991), "Statistics for Spatial Data," Wiley and Sons Inc., New York.
- Deutsch, C. V. and Journel, A. G. (1998), "GSLIB Geostatistical Software Library and Users Guide Second Edition," Oxford University Press, New York.
- Iman, R. L. and Conover, W. J. (1982), "A Distribution-Free Approach to Introducing Rank Correlation Among Input Variables," *Communications in Statistics, Part B - Simulation & Computation*, 11(3), pp. 311-334.
- McKay, M. D., Beckman, R. J. and Conover, W. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, Vol. 21, No. 2, pp. 239-245.
- O'Connell, M. A. and Wolfinger, R. D. (1997), "Spatial Regression Models, Response Surfaces, and Process Optimization," *Journal of Computational and Graphical Statistics*, Vol. 6, No. 2, pp. 224-241.
- Parzen, E. (1962), "Stochastic Processes," Holden Day, San Francisco.
- Patel, J. K., (1982), "Handbook of the Normal Distribution," Marcel Dekker, Inc., New York.

Sacks, J., Schiller, S. B. and Welch, W. J. (1989), "Designs for Computer Experiments," *Technometrics*, Vol. 31, No. 1, pp. 41-47.