THE SPATIAL CORRELATION FUNCTION APPROACH TO RESPONSE SURFACE ESTIMATION

Toby J. Mitchell
Max D. Morris

Oak Ridge National Laboratory
P.O. Box 2008, Building 6012, MS 6367
Oak Ridge, Tennessee 37830, U.S.A.

ABSTRACT

This is an expository paper which discusses an alternative to conventional response surface methodology for use in simulation experiments where the objective is to express an output variable (response) as a function of several input variables. The method is Bayesian in the sense that uncertainty about the true response function y is expressed by the random function Y, defined on the region of interest X in the space of the input parameters. If Y is Gaussian, straightforward formulas exist for updating Y given observations of $y(x^1), y(x^2), \cdots, y(x^n)$, which are available from n simulation runs at different settings $(x^i \in X)$ of the input parameters. The posterior mean of $Y(x)$, viewed as a function of x, serves as the estimated response function $\hat{y}$. The method is driven primarily by means of a chosen *spatial correlation function (SCF)*, which defines the prior correlation between the responses at any two points in the space of the input parameters. Once the SCF is chosen, the method is naturally adaptive -- $\hat{y}$ becomes more subtle and complex as more simulation runs are made, with no intervention required to add terms to a parametric model. Although much of the focus of this paper is on deterministic simulations, where we have had most of our experience, we shall show how modifications can be made to handle "random" responses. Some examples are discussed to illustrate the ideas and the nature of the results.

1 INTRODUCTION

Here we regard a simulation model as a computer program that maps a vector of inputs (system parameters or independent variables) into a vector of outputs. If one is interested in exploring the relationship between a relatively small set of input parameters $(x = x_1, \cdots, x_k)$, with values in some k-dimensional region of interest X, and a single "response" y that is computed from the output data, then we can view y as a function $y(x)$ whose values are determined by the model. This function may be deterministic or random, the latter occurring when some of the input variables are produced by random number generators. Our concern here will be primarily with the deterministic case; in Section 6, we shall indicate how the method can be easily extended to random outputs.

Response surface methodology (RSM) is based on the approximation of y over part of its domain X by a relatively simple function, usually a polynomial of degree one or two. Although RSM has various applications, its use in simulation seems to be primarily for the purpose of optimizing y over X. (See, e.g., Kleijnen 1975, pp. 79-80.) One approaches the problem of finding the maximum response, say, in the same way that a myopic person might climb a hill. A local approximation (usually linear) is made, and the direction of steepest ascent is determined. After following this path until it no longer improves the response, another approximation is made, and the climb continues. Eventually, a more elaborate approximation (e.g., a quadratic function) is used, and the "final" location of the optimum is estimated. For a comprehensive text on RSM and its use in empirical model building, see Box and Draper (1987).

In spite of the availability of RSM, optimization problems in simulation are not always tractable. Indeed, Bratley, Fox, and Schrage (1983, p. 28) assert that "it is difficult or impossible to optimize systematically via simulation", and they seem pessimistic about the usefulness of RSM in this context. They are concerned that the postulated form of the function may be grossly inaccurate and that, if it has more than about three

parameters, it may take far too many runs to estimate them well, particularly when y is random. These concerns, although valid, can often be mitigated by using RSM strategies that involve only local approximation in low-dimensional subspaces.

The purpose of this paper is to describe another method of response surface estimation, one which is an interpolator in the deterministic case, and which adapts well to functions that are not well represented by polynomials. In this approach, the function y is not modeled directly. Through the choice of a single "spatial correlation function", a model term is automatically included for each data point, so the choice of which terms to include in a polynomial model is avoided. The method can be applied in as many dimensions as desired, with any number of runs, and any configuration of the design points (points in X at which the simulation model is run). Of course, the *performance* of the method depends on dimension, sample size, design, and correlation function, as one would expect. Because of their flexibility, the response surfaces that are produced by this method are more useful as a global approximations to y than are polynomials, which are based on Taylor expansions.

Versions of this methodology have been in use for some time in spatial statistics, especially the "kriging" methods used in geostatistics. (See Cressie (1991) for a comprehensive text on spatial statistics.) Only recently has a literature on applications to simulation experiments begun to emerge, beginning with Sacks, Schiller, and Welch (1989), Sacks et al. (1989), and Currin et al. (1991). These papers provide extensive references to related work.

Because of its mixed ancestry, there are several different concepts and philosophies that have been used to support this method. Here we approach it from a Bayesian standpoint, as in Currin et al. (1991), while noting that classical frequentist concepts, as in kriging, for example, lead to a similar methodology (Sacks, et al. 1989). The cornerstone of both approaches is the use of spatial correlation functions, or more generally, spatial covariance functions. For the sake of neutrality and convenience, we shall use the term "SCF method" to refer generically to both approaches. It is the purpose of this paper to review a basic version of the SCF methodology and to describe some of its recent applications in simulation experiments.

2 METHOD

2.1 Random Functions

This method is based on the representation of

uncertainty about y by a random function Y (also called a "random field" or "spatial stochastic process"). Consider, for example, the simple linear model,

$$y=\beta_0+\beta_1x_1+\beta_2x_2 \qquad (2.1)$$

where we assign a 3-dimensional probability density function to the vector β of coefficients to represent our uncertainty about its value. Then the function (2.1) is random, and we could simulate realizations ("sample paths") of it by drawing from the distribution of β and substituting into (2.1). Now if we observed that $y = y^0$ at the point $x = x^0 = (x_1^0, x_2^0)$, we would want to disregard all the sample paths that did not pass through the point (x^0, y^0); what remains is a set of sample paths from the *posterior random function*. This represents our uncertainty about the function after observing it at one point. By the time we have observed it at three points (that do not all lie on the same line), there is only one sample path left. We now know the function exactly.

If the distribution of β is a multivariate normal, then the random function in (2.1) is a Gaussian random function, or Gaussian process. However, there is a problem with using random functions like (2.1) in practice. As we have seen, knowledge of the function value at three points implies knowledge of the function everywhere; this is not realistic.

In the SCF approach, one does not begin with a parametric model, but with the concept that the prior correlation between the values of Y at any two points should depend on the spatial relationship between the points. In particular, we shall require the correlation to depend only on the *difference* between the points. The prior mean and variance of Y are usually very simple, to express a form of prior ignorance; here we will always take $E(Y(x)) = \mu$ and $V(Y(x)) = \sigma^2$, no matter where x is. Thus, for any two points x and w in X,

$$Cov(Y(x),Y(w)) = \sigma^2 R(w-x),$$

where the *spatial correlation function* R usually depends on a set of parameters θ, whose values, like μ and σ, are estimated using the data.

2.2 Correlation Functions

Unlike conventional RSM, where the critical choice is a parametric model for the mean of Y, and there are only a few "standard" choices (e.g., linear and quadratic polynomials), here the critical choice is the correlation function, and none have really emerged as "standard". It is important to note that the choice of correlation function

is not completely arbitrary. It must be *valid* in the sense that the variance of any linear combination of the responses at any set of points in X must not be negative.

A common practice in choosing a correlation function in several dimensions is to adopt the "product correlation rule", i.e.

$$R(w-x) = \prod_{j=1}^{k} R_j(w_j - x_j),$$

where R_j is a valid correlation function in one dimension, and usually depends on one or more parameters. The choice of R_j is usually made based on considerations of smoothness, where smoothness reflects the number of times the random function Y is differentiable.

Table 1 gives some useful choices for $R_j(d_j)$, where $d_j = w_j - x_j$ and $\theta_j > 0$. (The subscript j, which indexes the inputs, is omitted from R, d, and θ for simplicity.) All of the correlation functions shown decline to 0 as $|d|$ increases. We omit processes without derivatives here because they are less well suited for simulation models, which (in our experience at least) tend to be relatively smooth functions of the inputs.

Table 1: Some Useful Correlation Functions

	R(d)	#derivs
(1)	$\exp(-(\theta|d|)^2)$	∞
(2)	$\exp(-\theta|d|)[1+\theta|d|+\theta^2|d|^2/3]$	2
(3)	$1.0 - 6(\theta|d|)^2 + 6(\theta|d|)^3, \theta|d|<0.5$ $2(1 - \theta|d|)^3, 0.5\le\theta|d|<1$ $0, \theta|d|\ge1$	1
(4)	$\exp(-\theta|d|)[1+\theta|d|]$	1

The Gaussian correlation (1) yields sample paths that are infinitely differentiable. Although it is sometimes criticized for being "too smooth", we have found its performance to be satisfactory enough to warrant its use as an automatic first choice in routine applications. (It may not do so well in cases where some of the data points are very close together, but this can be easily avoided through choice of design.) Correlations (2) and (4) are from the Matern (1986) class, which is useful because it includes random functions having any desired degree of smoothness. Correlation (3) is included because it produces response surfaces that are piecewise cubic splines

in any coordinate and because its range of influence (i.e., the range over which it is positive), can be controlled by θ.

Note that the parameters θ in Table 1 all appear as multipliers of d, and so come into play as scaling parameters. We generally allow each coordinate (input) to have its own scaling parameter. An exception occurs when a single input is represented by a point in more than one dimension (like "location" on a two-dimensional surface); then the meaning of the coordinate axes for representing that point may be arbitrary. In this case, one might require the correlation between the responses at two locations (with the other variables fixed) to depend, for example, on the Euclidean distance between them. The correlation functions (1), (2), and (4) in Table 1 can be used for this purpose, with $|d|$ taken as Euclidean distance.

2.3 Response Surface Estimation

Like the prior process, the posterior process after observation of y at n design points $x^1, x^2, \cdots, x^n$ is Gaussian; its mean is given by:

$$\hat{y}(x) = \mu + r'(x)C^{-1}(y - \mu J), \qquad (2.2)$$

where

y is the $n \times 1$ vector of observed responses, J is the $n \times 1$ vector of 1's, $C = \{R(x^i-x^j)\}$ is the $n \times n$ matrix of correlations among the design points, and $r = \{R(x-x^i)\}$ is the $n \times 1$ vector of correlations between x and the design points.

It is customary to substitute maximum likelihood estimates for the parameters μ, σ, and θ in (2.2), recalling that θ appears both in r and in C. The likelihood is given by

$$L = (2\pi)^{-n/2} |\tilde{C}|^{-1/2} \exp(-\tfrac{1}{2}(y-\mu J)'\tilde{C}^{-1}(y-\mu J)),(2.3)$$

where $\tilde{C} = \sigma^2 C$ is the matrix of *covariances* among the responses at the design points and depends on σ and θ. Maximization of (2.3) with respect to μ, σ, and θ must be done numerically. This is by far the most (computer) time-consuming part of the method, and its biggest disadvantage when compared with polynomial fitting by least squares. Some gain is possible by noting that, given θ, well known expressions for μ and σ^2 exist. (See, e.g., Currin et al. 1991.)

Formulas for the posterior variances and covariances among the responses at arbitrary points in X are given, e.g., by Currin et al. (1991). It should be noted that

these are for *fixed* values of the parameters, and therefore don't take into account the variability of the maximum likelihood estimates. Sacks et al. (1989) give a formula for the mean squared error (MSE) of $\hat{y}$ (analogous to the posterior variance of Y), which does at least account for the variability in $\hat{\mu}$.

Here, we shall restrict attention to the posterior mean function (2.2) which serves as the response surface estimator. Note that it is a linear combination of terms of the form

$$r_i(x) = R(x-x^i) = \prod_{j=1}^{k} R(x_j-x_j^i).$$

Thus, for example, the cubic correlation function (3) in Table 1 produces cubic splines.

3 DESIGN

There are various formal design criteria that can be used to reduce the problem of choosing the design points $\{x^i\}$ to a mathematical optimization problem. We will not discuss these here; see Sacks, Schiller, and Welch (1989), Sacks et al. (1989), and Currin et al. (1991) for some examples.

Research is currently going on in this area, and there is not yet a generally accepted design strategy. It appears that the design issue is not as critical for the SCF method as it is for fitting polynomials by least squares, but we know of no research that has examined this question thoroughly. (There are some anecdotal examples in Currin et al. 1991.)

More often than not, current practice is to select a *Latin hypercube design* because such designs are easily constructed and seem to spread the design points reasonably well throughout the design space. In the simplest implementation, an n-run Latin hypercube design in k parameters is obtained by writing down n equally spaced levels of each input as a column in an n×k matrix, and then permuting each column randomly. This is similar to the *lattice sampling designs* of Patterson (1954). A more general construction was proposed by McKay, Conover, and Beckman (1979), who were concerned with the estimation of the distribution of the response given a probability distribution on the inputs. Various refinements have been made since. Latin hypercube designs have become very popular for certain kinds of simulation activities, well beyond their original purpose. They are the designs on which several of the published applications of the SCF method are based.

4 APPLICATIONS

4.1 Screening

Since the SCF approach provides global approximations of a response surface, a global evaluation of the importance of the input parameters is possible. Morris (1991) suggested the mean and variance of the "elementary effect"

$$g_j = y(x_1, \cdots, x_j+\delta, \cdots, x_k) - y(x)$$

over randomly selected points x in X as good indicators of the importance of the jth input. A small mean and variance of g_j implies that the jth input has little effect on the response. A large mean coupled with small variance implies a linear effect, independent of the other inputs. A large variance implies either a nonlinear independent effect or interactions with other inputs. Morris's application was directly to the responses from a heat transfer model, but the same quantities can be easily computed using an estimated response surface.

In an alternative approach, Sacks et al. (1989) suggested plotting main effect and two-factor interaction functions. The main effect function for input u is

$$F_u(x_u) = E(y \mid x_u) - E(y),$$

where E(y) is the average response over X, and $E(y \mid x_u)$ is the average response when x_u is fixed. The fitted response $\hat{y}$ is used for y here. If F_u is approximately constant over the range of x_u this is a good indication that x_u has little effect. The form of F_u is often informative also, especially if x_u does not interact with other inputs.

The two-factor interaction function for inputs u and v is

$$F_{uv}(x_u, x_v) = E(y \mid x_u, x_v) - F_u(x_u) - F_v(x_v) - E(y).$$

Again, the form of this function is useful if other inputs do not interact with x_u and x_v.

Examples of these plots are shown in Bernardo et al. (1992) and in Welch et al. (1992).

Higher order interactions are obviously more difficult to represent in this way, and they may be quite numerous. An alternative way to assess main effects and interactions is to make predictions at the corners of a 2^k cube and then analyze the results as one would a 2^k factorial experiment (Kleijnen 1975, Ch. 4.3.). An example of this is described by Mitchell and Morris (1992) in connection with a deterministic model for a

combustion process.

There is also information in the maximum likelihood values of the scaling parameters. Setting $\theta_j \approx 0$ implies a prior belief that differences in the jth coordinate will have very little effect on the predictions; when the maximum likelihood procedure chooses such a low value for θ_j, it is an indication that the jth input is unimportant. Welch et al. (1992) exploit this phenomenon in their approach to screening and prediction by the SCF method.

4.2 Optimization

Bernardo et al. (1992) describe the optimization of an integrated circuit design using a CAD circuit design simulator. They used the SCF approach to approximate the output current y as a function of 11 design parameters (resistances and transistor sizes) and two uncontrollable "noise" parameters (voltage and temperature). The goal of the experiment was to find values of the design parameters such that y, maximized over the noise parameters, would be within 3% of 1 mA. The initial design consisted of 576 points in the 13 dimensional space. This did not involve as much computing as one might suppose, since at each of the 48 combinations of the design parameters, which were chosen according to a Latin hypercube design, the responses at the 12 combinations of the noise parameters could be produced relatively quickly. Also, this design structure was exploited to greatly reduce the cost of maximizing the likelihood in (2.3). After determining the estimate $\hat{y}$ of the response surface, and examining main effect and interaction plots (see Section 4.1), the investigators identified the two most influential design inputs and restricted further attention to those, plus two more that were of special engineering interest, plus the two noise inputs. A second simulation experiment was run, a 24-run Latin hypercube in the four remaining design parameters, restricted to a region where the predicted response was near the target. Again the responses at the 12 combinations of the noise parameters were produced in each case. After estimating a new response surface from these 288 data points, and validating it at 20 randomly selected sites in the reduced region, the investigators searched the surface numerically and found two points in the design parameter space that minimized their loss function, which had been chosen to reflect the deviation from target.

Some interesting features of this approach are:

1. The optimization strategy was an "outside-in" approach, where the response was approximated globally first, then attention was focused on a reduced region.

2. The response y and not the loss was approximated first. Then the loss was optimized using the *predicted* response surface for y.

3. The effect of the noise inputs (voltage and temperature) was modeled directly, rather than by simulating them randomly and observing the distribution of the response values. If the response surface approximation is adequate, simulation experiments can be performed, with random inputs for these parameters, but using the approximation (rather than the circuit simulator) to evaluate the response.

4.3 Inverse Problems

Here we describe a simulation experiment on a model of groundwater flow, in which the purpose of the experiment was to identify the values of the input parameters that lead to physically untenable results. This problem came about as the result of an uncertainty analysis, in which randomly chosen values of seven inputs (in this case six conductivities and a "boundary flux" coefficient) were submitted to the model, yielding distributions for various response values. One response of interest was the pressure at a particular location; this is y in this example. One hundred runs of the simulator were made, and the appropriate plot of the cumulative distribution function of y was produced. It was observed that in nearly half the cases, the pressure was higher than y*, the physical upper bound. Attention then turned to trying to determine which combinations of the seven inputs lead to such unrealistic response values.

We first used the SCF method to fit a response surface (y as a function of the seven inputs) to the data that had already been generated. The Gaussian correlation function ((1) in Table 1) was used throughout, with scaling parameters selected by maximum likelihood. The fitted response $\hat{y}$ was then evaluated at 4000 randomly chosen points in X; of the points where $\hat{y}$ was within 10% of the critical pressure y*, 20 were chosen randomly as a set of validation points. The rationale was to focus on the boundary of the surface that separates the unrealistic results from the realistic ones. The model was then run at these 20 points, and comparisons were made with the predictions there. The root mean squared error (after taking natural logs of pressure) was 0.24, which corresponds roughly to a 24% error. Although this is not bad in view of the dimension of X and the fact that the range of y over X is more than two orders of

magnitude, we did not feel it was adequate for our purpose. (By way of comparison, the root mean squared error at the 20 validation points, for a linear polynomial fit by least squares to the original data, was 0.52; surprisingly the corresponding error for a quadratic polynomial was 0.67.)

We then added the results from the validation runs to the data base, re-estimated the response surface, and chose 20 new validation runs in a similar way. The results were disappointing: the root mean squared error of prediction at the 20 new validation points was 0.28, a bit larger than before. Again, we plodded ahead, adding the validation points to our data base, re-estimating the response surface, and choosing 20 new validation points. The results this time were much better: all the predictions at the validation sites were within 21% of the true values, and most of them were within 6%. The root mean squared error (after taking natural logs of pressure) was 0.09. The validation points were then added to the data base, and the response surface was re-estimated. This can now be used to screen out unrealistic combinations of inputs in future uncertainty analyses with this simulation model.

5 ADDITIVE MODELS

When X has more than a few (three or four) dimensions, good global approximation to y in an affordable number of runs is likely to be impossible, unless there is some simplicity of structure in y. One approach is to seek an *additive model* for y. In the SCF approach, for example, we can let

$$Y = \sum_1^k Y_j(x_j), \qquad (5.1)$$

where the Y_j's are one-dimensional Gaussian processes, independent of one another, with $E(Y_j(x_j)) = \mu_j$ and

$$Cov (Y_j(x_j), Y_j(w_j)) = \sigma_j^2 R_j(w_j - x_j).$$

Then Y is a Gaussian process, with

$$\mu = \sum_1^k \mu_j$$

and

$$Cov(Y(x), Y(w)) = \sum_1^k \sigma_j^2 R_j(w_j - x_j).$$

The estimated response surface is again given by (2.2), but with C replaced by $\tilde{C}$, the matrix of *covariances* among the responses at the design points, and r replaced by $\check{r}$, the vector of covariances (rather than correlations) between x and the design points. The likelihood (2.3) must be maximized over the σ_j's, as well as over μ and θ.

The individual components of the response surface can be recovered (up to an additive constant) by evaluating the posterior mean of $Y_j(x_j)$, which is given by (2.2) with C replaced by $\tilde{C}$ and r replaced by the vector of covariances between $Y_j(x_j)$ and the $Y(x^i)$'s:

$$r_i(x) = Cov (Y_j(x_j), Y(x^i)) = \sigma_j^2 R_j(x_j - x_j^i).$$

Technically, the first μ in (2.2) should also be replaced by μ_j, but this is indeterminate (i.e., the likelihood is maximized for any set of μ_j's that sum to $\hat{\mu}$).

Of course, (5.1) can be generalized so that the additive components are functions of more than one input.

6 RANDOM RESPONSES

In many simulation models, the response y is random. Under the SCF approach, one assumes the following additive model:

$$Y_i = Z(x^i) + \varepsilon_i,$$

where x^i is the point in X corresponding to the ith run, Z is a Gaussian process which represents uncertainty about the mean of y(x), as described in Section 2, with $E(Z(x)) = \mu$ and

$$Cov (Z(x), Z(w)) = \sigma^2 R(w - x),$$

and the ε_i's are Gaussian random variables, with mean 0 and variance α^2, independent of each other and of Z. Then

$$\tilde{C} = \sigma^2 C + \alpha^2 I$$

where $C = \{R(x^i - x^j)\}$ is the matrix of correlations among the Z's at the design points. Further, since we want to determine the posterior mean of Z(x),

$$\check{r}_i(x) = Cov (Z(x), Y(x^i)) = Cov (Z(x), Z(x^i)) = \sigma^2 R(x - x^i).$$

The fitted response surface is then given by (2.2) with $\check{r}$ and $\tilde{C}$ replacing r and C, respectively. The maximum likelihood estimates of the parameters ($\mu, \sigma^2, \alpha^2, \theta$) are again found by maximizing (2.3).

REFERENCES

Bernardo, M.C., Buck, R., Liu, L., Nazaret, W.A., Sacks, J., and Welch, W.J. 1992. Integrated Circuit Design Optimization Using a Sequential Strategy. *IEEE Transactions on Computer-Aided Design* 11:361-372.

Box, G.E.P. and Draper, N.R. 1987. *Empirical Model-Building and Response Surfaces.* New York: John Wiley and Sons.

Bratley, P., Fox, B.L., and Schrage, L.E. 1983. *A Guide to Simulation.* New York: Springer-Verlag.

Cressie, N. 1991. *Statistics for Spatial Data.* New York: John Wiley and Sons.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. 1991. Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments. *Journal of the American Statistical Association* 86:953-963.

Kleijnen, J.P.C. 1975. *Statistical Techniques in Simulation.* New York: Marcel Dekker.

Matern, B. 1986. *Spatial Variation (Second Edition), Lecture Notes in Statistics, No. 36.* New York: Springer.

McKay, M.D., Conover, W.J., and Beckman, R.J. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* 21:239-245.

Mitchell, T.J. and Morris, Max D. 1992. Bayesian Design and Analysis of Computer Experiments: Two Examples. *Statistica Sinica,* in press.

Morris, M.D. 1991. Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics* 33:161-174.

Patterson, H.D. 1954. The Errors of Lattice Sampling. *Journal of the Royal Statistical Society, Series B* 16:140-49.

Sacks, J., Schiller, S.B., and Welch, W.J. 1989. Designs for Computer Experiments. *Technometrics* 31:41-47.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. 1989. Design and Analysis of Computer Experiments (with Discussion). *Statistical Science* 4: 409-435.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., and Morris, M.D. 1992. Screening, Predicting, and Computer Experiments. *Technometrics* 34:15-25.

AUTHOR BIOGRAPHIES

TOBY MITCHELL is a Research Staff Member at Oak Ridge National Laboratory. His research interests include Bayesian inference, particularly its application to the design and analysis of computer experiments.

MAX MORRIS is a Research Staff Member at Oak Ridge National Laboratory. His research interests include the design and analysis of experiments, including computer experiments.