ornl

ORNL-6498

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

A Bayesian Approach to the Design and Analysis of Computer Experiments

Carla Currin
Toby Mitchell
Max Morris
Don Ylvisaker

OPERATED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A03; Microfiche A01

ORNL-6498

Engineering Physics and Mathematics Division

Mathematical Sciences Section

A BAYESIAN APPROACH TO THE DESIGN AND ANALYSIS OF COMPUTER EXPERIMENTS

Carla Currin
Bryn Mawr College
Bryn Mawr, Pennsylvania 19010

Toby Mitchell
Max Morris
Mathematical Sciences Section
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831-8083

Don Ylvisaker
Department of Mathematics
University of California at Los Angeles
Los Angeles. California 90024

Date Published: September 1988

This work was supported by the
Applied Mathematical Sciences Research Program
U.S. Department of Energy
Office of Energy Research
and
NSF Grant DMS 86-02018

Prepared by the
Oak Ridge National Laboratory
Oak Ridge, Tennessee 3783 1
operated by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-84OR21400

TABLE OF CONTENTS

A Bayesian Approach to the Design and Analysis of Computer Experiments

by Carla Currin, Toby Mitchell, Max Morris, and Don Ylvisaker

Abstract

We consider the problem of designing and analyzing experiments for prediction of the function y(f), $t \in T$, where y is evaluated by means of a computer code (typically by solving complicated equations that model a physical system), and T represents the domain of inputs to the code. We use a Bayesian approach, in which uncertainty about y is represented by a spatial stochastic process (random function); here we restrict attention to stationary Gaussian processes. The posterior mean function can be used as an interpolating function, with uncertainties given by the posterior standard deviations. Instead of completely specifying the prior process, we consider several families of priors, and suggest some cross-validational methods for choosing one that performs relatively well on the function at hand. As a design criterion, we use the expected reduction in the entropy of the random vector y (T^*), where $T^* \subset T$ is a given finite set of "sites" (input configurations) at which predictions are to be made. We describe an exchange algorithm for constructing designs that are optimal with respect to this criterion. To demonstrate the use of these design and analysis methods, several examples are given, including one experiment on a computer model of a thermal energy storage device and another on an integrated circuit simulator.

1. Introduction

1.1 Computer models and computer experiments.

There is widespread and growing use of computer models as tools in scientific research. Some are simulations of events and processes while others are programs for numerically solving equations that are derived from physical assumptions and laws. As surrogates for physical or behavioral systems, computer models can be subjected to experimentation, the goal being to predict how the corresponding real system would behave under certain conditions. Complex models often require long running times, however; even as computers become more and more powerful, researchers adjust quickly and develop more extensive and demanding models. The result is that computing time very often severely limits the size and scope of computer experiments. The research reported here, on the design and analysis of such experiments, is motivated by the goat of getting information from computer models as efficiently as possible.

Here we regard a computer model as a computer program that maps a vector of input variables (parameters) t into a vector of output variables y , where t and y are physically meaningful. For example, t might specify the boundary conditions and coefficients in a set of complicated differential equations, which are solved numerically by the computer program to produce a record of the state of some physical system at each of many points in space and time. From the mass of data that represents the solution of the equations, various responses y of interest are determined. We can therefore view y as a function y(r) over some domain T in the space of the input variables. This function is deterministic: if the program is run twice (on the same computer) with the same value of t , the same value of y will result.

We consider a *computer experiment* to be a collection of runs of the computer model, made for the purpose of investigating $y(t)$ for $t \in T$. For convenience. we shall consider T to be defined only by the *design variables,* i.e., those variables that are changed during the course of the experiment. In a typical experiment of n runs, the i^{th} computer run is made using inputs $t_i \in T$, $i = 1, 2, , n$; this collection of input configurations is called the *experimental design.*

There are several important general classes of problems that can be approached through computer experiments. Some of the major ones are:

(1) Prediction: Given t, predict y .

(2) Sensitivity analysis: Identify the important and the negligible input variables.

(3) Uncertainty analysis: Determine how uncertainty about t affects y Equivalently, determine the variability in y caused by random variability in t.

(4) Optimization: Find the t at which y is "best" in some sense.

(5) Root finding: Find a t that yields a specified y

(6) Integration of output: Find the average y that results when t is randomly drawn from a known input distribution.

Of course, these are interrelated. Perhaps the most fundamental is the problem of prediction, which relates to all the others in addition to being of interest in its own right. This is the subject of this paper.

We shall restrict attention here to the prediction of a scalar (univariate) y For a given design, multidimensional y's can be predicted by applying the scalar predictions separately to each component, although this would ignore potentially useful information about relationships among the components. We have not considered the important question of designing experiments for the purpose of predicting multiple, interrelated responses.

1.2 The prediction problem.

We consider a solution to the prediction problem to include a prediction equation $\hat{y}(t)$, formulas for evaluating the uncertainty of prediction, and rules for choosing the design. Because of the nature of our approach, which is described below, our method is quite similar to interpolation, in that the prediction of y will be identical to the observed y at values oft for which the model has been run, At other values of t, our prediction will take the form of a probability distribution, the mean of which, expressed as the function $\hat{y}(t)$, can be used as a prediction equation.

A frequently used approach to the prediction pmblem in computer experiments is based on the response surface methods that have so often been successful in physical experiments. (See, e.g., Baker and Bargmann, 1985.) Typically, one conducts the experiment using a standard response surface design (e.g, a fractional factorial or central composite design). A response surface model (e.g. a second order polynomial) is fitted to the data by the method of least squares, and the fitted model is then used for purposes of prediction. Measures of uncertainty, if given, are usually standard errors of prediction derived from classical least squares theory or confidence intervals based on normal regression theory.

Our reservations about this kind of approach to the problem of predicting deterministic functions are:

1. The class of approximating functions is not flexible enough. This is not a major problem for some applications, e.g., where. the predictive approximation is to be used for an analysis in which T is small enough so that a first- or second- order Taylor approximation is adequate. For more general applications, however, more flexible functions are needed. Extending the class of functions to higher-order polynomials is seldom practical, because of the large number of terms whose coefficients must be estimated. Moreover, any approach based on choosing a class of functions has the inherent limitation that one can do no better than to find the best approximation toy within that class.
2. The estimation procedure (least squares) is not well justified. Although it retains some heuristic appeal for this problem, its statistical justification is lost because of the absence of random error. What is of interest in the prediction problem is prediction at the points of T not run in the experiment; to our knowledge, there is no argument that supports least squares estimation for that purpose.
3. There is no theory that supports statements of uncertainty about the predictions. Confidence intervals in classical response surface methodology are based on the assumption that the "true" response function is in the assumed class, and that departures of the data from that function represent independent random variables. In the prediction problem, it is highly unlikely that the "true" response function will be in the assumed class, and, as we have already remarked, there is no random error.
4. Most standard designs have been developed using criteria appropriate for standard statistical models and inappropriate for the prediction problem when there is no random error.

1.3 A Bayesian approach.

We approach the problem from a Bayesian point of view, under which uncertainty about the function y is expressed by means of a probability distribution over all possible response functions. Random functions (stochastic processes. random fields) have been studied for a long time, and we borrow notation and nomenclature from that source. The initial (prior) process will generally be very diffuse, to indicate a high degree of uncertainty about the function y (r) given t. As data from a computer experiment become available, the prior process can be updated, under the rules of conditional probability. to yield the posterior process, which we shall frequently call the *predictive* process. The mean of the predictive process, which is a function of t , serves as a prediction equation, and the standard deviation, also a function of t, serves as a measure of uncertainty of prediction. Measures of information based on the predictive process can he used to establish design criteria, and computational optimization algorithms can be used to choose good designs.

The main ideas that underlie this approach are:

(1) The use of stochastic process models to make predictions about deterministic functions, and

(2) The adoption of a Bayesian approach to derive such models and to guide the formulation and solution of prediction and design problems.

These are not new ideas, especially (1), which has been applied extensively in the analysis of spatial data, and supports, for example, the "kriging" methods used in geostatistics. (See Ripley's (1981) book on spatial statistics or, for an introduction to the kriging literature in particular, the introductory sections of Cressie's (1986) article.) The prediction problem in kriging is usually formulated as the problem of making inferences about the realization of a spatial stochastic process Y(r), given the values of that process at a set of "sites" $t_1, \ldots, t_n$. See Ylvisaker (1987) for a discussion of problems of this general type and of the associated design problems. Recently, Shewry and Wynn (1986) and Sacks and Schiller (1987) proposed and used design optimality criteria based on spatial stochastic process models to compute optimal designs for prediction in various settings. Sacks, Schiller, and Welch (1988) applied such models to the design and analysis of computer experiments, which is the application of interest here. For $\hat{y}(t)$, they used the best (for squared error loss) linear unbiased predictor; for a design criterion, they used the mean squared error of prediction, integrated over the region of interest. Their examples included experiments on computer models for the homogenous pyrolysis of propane and the combustion of methane.

Kimeldorf and Wahba (1970) were the first, as far as we know, to use a stochastic process in an explicitly Bayesian sense, for the purpose of predicting a fixed but unknown function. They considered the correspondence between the prediction equation (the mean of the posterior process) and smoothing splines. In kriging, Bayesian approaches are still not common, the recent paper by Kitanidis (1986) being one of the few examples.

In this paper, we shall present our basic approach and give a simple example in one dimension (Section 2). We shall then discuss a design criterion and our design construction algorithm (Section 3). Both the design and the analysis are driven by the prior process, the choice of which presents a difficult problem in practice. Hem we strive for a semiautomatic Bayesian method,

using "impartial" priors to the extent possible and letting the data help choose the prior, rather than trying to faithfully model the experimenter's prior feelings. Although we are far from settling on a "best" way of making the choice of prior process, we discuss some criteria and describe our current approach (Section 4). Finally, several examples will be discussed (Section 5), including one experiment on a computer model of a thermal energy storage device and another on an integrated circuit simulator.

2. Prediction

2.1 The prior process.

We represent "knowledge" about the unknowo function y (t) by a stochastic process Y (t), where

(P1). Y(r) has a normal distribution with mean μ and variance σ^2 (the same for all t), and

(P2). For any pair of sites $t \in T$, $s \in T$, the correlation between Y (t) and Y(s) is a function only of the vector of differences $d = t - s$, i.e.,

$$\rho_{ts} = Corr(Y(t), Y(s)) = R(t-s) = R(d), \tag{2.1}$$

where $R(d) = R(-d)$ and $R(0) = 1$.

The properties (P1) and (P2) define *Y(r)* as a stationary Gaussian stochastic process. Normality is chosen for convenience; the posterior process is easily derived, as noted below. Stationarity is desirable from the objective Bayesian point of view as a way of expressing a form of prior exchangeability: the prior distribution of the response y(r) is the same for all t, and the prior distribution of the difference y $(t)-y(s)$ depends only on the difference between t and s

From a Bayesian viewpoint, the correlation between Y (t) and Y (s) in (P2) expresses the effect that exact knowledge of $Y(s)$ has on knowledge of Y(r). Given $Y(s) = y(s)$, Y(t) has a normal distribution with mean $\mu_{t|s} = \mu(1-R(d))+y(s)R(d)$ and variance $\sigma^2_{t|s} = \sigma^2[1-R^2(d)]$. Clearly, the choice of *R* in (2.1) is not arbitrary; *R* must be "legal" in the sense that, for any finite set of sites in T, the covariance matrix generated by *R* must be nonnegative definite.

2.2 The posterior (predictive) process.

The posterior process, given the set of observed responses $y(D)$ on the set of design sites $D \subset T$, is easily obtained as follows.

Let

$$C_D = Corr(Y(D), \text{Y(D)})$$

be the n x n matrix whose elements are the prior correlations between the responses at all pairs of design sites. For t $\in T$, let

$$r_D(t) = Corr(Y(t), Y(D))$$

be the n -vector of prior correlations between Y(t) and Y *(D).*

Then the posterior distribution of Y (r) is normal with mean:

$$\mu_{t\,|D} = \mu + r_D^T(t) C_D^{-1}(y_D - \mu J) \tag{2.2}$$

and variance

$$\sigma^2_{t|D} = \sigma^2[1 - r_D^T(t) C_D^{-1} r_D(t)], \tag{2.3}$$

where J in (2.2) is an n-vector of 1's, and y_D is the set of observed responses *y(D)* written as vector.

For t and s in T, the posterior covariance of Y (t) and Y (s) is

$$\Sigma_{ts\,|D} = \sigma^2[\rho_{ts} - r_D^T(t) C_D^{-1} r_D(s)] \tag{2.4}$$

All knowledge about y(t) given the data and the prior process is embodied in the posterior process defined by (2.2)-(2.4), which is Gaussian like the prior process, but is no longer stationary. Since we shall use the posterior process for prediction, we shall often refer to it as the "predictive process." When we consider the mean of this process as a function of I, we shall denote it by $\hat{y}_D(t)$ or simply j(r): this is an interpolating function, since it passes through the observed y 's. The posterior variance (2.3) can be used as a measure of uncertainty of prediction at site t; it is necessarily zero at the observed sites.

The computation of (2.2)-(2.4) is mainly a matter of inverting Co, or solving a set of n equations in n unknowns. This ordinarily takes very little time, relative to the time it would usually take for the computer model to generate a single response. Moreover, Co does not depend on t, so predictions can be generated very quickly for a large number of sites, once the n-run experiment on the computer model has been completed.

2.3 Linear correlation function in one dimension.

As a simple one-dimensional example with $T = [0,1]$, consider the correlation function:

$$R(d) = 1 - (1-\rho)|d|, \tag{2.5}$$

where $0 < \rho < 1$ is $Corr(Y(0), Y(1))$. In this case, the i^{th} element of $r_D(t)$ is $1-(1-\rho)|t-t_i|$, so $\hat{y}(t)$ is a linear spline interpolating function. (See equation (2.2).)

Remark 2.1. Negative values of ρ are permissible. and can lead to good predictive distributions in some cases, but we shall avoid them here because they are counter to the intuitive notion that the posterior variance of Y(r) given Y (s) should increase with I *r-s I*

Examole 1.

Consider an experiment consisting of five runs, equally spaced at intervals of 0.25 in T = [0,1], where the observed *values* ofy are 1.0, 0.86, 0.63, 0.49, and 0.39. Figure 1 shows the mean and

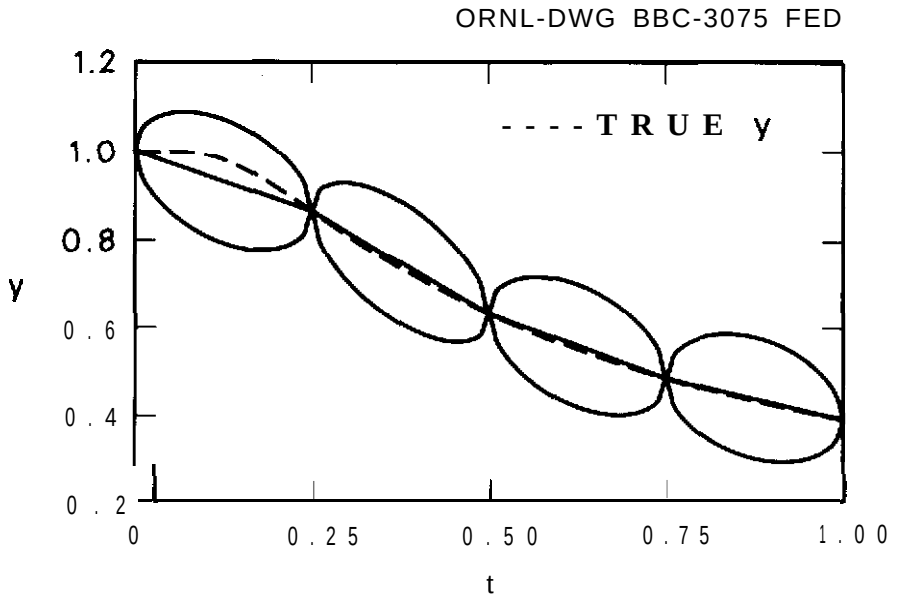


Figure 1. **Predictive mean $\hat{y}(t)$ and 95% predictive probability bounds ($\hat{y}(t) \pm 1.96\sigma_{t|D}$) after 5 runs (Example 1). The response data were generated by the function $y(t) = 1 - e^{-1/(2t)}$. The prior correlation function is linear (Equation 25) with $\rho = 0.000817$; the prior mean and standard deviation are $\mu = 0.70$ and $\sigma = 0.20$.**

95% probability bounds ($\hat{y}(t) \pm 1.96\sigma_{t|D}$) as functions of t when p = 0.000817, μ = 0.70. and a = 0.20. These values of the parameters of the prior process were not, in fact, specified a *priori,* but were chosen to maximize the likelihood (Section 4.2.3).

The function that we used to generate the data for this example is a survival function:

$$y(r) = 1 - e^{-1/(2t)}, \tag{2.6}$$

which is approximately linear over most of [0,1], but has zero slope at the origin. It is shown by the dashed line in Figure 1.

2.4 "Smoothed" correlation functions in one dimension.

In Figure 1, the lack of smoothness of $\hat{y}(t)$ at the data points and the rapid change in the width of the probability intervals there is due to the absence of prior information about the derivatives of $y(t)$. It can easily be shown that, unless *R'(0)* = 0, the prior variance of Y' (t) is infinite, This is the case for the linear correlation function given in (2.5). since *R'(d)* is discontinuous at the origin.

We can make the prior process smoother by supposing that the first derivative Y'(f) is a stationary Gaussian process having the linear correlation function given by (2.5) but with p replaced by γ, where $\gamma = Corr\ (Y'\ (0), Y'(1))$. (We shall reserve the notation ρ for the correlation between Y(0) and Y(l).) Mitchell, Morris, and Ylvisaker (1988) have found necessary and sufficient conditions for the existence of stationary Y having such a derivative process. The correlation function of Y is given by:

$$R(d) = 1-(a/2)d^2+(b/6)|d|^3 \tag{2.7}$$

where a and *b are* positive parameters that satisfy:

$$b^2 - 6ab + 12a^2 \leq 24b. \tag{2.8}$$

Since. p = $1-(a/2)+(b/6)$ and $\gamma = 1-b/a$, *(2.8)* can be expressed in terms of ρ and yas:

$$\rho \geq (5\gamma^2+8\gamma-1)/(\gamma^2+4\gamma+7). \tag{2.9}$$

This region in ρ and γ is shown in Figure 2. In this paper we restrict further to $\rho > 0$, $\gamma > 0$; see Remark 2.1.

Since $\hat{y}(t)$ is a linear combination of n functions of the form $R(t-t_i)$, the interpolating function that follows from the choice of cubic R, Equation 2.7, is seen to be a cubic spline.

Figure 3 shows the results of applying the cubic correlation function to the data in Example 1 above; compare with Figure 1. The parameters of the prior process, also chosen as in Section 4.2.3, are ρ = 0.0441, γ = 0.149, μ = 0.72, and σ = 0.34. The interpolating equation is smoother

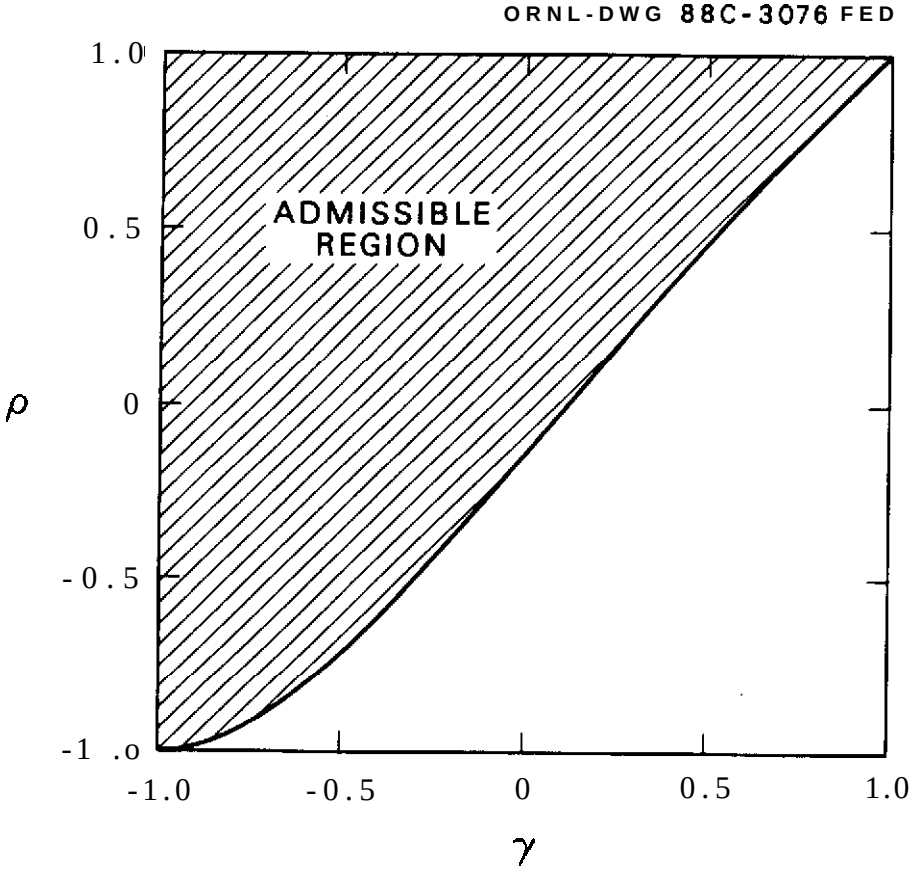


Figure 2. Admissible region for ρ and γ for the cubic correlation function. (See Equations 2.7-2.9.)

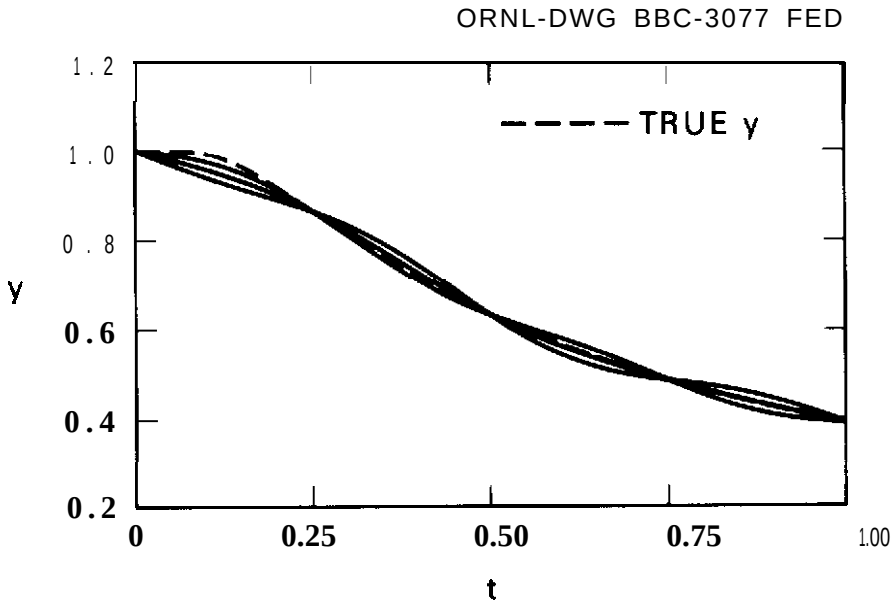


Figure 3. Predictive mean j(t) and 95% predictive probability bounds ($\hat{y}(t) \pm 1.96\sigma_{t|D}$) after 5 runs (Example **1), when the prior correlation function is cubic with ρ = 0.0441 and $\gamma = 0.149$, and the prior mean and standard deviation are $\mu = 0.72$ and $\sigma = 0.34$. The response data were generated by the function $y(t) = 1 - e^{-1/(2t)}$.**

in Figure 3, and the 95% probability hounds are much narrower. They seem a bit too narrow, in fact, since the true function falls outside of them in the range [0, .25]. (One should not however, interpret the 95% probability bounds as a confidence envelope for the whole response curve, since they are based on pointwise probability statements.)

Further smoothings can be made, as in Mitchell, Morris, and Ylvisaker (1988). We have also considered a few other families of correlation functions (Section 4.1).

2.5 Extension to **more** dimensions

Suppose now that there are two design variables and we want to be able to predict at sites within the unit square. Consider thc three sites t , s , and u in Figure 4 From the development for one dimension above, we can transfer information from s to u , i.e., we can predict $y(u)$ given y(s), and similarly from u to t We shall adopt this as the way to transfer information from s to t , i.e., we require:

$$p[y(t) \mid y(s)] = \int p[y(u) \mid y(s)]\, p[y(t) \mid y(u)]\, dy(u) \qquad (2.10)$$

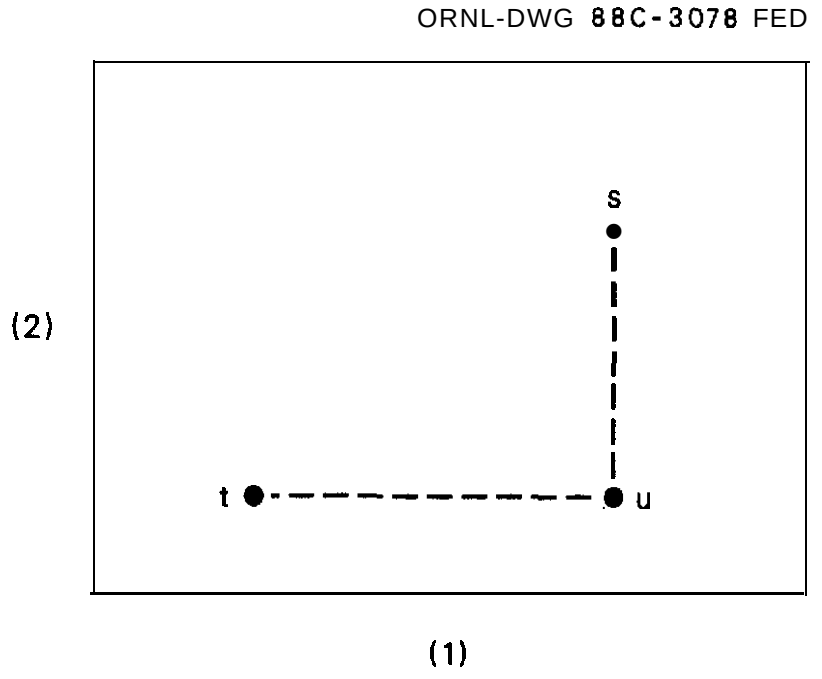


Figure 4. Under the product correlation rule, $\rho_{ts} = \rho_{tu}\rho_{us} = R_1(t_1 - s_1)R_2(t_2 - s_2)$.

where, e.g., $p[y(t) \mid y(u)]$ refers to the conditional density function of Y(r) given $Y(u)$. (We use p here generically to represent density functions.)

Given Y(t), $Y(u)$, and $Y(s)$ are jointly Gaussian, (2.10) can hold if and only if

$$\rho_{ts} = \rho_{tu}\rho_{us}, \tag{2.11}$$

where, e.g., ρ_{tu} is the correlation between $Y(t)$ and $Y(u)$, which we require, as above, to be a function only of the difference between t and u. Thus, (2.11) becomes

$$\rho_{ts} = R_1(t_1 - u_1)R_2(u_2 - s_2) = R_1(t_1 - s_1)R_2(t_2 - s_2), \tag{2.12}$$

where R_1 and R_2 are correlation functions for one-dimensional processes.

The same reasoning leads us in k dimensions to the *product* correlation rule, by which we define

$$\rho_{ts} = \prod_{j=1}^{k} R_j(t_j - s_j) \tag{2.13}$$

where t and s am in $\mathbb{R}^k$ and R_j, j =1, 2, ..., k, are correlation functions for one-dimensional processes. This rule has been used previously for prediction in spatial settings; see Ylvisaker (1975).

Remark 2.2. Our rationale for the product correlation rule is not a very strong one, although the notion of transmitting information along paths in which all but one variable is held fixed may have some appeal for those who like to think in terms of one-factor-at-a-time experimentation, One consequence of this notion is that, in Figure 4,

$$p[y(t) \mid y(s), y(u)] = p[y(t) \mid y(u)]; \tag{2.14}$$

this condition can be shown to be equivalent to (2.10) and (2.11). At present, we use the product correlation rule primarily for expediency, and we have not yet encountered any obvious pitfalls.

Remark 2.3. In situations where a single variable is represented by a point in several dimensions (like "location" on a two-dimensional surface), the selection of the coordinate axes for representing that point may be arbitrary. Then one might modify (2.13) by requiring the correlation between the responses at two locations (with the other variables fixed) to depend, for example, on the Euclidean distance between them. There am examples of such correlation functions in the literature on kriging. When each variable has a distinct physical meaning, however, the use of a definition of "distance" between two sites as a basis for choosing the form of the correlation function loses its intuitive appeal.

In this paper, we shall adopt the product correlation rule as given in (2.13). For example, in k dimensions, the linear correlation (2.5) becomes

$$R(d) = \prod_{j=1}^{k} (1 - (1 - \rho_j) \mid d_j \mid) \tag{2.15}$$

We generally allow each dimension to have its own parameter(s), although this complicates the problem of "estimating" them (Section 4.3). No matter what correlation function is chosen, the formulas for the properties of the posterior process remain the same as for the case of one design variable (see (2.2)-(2.4)).

An example of the appearance of the interpolating function that arises from the product of linear correlations is shown in Figure 5. where T is the unit square and there. are three observations as shown. Within each elementary rectangular piece of the grid generated by the three sites, $\hat{y}(t)$ can be (at most) bilinear; here it is linear in every piece. Similarly, the product of cubic correlations would produce bicubic functions in each piece.

3. Design

3.1 Design criterion

Suppose we want to design an experiment in n runs for prediction at a finite set of n^* sites $T^* \subset T$, where $n^* >$ n. After the experiment is run, knowledge of y at these sites will be embodied in the n^* -dimensional normal distribution of $Y(T^*$ ID) generated by the predictive process there. The mean $\mu_{T^* \mid D}$ and the covariance matrix $\Sigma_{T^* \mid D}$ of this distribution can be obtained using (2.2)-(2.4).

We would like to design the experiment to minimize, in some sense, the "amount of uncertainty" in $Y(T^*$ ID). To quantify this, we shall use Shannon's (1948) entropy, which, for a general multidimensional random variable X, is defined as

$$H_X = E[-\ln p_X(X)] + c, \tag{Entropy}$$

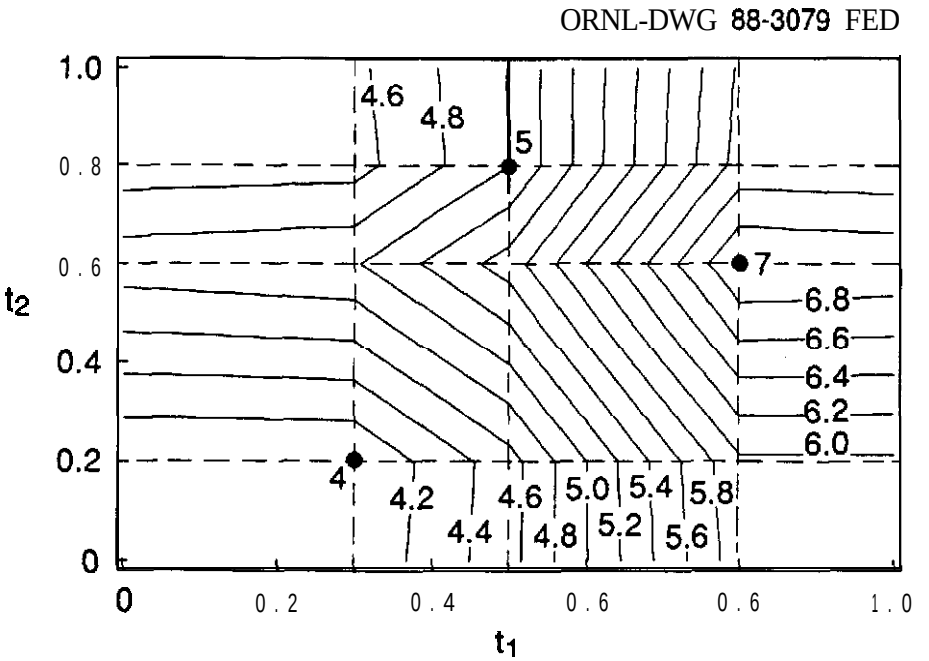


Figure 5. Contours of constant predictive mean ŷ(t) after observing y(03,0.Z) = 4, y(0.5, 0.8) = 5, and y(0.8, 0.6) = 7, where the prior correlation function is a product of two one-dimensional linear correlations with $\rho_1 = \rho_2 = 0.8$, and $\mu = 5$.

where p_X is the density function associated with X, E is the expectation, and c is an irrelevant constant. The entropy is always nonnegative; the lower me entropy, the more precise is the knowledge represented by X Lindley (1956) proposed using the expected reduction in entropy as a criterion for design. This has been done, e.g., by Box and Hill (1967) and Borth (1975) for model discrimination, by Shewry and Wynn (1986) for spatial sampling, and by Mitchell and Scott (1987) for group testing.

In the present setting, we take X to be $Y(T^*$ ID); this is multivariate normal with variance-covariance matrix $\Sigma_{T^*|D}$, so

$$H_{T^*|D} = 0.5 \ln \det \Sigma_{T^*|D} + c^*,$$

where c^* does not depend on the data. In fact, $H_{T^*|D}$ depends only on the design sites in this case, and not on the values of the responses. By choosing D to minimize $\det \Sigma_{T^*|D}$, we will therefore ensure that, after the experiment, the amount of uncertainty about the responses on T^* will be as small as possible.

In general, it can be shown that $H_{T^*|D}$ can be minimized over designs in T by choosing *D as the* subset of T^* on which *the prior* entropy H_D is *maximized* (Shewry and Wynn (1986).) For Gaussian priors, this criterion becomes

max det Co (D-optimality)

over all n-run designs D We shall call this D -optimality because, like the usual D -optimality criterion in the linear model setting, it minimizes the posterior generalized variance of the unknowns that one is trying to estimate.

Remark 3.1. For Gaussian prior processes, to minimize the entropy after the $(n+1)^{th}$ run, given n previous runs, choose the $(n+1)^{th}$ site to be one at which the predictive variance (after the first n runs) is maximum. This follows from Shewry and Wynn's result.

3.2 Design algorithm

Given a correlation function, a D-optimal design can. in principle, be found before any data on y arc taken, since the optimality criterion does not depend on y. Except in a few special cases, however, there seem to be few theoretical results available for finding such designs. The designs constructed for this paper were obtained fmm a computer algorithm adapted from DETMAX (Mitchell, 1974). which was first developed for the purpose of constructing D-optimal designs for linear regression. The optimization method is based on a series of "excursions," which are sequences of designs in which each design differs from its predecessor by the presence or absence of a single site. The first and last designs in an excursion have n sites; the intermediate designs all have fewer sites. (This restriction to designs with n or fewer sites was put in to avoid numerical problems associated with the nearly singular Co matrices that sometimes arose when the number of sites became large. It ensures that Co for any design D encountered during the excursion is at least as well conditioned as the starting design.)

The first step of each excursion removes a site from the best current design. At subsequent steps, a site is added, unless the design at that step has already been declared a "failure design," in which case a site is removed. (All designs encountered since the most recent successful excursion are designated as failure designs.) For the purpose of checking a design for equivalence to a failure design, only the determinants of their correlation matrices arc compared; thus false equivalence may occasionally be declared. All additions and deletions are made with the goal of maximizing the determinant of the correlation matrix for the resulting design. By Remark 3.1, the best site t to add to an existing design D is the one at which the variance function $\sigma^2_{t|D}$ is greatest. It can also be shown that the largest determinant after deletion of a site in D can be achieved by choosing that site to be the one associated with the greatest element of the diagonal of C_D^{-1}.

The search for the best site to add is conducted over a grid in T. Except when T has few dimensions or the grid is very coarse, it is not practical to make the search exhaustive. Instead we have incorporated a multiple search procedure that can best be envisioned by thinking of a set of n hikers trying to climb a hill. Each hiker starts at one of the n current design sites; at each of these the variance function is zero. The algorithm proceeds by stages, where in each stage, each hiker takes one step in the direction that allows him to increase his altitude the most. We restrict him to consider only the $2k$ neighboring grid points associated with a change in exactly one of the k design variables, and of course we don't let him step outside of T. Under this procedure, the variance function (2.3) is evaluated at (at most) $2nk$ sites in each stage. Sometimes, two

hikers will merge, in which case they continue as one. The search ends when all hikers have stopped at (local) maxima; the site that corresponds to the largest of these is taken to be the best site to bring into the design at the current point in the excursion.

The number of excursions made during each search ("try") is determined by restricting the maximum allowed deviation from the nominal number of runs (n). the maximum allowed number of successive excursions that fail to improve I Co I, and the maximum allowed number of "failure designs." (We generally set these restrictions to 4, 10, and 20, respectively.) When one of these constraints causes the search to end, a check for local optimality is made by removing each design site in turn and attempting to replace it by another, using the "hikers" algorithm. If the latter succeeds in finding the global maximum of the variance function in each case, then D is locally optimal in the sense that it cannot be increased by moving a single site. However, the success of the "hikers" algorithm is not guaranteed, and even if it were, the search would not necessarily produce a global optimum.

Figure 6 gives an example of a design (on a 6^5 grid) generated by our algorithm **for the case** $n = 6$, $k = 5$, for the linear correlation function with $\rho_j = .99$ for all j (When generating designs in the absence of previous data, we usually choose the same correlation function for each dimension.) This design exhibits some interesting geometrical structure, as shown by the intersite distance graph in Figure 6. Because of the high value of ρ, there is a large region in the middle of T in which there are no design sites; predictions here rely heavily on information from the surrounding design sites. This characteristic is even more pronounced for smoother correlation functions. If we use the cubic correlation with $\rho = .99$ and $\gamma = .99$ in the same case, all six sites in the optimal design am on corners of the S-cube. In fact, this design turns out to bc equivalent to the D-optimal first order regression design in 5 factors and 6 runs (Galil and Kiefer, 1980).

At the other extreme, designs that infiltrate T to a greater extent can be constructed by using correlation functions $R(d)$ that decrease rapidly with Id I. We favor such designs as initial designs in a stagewise approach, in which the correlation function that is used to generate the design sites at each stage may change during the course of the experiment, The cross-validational methods of Section 4 can be used to help select the correlation function to be used at each stage after the first. Examples of this kind of design strategy will be given in Section 5.

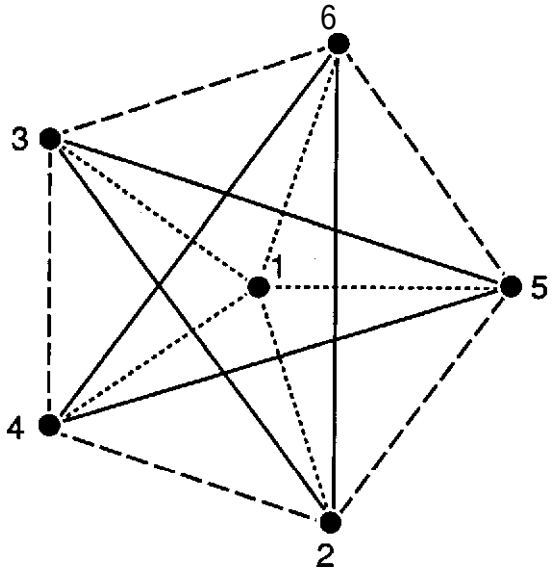
4. Choice of prior process

4.1 Alternative prior processes

We have made no attempt so far to investigate non-Gaussian or non-stationary prior processes. Within the stationary Gaussian family, we have used on occasion three correlation functions other than the linear and cubic correlations already described in Section 2, mostly on examples in one or two dimensions. They are the exponential, the smoothed exponential, and the Gaussian. These are all defined on $T = [0,1]$, but can be extended to general hyperrectangular regions by scaling the variables and applying the product correlation rule (2.13). In all three of these, ρ refers, as usual, to $Corr(Y(0), Y(1))$ and γ refers to $Corr(Y'(0), Y'(1))$.

ORNL-DWG 88M-3080R FED

SITE	t_1	t_2	t_3	t_4	t_5
1		0	0	0	0
2	0.			1	
3	0.0	016	11	0	01
4	1	1	0.6	0.6	0
5			0		1
6	01	01	0	i	0.6



INTERSITE DISTANCE (ON GRID)

2.6
2.8
3.2

Figure 6. **Our allegedly D-optimal design for five design variables and six runs, on a 6^5 grid, based on the product linear correlation function (Equation 2.15) with common $\rho = 0.99$. The graph below the design depicts the intersite distances, where the distances are defined by $d(t, s) = \sum_{j=1}^{5} |t_j - s_j|$.**

4.1.1 Exponential correlation.

The exponential correlation function

$$R(d) = \rho^{|d|}, \quad 0 < \rho < 1 \quad , \tag{4.1}$$

is the one associated with the well known Ornstein-Uhlenbeck process (Parzen, 1962, pp. 96-97). It is useful for design because it can be made to decrease very sharply in Id I by choosing ρ near 0. This leads to designs that fill in T as much as possible, since. what little information can be drawn at site t from the observations at the design sites must come primarily from its nearest neighbors. At the other extreme, as ρ approaches 1, (4.1) approaches the linear correlation (2.5), and, in one dimension, $\hat{y}(t)$ approaches a piecewise linear function through the observed data points. This can be seen by differentiating (2.2) twice with respect to t and noting that r,"(t) = $(\ln \rho)^2 r_D(t)$. Also, $\sigma^2_{t|D}$ approaches 0 in the limit, since it is no larger than $\sigma^2_{t|d}$ for $d \in D$ and the correlation between Y (t) and Y(d) tends to 1. A similar argument can be used in higher dimensions to show that the product exponential correlation leads to a k-linear spline interpolating function as the ρ_j 's approach I.

Remark 4.1. For experiments on [0,1] the D-optimal design for the exponential correlation function (4.1) can be derived theoretically. In this case, we can write

$$\det C_D = \prod_{i=1}^{n-1} (1 - R^2(t_{i+1} - t_i)),$$

from which it follows that, for a design to be D-optimal, the correlations between two adjacent design sites must all be equal, and as small as possible. Therefore. no matter what the value of ρ, the D -optimal design is equispaced and includes the two extreme sites at $t=0$ and $t=1$.

4.1.2 Smoothed exponential correlation.

The smoothed exponential correlation

$$R(d) = \frac{1-\gamma^{|d|}+|d|\ln\gamma}{-\ln\gamma-1+\gamma}(1-\rho)+1, \quad -1 < \rho < 1,\ 0 < \gamma < 1. \tag{4.2}$$

was obtained by Mitchell, Morris, and Ylvisaker (1988) from the exponential in the same way that the cubic correlation was derived from the linear (Section 2.4). Again there is a necessary and sufficient constraint on ρ and γ:

$$\rho \geq -1 + 2(1-\gamma)/(-\ln\gamma)$$

This region is shown in Figure 7; for the examples in this paper we shall require further that $\rho > 0$.

4.1.3 Gaussian correlation.

The Gaussian correlation

$$R(d) = \rho^{d^2}, \quad 0 \leq \rho < 1. \tag{4.3}$$

was used in the examples of Sacks, Schiller and Welch (1988). It is very smooth in the sense that it puts all of its probability mass on analytic functions; all derivatives of Y(r) have finite variances.

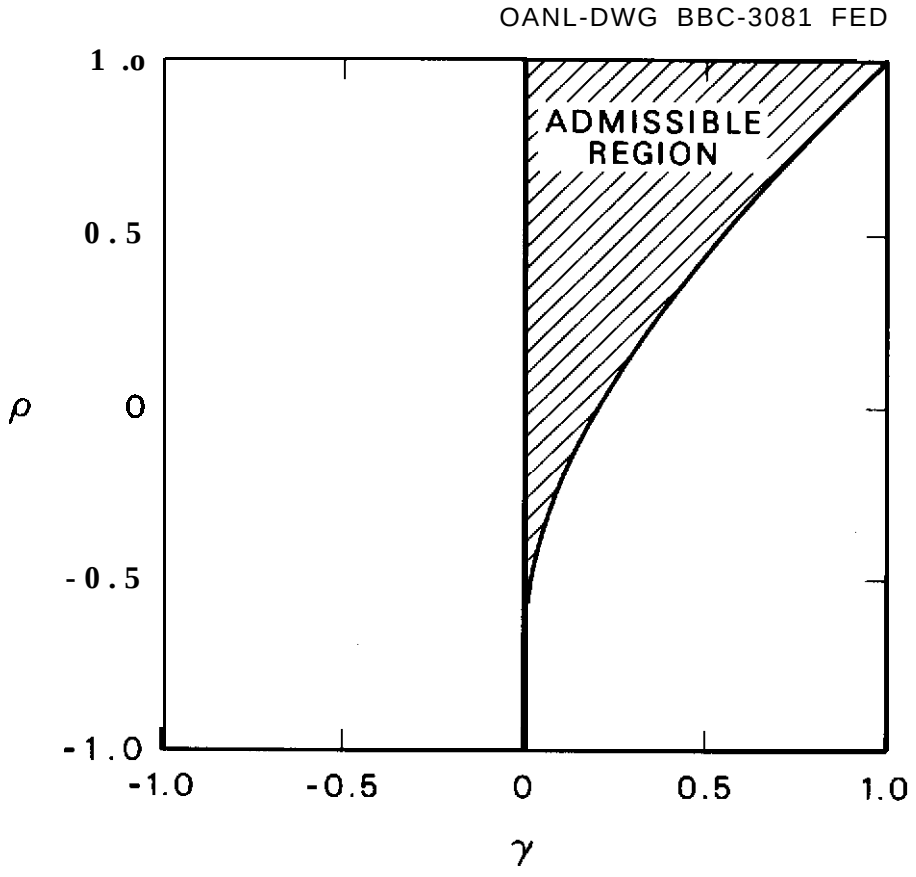


Figure 7. Admissible region for ρ and γ for the smoothed exponential correlation function (Equation 4.2).

We have not **considered** enough examples to be able to make any general recommendations about which type of correlation function to use in a given situation. We currently lean toward the cubic or the smoothed exponential, with a slight preference for the former because of its more direct connection with cubic splines. The cross-validational methods described below can be of value in making the choice, since cross-validational criteria can be compared across families of prior processes as well as within families.

4.2 Cross-validation

It is often possible to choose prior values of μ and σ, based on one's knowledge of the overall response level and the **expected** magnitude of departures from that level; even so, the posterior process can be quite sensitive to these values. It is generally more difficult to choose, *a priori*, a completely specified correlation function. Our preference is to make all of these choices using the three cross-validational methods described in Sections 4.2.1-4.2.3 below.

We consider the method of inference described in Section 2 to be actually a *family* of methods, where each method corresponds to a particular value of (μ, σ) and a completely specified correlation function R We then ask which method performs "best" in a cross-validational sense. The criteria. we have used arc described below. Our hope was that one criterion would appear superior, but we have tentatively concluded that we do better by considering all three jointly in an informal, subjective way. This will be described more fully in Section 4.3: in this section we present the mechanics of the computations.

Remark 4.2. A fully Bayesian approach would require a completely specified prior process, rather than a family of prior processes. This could be accomplished here by specifying a prior distribution on (μ, σ, R). It is not difficult to derive the posterior process when μ and log σ are given the usual "noninformative" improper uniform priors. For fixed R , the posterior distribution of $y(t)$ is Student's t, as one would expect, and the entropy criterion becomes equivalent to the maximization of $|C_D|\,(J^T C_D^{-1} J)$, where J is the n-vector of 1's. We have not yet explored approaches in which a prior distribution on R is specified, primarily because of the lack of an obvious candidate for one that is "noninformative."

42.1 "Leave-one-out" predictive density

We consider the n sets of data that result from leaving out a single site and measure, in each case, the effectiveness of the method for predicting the value of y at the deleted site. Let $p_i(\,|\,\mu, \sigma, R)$ represent the predictive density for y_i, based upon the i^{th} "training sample," i.e.. the data set that excludes the i th observed site. The mean and variance of this distribution are:

$$Pi = y_i - q_i(g_i - \mu w_i) \tag{4.4}$$

$$\sigma_i^2 = \sigma^2 q_i \tag{4.5}$$

where

$$g = g(R) = [C_D(R)]^{-1} y_D, \tag{4.6}$$

$$w = w(R) = [C_D(R)]^{-1} J, \tag{4.7}$$

and $q = q(R)$ is the inverse of the diagonal of $C_D^{-1} = [C_D(R)]^{-1}$.

We shah define the *predictive deficiency* of this distribution to be the negative log of its density at the observed y_i :

$$-\ln p_i(y_i\,|\,\mu, \sigma, R) = \frac{1}{2}[\ln(2\pi) + \ln \sigma_i^2 + \frac{(y_i - \mu_i)^2}{\sigma_i^2}] \tag{4.8}$$

'The average deficiency over all n design sites is:

$$\Phi_{pd} = \Phi_{pd}(\mu, \sigma, R) = \frac{1}{2}[\ln(2\pi) + \frac{1}{n}\sum_{i=1}^{n}\ln \sigma_i^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma_i^2}] \tag{4.9}$$

$$= \frac{1}{2}[\ln(2\pi) + \frac{1}{n}\sum_{i=1}^{n}\ln q_i(R) + \ln \sigma^2 + \frac{1}{n\sigma^2}\sum_{i=1}^{n} q_i(R)(g_i(R) - \mu w_i(R))^2].$$

Remark 4.2. The predictive deficiency (4.9) is essentially the same as that used by Geisser and Eddy (1979). In effect, we am treating the predictive distributions as though they had independently generated the observed y 's, and are estimating their parameters by the method of maximum likelihood. The general objective here is to choose these parameters so that the observed y's look as though they could reasonably have been drawn from their respective predictive distributions.

For fixed R , Φ_{pd} can be minimized with respect to μ and σ by

$$\hat{\mu}_{pd}(R) = \frac{\sum_{i=1}^{n} q_i(R) w_i(R) g_i(R)}{\sum_{i=1}^{n} q_i(R) w_i^2(R)} \tag{4.10}$$

$$\hat{\sigma}_{pd}^2(R) = \frac{1}{n}\sum_{i=1}^{n} q_i(R)(g_i(R) - \hat{\mu}_{pd} w_i(R))^2 \tag{4.11}$$

and the average deficiency for fixed R **becomes**

$$\Phi_{pd}^*(R) = \Phi_{pd}(\hat{\mu}_{pd}(R), \hat{\sigma}_{pd}(R), R) = \frac{1}{2}[\ln(2\pi) + \frac{1}{n}\sum_{i=1}^{n}\ln q_i(R) + \ln \hat{\sigma}_{pd}^2(R) + 1] \tag{4.12}$$

which is to be minimized over R

4.2.2 "Leave-one-out" squared bias

Sometimes, only the posterior mean is of interest. In this case, one may want to measure the performance of a correlation function by considering only the leave-one-out residuals:

$$e_i = y_i - \mu_i = q_i(g_i - \mu w_i) \tag{4.13}$$

We shah call e_i the (predictive) bias at the i^{th} site. The average squared bias

$$\Phi_b = \Phi_b(\mu, R) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} q_i^2(R)(g_i(R) - \mu w_i(R))^2 \tag{4.14}$$

is minimized for given R by

$$\hat{\mu}_b(R) = \frac{\sum_{i=1}^{n} q_i^2(R) w_i(R) g_i(R)}{\sum_{i=1}^{n} q_i^2(R) w_i^2(R)} \tag{4.15}$$

We can then search for the R that minimizes $\Phi_b^*(R) = \Phi_b(\hat{\mu}_b(R), R)$.

Remark 4.3. The predictive mean squared error:

$$\frac{1}{n}\sum_{1}^{n}(e_i^2 + \sigma_i^2) = \Phi_b + \frac{\sigma^2}{n}\sum_{1}^{n} q_i,$$

is always minimized by setting $\sigma = 0$, and so reduces here to the squared bias criterion.

When using the squared bias criterion to choose $\mu = \hat{\mu}_b$ and $R = \hat{R}_b$, *we* suggest that σ^2, which does not affect the bias, be chosen to minimize $\Phi_{pd}(\hat{\mu}_b, \sigma, \hat{R}_b)$, i.e., use (4.11) with $\hat{\mu}_b$ instead of $\hat{\mu}_{pd}$, and with $R = \hat{R}_b$.

42.3 Maximum likelihood

The two methods described so far are both forms of cross-validation based on training samples of size n-l. An alternative approach, which we like to view as an extension to training samples of different sizes, is to define the predictive deficiency to be n^{-1} times the negative log likelihood:

$$\Phi_l = \Phi_l(\mu, \sigma, R) = -\frac{1}{n}\ln p(y_D \mid \mu, \sigma, R)$$

$$= \frac{1}{2}[\ln(2\pi) + \ln\sigma^2 + \frac{1}{n}\ln|C_D(R)| + \frac{1}{n\sigma^2}(y_D - \mu J)^T [C_D(R)]^{-1}(y_D - \mu J)]$$

$$= \frac{1}{2}[\ln(2\pi) + \ln\sigma^2 + \frac{1}{n}\ln|C_D(R)| + \frac{1}{n\sigma^2}\sum_{i=1}^{n}(y_i - \mu)(g_i(R) - \mu w_i(R))] \tag{4.16}$$

Then

$$\hat{\mu}_l(R) = \frac{\sum_{i=1}^{n} g_i(R)}{\sum_{i=1}^{n} w_i(R)} \tag{4.17}$$

$$\hat{\sigma}_l^2(R) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_l(R))(g_i(R) - \hat{\mu}_l(R) w_i(R)) \tag{4.18}$$

Substituting into (4.16). the minimum deficiency for fixed R is:

$$\Phi_l^*(R) = \Phi_l(\hat{\mu}_l, \hat{\sigma}_l, R) = \frac{1}{2}[\ln(2\pi) + \frac{1}{n}\ln|C_D(R)| + \ln\hat{\sigma}_l^2(R) + 11 \tag{4.19}$$

which is to he minimized over R

The likelihood deficiency Φ_l can be written as a sum of average predictive deficiencies in the sense of Section 4.2.1, where the "training samples" consist of all subsets of n-l or fewer observations. This can be seen by writing the likelihood in n ! ways as

$$L = p(y_{i_1})p(y_{i_2}|y_{i_1}) \cdots p(y_{i_n}|y_{i_1}, y_{i_2}, \cdots, y_{i_{n-1}}) \tag{4.20}$$

where $i_1, i_2, , i_n$ is a permutation of $1, 2, , n$. Taking logs on both sides and averaging over all n ! equations yields

$$\Phi_l = \bar{\Phi}_{pd}^{(1)} + \bar{\Phi}_{pd}^{(2)} + \cdots + \bar{\Phi}_{pd}^{(n)}, \tag{4.21}$$

where $\bar{\Phi}_{pd}^{(j)}$ is the average of all deficiencies of the form (4.9), taken over all subsets of j sites. Note that $\bar{\Phi}_{pd}^{(n)}$ is the same as Φ_{pd}.

4.3 Optimization

Our current procedure for choosing μ, σ, and R in practice is rather primitive. We first choose a number of candidates for R within a given family of correlation functions by picking ρ_j (and γ_j if necessary), j = I, 2, , k, from a uniform distribution. (In the examples of the next section, we use 800 such candidates.) Each candidate is then evaluated with respect to each of the three criteria described in Section 4.2. At the end of this search, any process that was best in its family under any of the three criteria becomes a "finalist." Since we consider five different families and three different criteria, there are fifteen finalist processes. These are then evaluated subjectively by considering the values of Φ_{pd}, Φ_b, and Φ_l for each one. Usually, several can he rejected immediately because there is another that is better with respect to all three criteria. Others are then rejected because they are clearly weak with respect to at least one criterion. This usually leaves a manageable subset from which to choose one.

5. Examples

In this section we discuss the application of the methods of this paper to four examples. In the first two, the data are generated by known test functions, although we shall treat them as unknown functions evaluated by a computer model. In the last two examples, real computer models am used. In all of these examples, the random search method described in Section 4.3 was employed to present 15 "finalist" processes. from which one was chosen as the prior process on the basis of overall cross-validational performance.

5.1 Sine function.

The data were generated by me function

$$y(t) = \sin(2\pi(t - 0.1)) \tag{5.1}$$

at the sites $t = 0, 0.25, 0.5, 0.75, 1$.

Of the 15 finalists presented by the random search method the process that minimized Φ_l within the Gaussian correlation family was chosen; it performed well under all three criteria. A plot of the posterior mean and me upper and lower 95% probability bounds is shown in Figure 8.

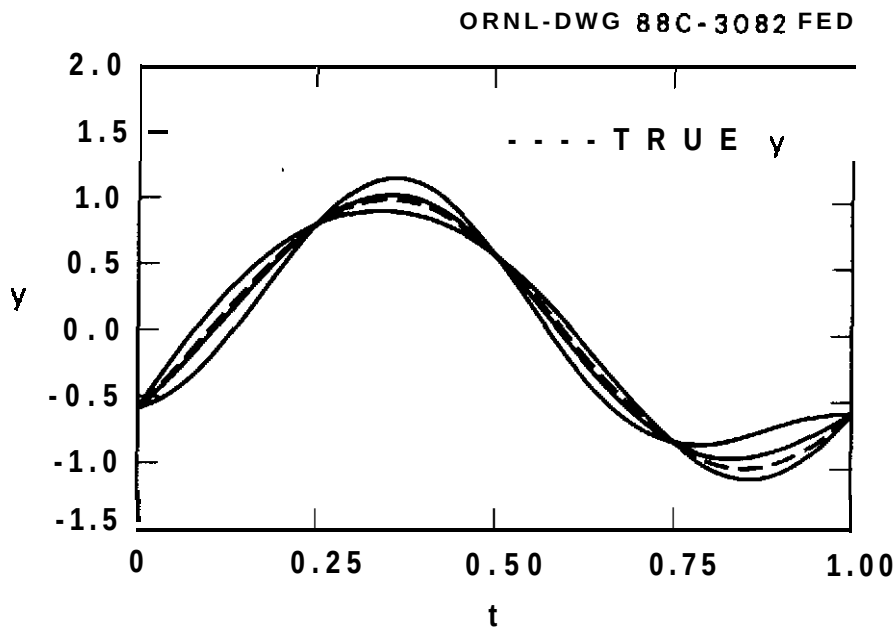


Figure 8. Predictive mean $\hat{y}(t)$ and 95% predictive probability bounds ($\hat{y}(t) \pm 1.96\sigma_{t|D}$) after 5 runs (Section 5.1). The response data were generated by the function $y(t) = \sin(2\pi(t - 0.1))$. The prior correlation function is Gaussian (Equation 4.3) with $\rho = 0.000817$; the prior mean and standard deviation are $\mu = -0.241$ and $\sigma = 0.917$.

5.2. **Test function in two dimensions.**

Here we again pretended that $y(t)$ was an unknown function generated by a computer model, but we used a known function to generate the response values:

$$y(t_1, t_2) = (1 - e^{-1/(2t_2)}) \frac{2300t_1^3 + 1900t_1^2 + 2092t_1 + 60}{100t_1^3 + 500t_1^2 + 4t_1 + 20} \qquad (5.2)$$

For prediction of y(t) on the unit square $T: 0 \le t_j \le 1$, j =1, 2, we adopted a general approach that does not require much prior knowledge about y. We first designed the experiment using an exponential correlation function with $\rho = .0001$ (Section 4.1.1). The best design on a 20×20 grid produced by our algorithm in ten tries is shown in Figure 9. All ten tries gave slightly different determinant values. so it is unlikely that this design is truly optimum. There seemed to be little point in undertaking more tries, however, especially since the computing time per try was about 45, seconds on a Cray X-MP. We did try various grid sizes, to avoid penalizing ourselves by choosing too coarse a grid. We found that 20×20 was sufficient: finer grid sizes require increasingly longer computation times with little apparent benefit.

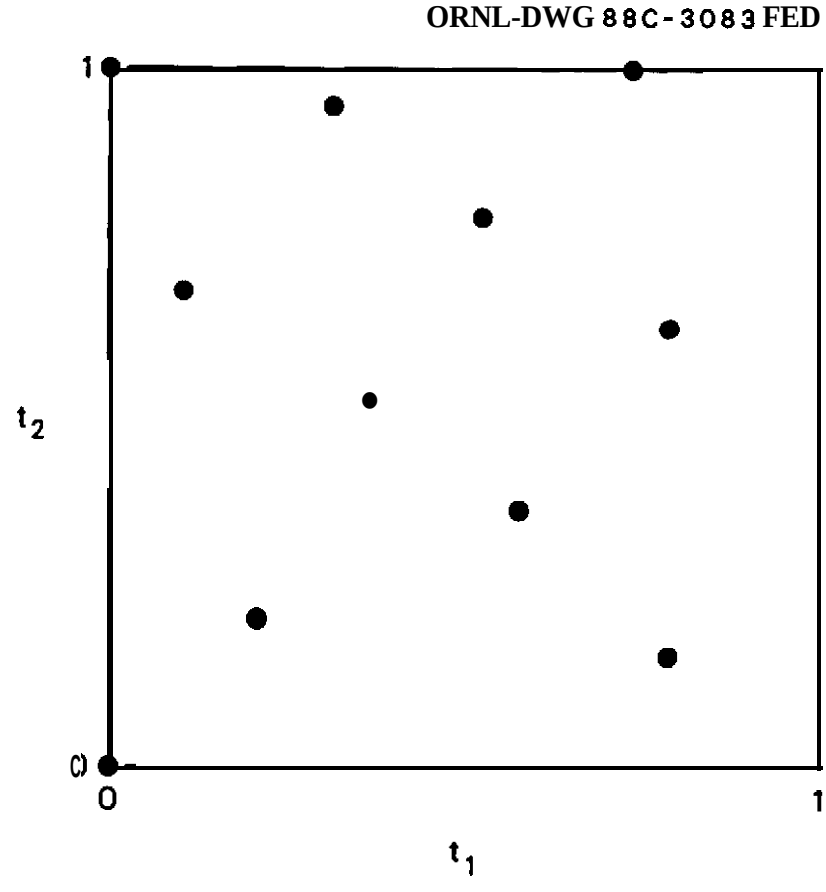


Figure 9. **Design for k = 2 and n = 16, used in example of Section 5.2. This was the best design (under the entropy criterion) produced by our algorithm in ten tries on a 20x20 grid, given a product exponential correlation function with each $\rho = 0.0001$.**

The response data were generated by the function (5.2) and analyzed using the five correlation functions and three cross-validational criteria that we have discussed. Of the 15 finalists, the two that seemed the best overall were (A) the process that minimized Φ_b within the smoothed exponential correlation family, and (B) the process that minimized Φ_l within the cubic correlation family. Contours of constant y for (A) and (B) are shown in the first two panels of Figure 10; the contours of the true response (5.2) are in the third panel. The maximum error of $\hat{y}$ for (A) on an 11x11 grid is 2.45, and the root mean squared (RMS) error on the same grid is 0.50. The corresponding values for (B) are 2.36 and 0.48. By way of comparison, we fit several polynomial models by least squares to the 16 sites on the 4x4 grid that covers the unit square.

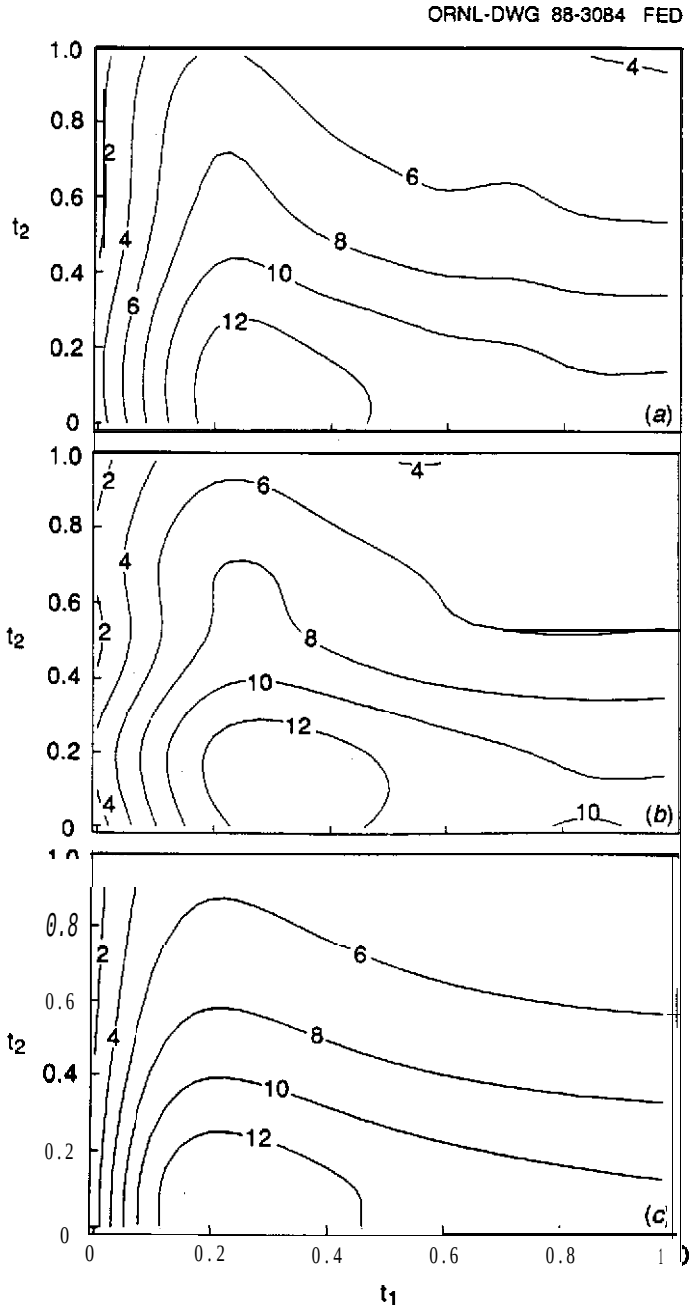


Figure 10. (a) **Contours of constant $\hat{y}(t_1, t_2)$ after 16 observations of the function (Equation 5.2), where the prior correlation function is a product smoothed exponential with ρ_1 = 0546, γ_1 = 0.245, ρ_2 = 0.762, γ_2 = 0.770, and where μ = -23.510 and σ = 20.318.**

(b) **Contours of constant $\hat{y}(t_1, t_2)$ after 16 observations of the function (Equation 5.2), where the prior correlation function is a product cubic with ρ_1 = 0.0267, γ_1 = 0.0813, ρ_2 = 0.754, γ_2 = 0.714, and where μ = -7.991 and σ = 23.213.**

(c) **Contours of constant $y(t_1, t_2)$ for the function (Equation 5.2).**

The extent of error in the fitted surfaces (as measured on an 11x11 grid) is shown in the last two columns of the following table:

Polynomial	Error D.F.	Error S.S.	R^2	Max. Error	RMS Error
Quadratic	10	26.14	86.3	5.88	1.94
Cubic	6	2.18	98.9	4.04	1.12
Bicubic	0	0.00	100.0	3.63	1.02

5.3. Thermal energy storage system example (two dimensions).

We now discuss an experiment that we conducted as a demonstration exercise using the model TWOLAYER, which was created by Dr. Alan Solomon and his colleagues at the Oak Ridge National Laboratory. TWOLAYER models heat transfer into, out of, and through a wall containing two layers of possibly different phase change materials. Heat is applied to the wall during a 10 hour charge cycle, during which time some of the phase change material melts. During the following 14 hours (the discharge cycle) heat is released from the wall naturally as the phase change material solidifies. Model inputs include layer dimensions, thermal properties of the materials, and characteristics of the heat source.

Our experiment was conducted to determine the effect of the melting temperature (t_1) and thickness (t_2) of one of the layers on a "utility index" (y), which is the proportion of phase change material that changes phase during a certain period of heat discharge. The region of interest was defined by $40 \le t_1 \le 160$ and $0.03 \le t_2 \le 0.07$, which we transformed (coded) to the unit square $[0,1]^2$.

For our initial experiment, we chose an 8-run design, generated to be optimal on a 13x13 grid for the exponential correlation with $\rho = .0001$. The design points **and the responses were:**

t_1	t_2	Y
0.0000	0.0000	0.6122
0.0000	1.0000	0.4290
1.0000	0.0000	0.0000
1.0000	1.0000	0.0000
0.1667	0.5000	0.3623
0.5000	0.1667	0.0898
0.5000	0.8333	0.0350
0.8333	0.5000	0.0000

Of the 15 prior processes presented as "optimal" by the random search method, the one that minimized Φ_l within the product smoothed exponential correlation family and the one that minimized Φ_l within the product cubic correlation family had the best overall cross-validational performance. There was very little difference between them, and in fact they shared the same values of the correlation parameters. Figure 11 shows the contours of constant $\hat{y}$ for the posterior process derived from the smoothed exponential correlation.

This process was used as a basis for choosing three additional sites, again using the entropy criterion. We restricted the new sites to the region $t_1 \le 0.5$, since we were not very interested in y

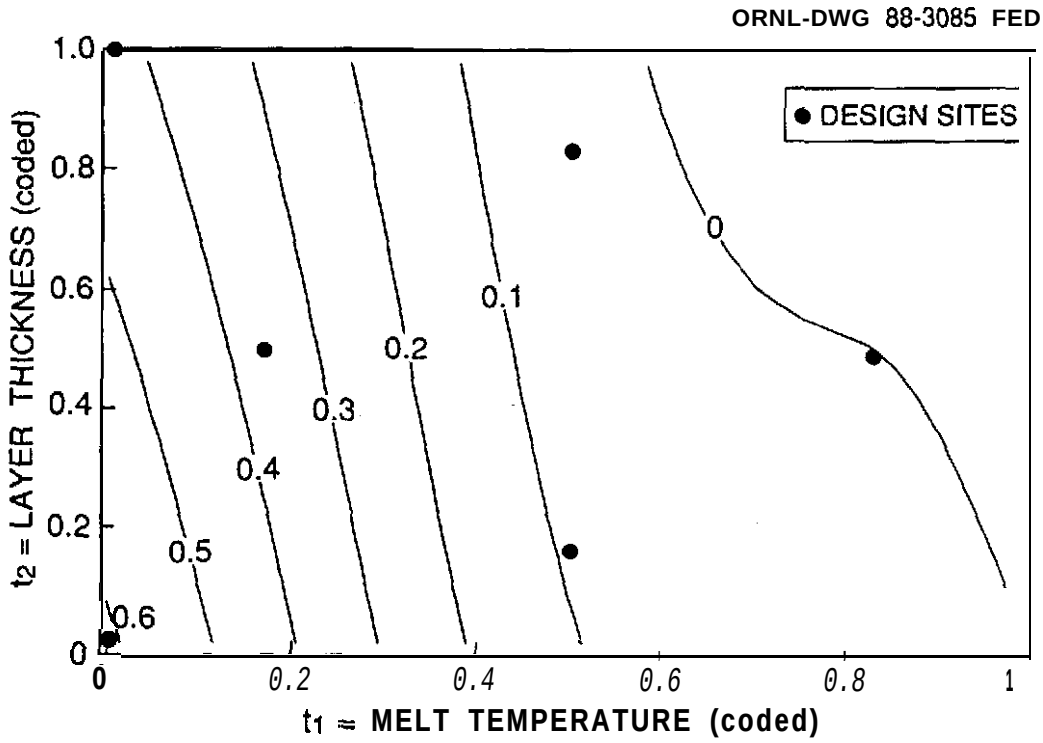


Figure 11. Contours of constant $\hat{y}(t_1, t_2)$ **after 8** observations of the utility **index y produced by the computer model** TWOLAYER, where the correlation **function is a product smoothed exponential with** $\rho_1 = 0.175$, $\gamma_1 = 0.159$, $\rho_2 = 0.924$, $\gamma_2 = 0.824$, **and where** $\mu = 0.470$ **and** $\sigma = 0.328$.

at or near 0, but the entropy criterion was based on all 11 sites. The new sites and the response values there were:

t_1	t_2	Y
0.25	0.0	0.5288
0.25	1.0	0.2503
0.3333	0.5	0.2306

After repeating the search for a "best" prior process, we **settled** on the one that gave the lowest value of Φ_l within the product linear correlation family. The contours of constant $\hat{y}$ derived from this process are shown in Figure 12.

Remark 5.1. Like many model codes, TWOLAYER produces only an approximate solution to the differential equations of the model. Here the approximation is not very good, since we adjusted the parameters of the solution method to reduce the amount of computer time needed to produce the response. As a result, tbe response surface $y(t_1, t_2)$ has plateaus and finite jumps, very much like a two-dimensional step function. The prior processes we have described here are not well suited for accurate prediction of this kind of function, although the main features of the response surface (except for the discontinuities) are well conveyed by Figure 12.

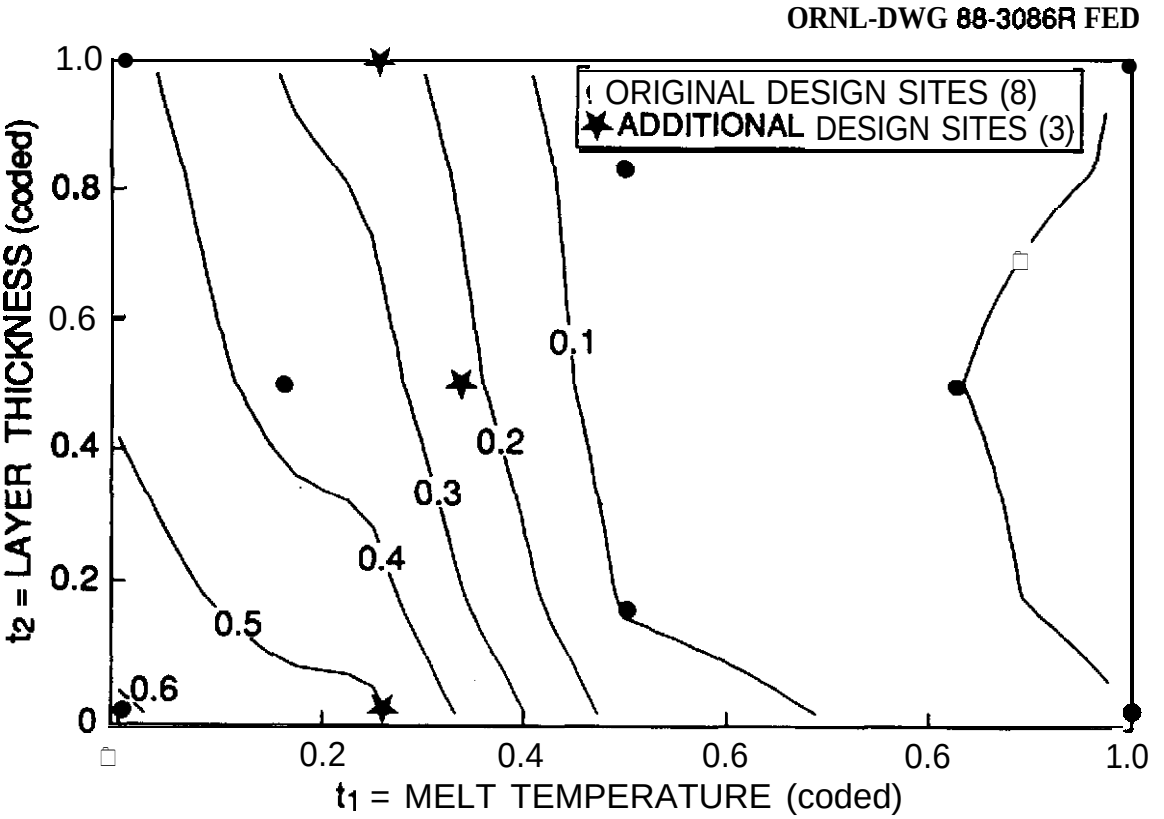


Figure 12. Contours of constant $\hat{y}(t_1, t_2)$ after 11 observations of the utility index y produced by the computer model TWOLAYER, where the correlation function is a product linear with $\rho_1 = 0.004527$, $\rho_2 = 0.788$, and where $\mu = 0.260$ and $\sigma = 0.218$.

5.4 Circuit simulation example (six dimensions).

This experiment was run on a computer model similar to the one described by Welch, et al. (1988). The model is used to help design an integrated circuit, in this case a CMOS VLSI clock driver. From a master clock, the circuit generates two **output** clocks of opposite polarities. The objective of this experiment is to determine the effect of six transistor widths on the "clock skew," which is a measure of the degree of asynchronization between the clocks.

We decided to do the experiment in two sets of 16 runs, with the analysis at the end of the first set used to guide the design for the second set.

Table 5.1 shows the design sites for the first 16 runs **and the response** values (clock skew) found at those sites. The actual values of the design variables have been shifted and scaled to make $T = [0,1]^6$. This design was generated using a product exponential correlation with $\rho_j = 0.1$ for

all j, following the same philosophy that we used in the examples above. The search was restricted to a 5^6 grid, to save computer time. The design shown here is the best one found by the algorithm in 10 tries, which took a total of about 20 minutes on a Cray X-MP. (At the time, the design algorithm was such that values of ρ_j much less than 0.1 would have resulted in much longer search times. Since then, we have modified the algorithm so that 10 tries with $\rho_j = 0.01$. e.g., would require about 25 minutes.)

Table 5.1. Design Sites and Response Values for Runs l-16 of Experiment on Circuit Simulator.

t_1	t_2	t_3	t_4	t_5	t_6	Y
1.00	0.00	**0.75**	0.00	**0.50**	**0.50**	-1.3480
0.00	1.00	**1.00**	0.00	**0.00**	**0.00**	**-0.9880**
0.00	0.00	**0.00**	0.00	1.00	1.00	**-0.8510**
0.75	**0.50**	**0.25**	**0.75**	1.00	**0.75**	**-0.3150**
1.00	**0.00**	1.00	1.00	1.00	**0.00**	**-0.5709**
1.00	1.00	1.00	0.00	1.00	1.00	**-1.2960**
0.50	**0.25**	0.00	0.00	**0.00**	**0.25**	-1.0190
1.00	**1.00**	**0.75**	1.00	0.00	**0.50**	-1.1351
0.00	**0.00**	**0.50**	1.00	0.00	**0.00**	-1.1501
0.25	**0.50**	**0.75**	**0.25**	1.00	**0.00**	-0.1160
0.00	1.00	**0.00**	1.00	1.00	**0.00**	**0.1627**
1.00	0.00	**0.00**	1.00	**0.25**	1.00	**-0.7740**
0.25	0.00	1.00	**0.25**	**0.00**	1.00	**-2.3570**
0.00	**0.75**	1.00	1.00	**0.75**	1.00	**-0.9529**
0.00	**1.00**	0.00	**0.50**	**0.00**	1.09	**-0.7490**
1.00	1.00	**0.00**	**0.25**	**0.50**	0.00	**0.3390**

Of the 15 finalist candidates for best prior process, the three based on the product cubic correlation function had the best cross-validational performance for these data. The same correlation parameters were optimal (in 800 random choices) for all three cross-validation criteria; they am given in the following table:

j	ρ_j	γ_j
1	**0.996**	**0.537**
2	0.910	0.0103
3	**0.700**	**0.571**
4	**0.428**	**0.0268**
5	**0.589**	**0.512**
6	**0.690**	**0.0694**

Of the three finalist processes that had this correlation function, we chose the one that was optimal with respect to Φ_b, since it was either in first or second place when judged by each of the three criteria. The optimal prior mean and standard deviation were μ = -1.339 and σ = 0.570.

We used this process to generate the next set of 16 runs, again from a 5^6 grid. The entropy criterion was based on all 32 runs. The algorithm made three searches, taking a total of between 5 and 10 minutes on the Cray X-MP. The best of the three resulting designs is shown in Table 5.2, together with the observed responses.

Table 5.2. Design sites and Response Values for Runs 17-32 of Experiment on Circuit Simulator.

t_1	t_2	t_3	t_4	t_5	t_6	Y
0.00	1.00	1.00	0.75	0.00	1.00	-1.5615
1.00	1.00	0.00	1.00	0.00	0.00	-0.2806
1.00	0.00	1.00	1.00	0.00	1.00	-2.2942
0.00	1.00	0.00	0.00	1.00	0.50	-0.0560
1.00	0.00	0.00	0.00	1.00	0.00	-0.0060
1.00	1.00	1.00	0.00	1.00	0.00	-0.2680
1.00	0.00	0.00	0.00	0.00	1.00	-1.6800
0.00	0.00	0.00	1.00	1.00	1.00	-0.3991
1.00	0.00	0.00	1.00	1.00	0.00	0.0665
1.00	0.00	1.00	0.50	0.00	0.00	-1.3671
1.00	1.00	1.00	0.75	1.00	0.25	-0.4492
1.00	1.00	0.00	0.50	1.00	1.00	-0.1300
1.00	0.00	1.00	0.50	1.00	1.00	-1.5500
0.00	1.00	1.00	1.00	0.00	0.00	-1.0526
0.00	0.00	0.00	0.50	0.00	0.50	-0.9930
1.00	1.00	1.00	0.00	0.00	1.00	-1.9940

Of the 15 finalist processes found by the random search procedure, only the three associated with the product cubic correlation function were admissible. All three had the same correlation parameters; we used the likelihood criterion to specify μ =-1.946 and σ = 1.005. The correlation parameters were:

j	ρ_j	γ_j
1	0.938	0.571
2	0.960	0.596
3	0.864	0.488
4	0.757	0.559
5	0.806	0.719
6	0.890	0.506

Because this particular computer model is relatively fast running, it was feasible to evaluate the predictive process at 100 test sites, chosen randomly in the 6-cube. On these sites, the empirical mot mean squared error was 0.163 and the maximum absolute error was 0.369. For comparison, we also fit a quadratic polynomial in 6 dimensions by the method of least squares. The value of R^2 was 0.9993, indicating a very close fit to the observed data, although this is due in part to the large number of terms in the polynomial (28) relative to the number of observed sites (32). At the 100 random test sites, the empirical mot mean squared error of the fitted values was 0.206, and the maximum absolute error was 0.406.

Although the mean of our predictive process did fairly well, the 95% probability bounds implied a greater degree of certainty than was warranted. At the 100 test sites, the predictive standard error was typically between .06 and .07.

We carry this example a bit further by doing a "predictive factorial analysis," in which the main effects and interactions of the design variables am estimated. This information. which exposes some of the main features of the response surface, is not available directly from me observations, but can be predicted by considering the 64 sites at the comers of the cube $[0, 1]^6$. Each factorial effect is a linear combination of the responses at these sites, and therefore has a normal predictive distribution whose mean and variance can be calculated in the usual way, using (2.2)-(2.4) to supply the means, variances, and covariances of the components.

Of the 63 main effects and interactions, those whose magnitude exceeded twice their standard deviation are given in Table 5.3. Factorial effects here are defined as in Box, Hunter, and Hunter (1978, Chapter 10). Each is a linear combination of the $\hat{y}$ values at the 64 comers of T, where the coefficients in each linear combination are $\pm 1/32$.

Table 5.3. Largest Effects in the Predictive Factorial Analysis, After 32 Runs in the Circuit Simulator Experiment.

Effect	Mean	Std. Dev.
5	0.762	0.023
3	-0.720	0.024
6	-0.672	0.021
2	0.416	0.021
4	0.217	0.025
46	0.189	0.035
36	-0.148	0.020
13	-0.112	0.033
16	-0.098	0.031
34	-0.068	0.028
25	-0.062	0.024
345	0.060	0.028
35	0.058	0.022
24	-0.057	0.032

Them is clearly a danger of overinterpreting these results, since we are predicting 63 effects from only 32 data points. However, we think it is useful to consider the largest effects, if only to suggest ways to plot the response.

Our tentative conclusions hem arc that t_3 and t_6 arc the most important variables, since they have strong main effects and occur in the largest interactions, and that t_5 has a strong effect that depends only *slightly* on the other variables. To investigate the effects of these variables in more detail, we plotted $\hat{y}(t)$ as a function of t_3 and t_6 with the other variables fixed at 0.5 (Figure 13), and j(t) as a function of t_5 , again with the other variables fixed at 0.5 (Figure 14). In Figure 14, we also show the upper and lower 95% probability bounds.

In general, we were pleased with our results in this example, especially since no special assumptions about the form of the response function were made. We expect that further development of useful correlation functions, particularly those that can exploit simplicities in the response function like approximate additivity or effect sparsity (Box and Meyer, 1986) will improve the effectiveness of Bayesian predictive methods in higher dimensions.

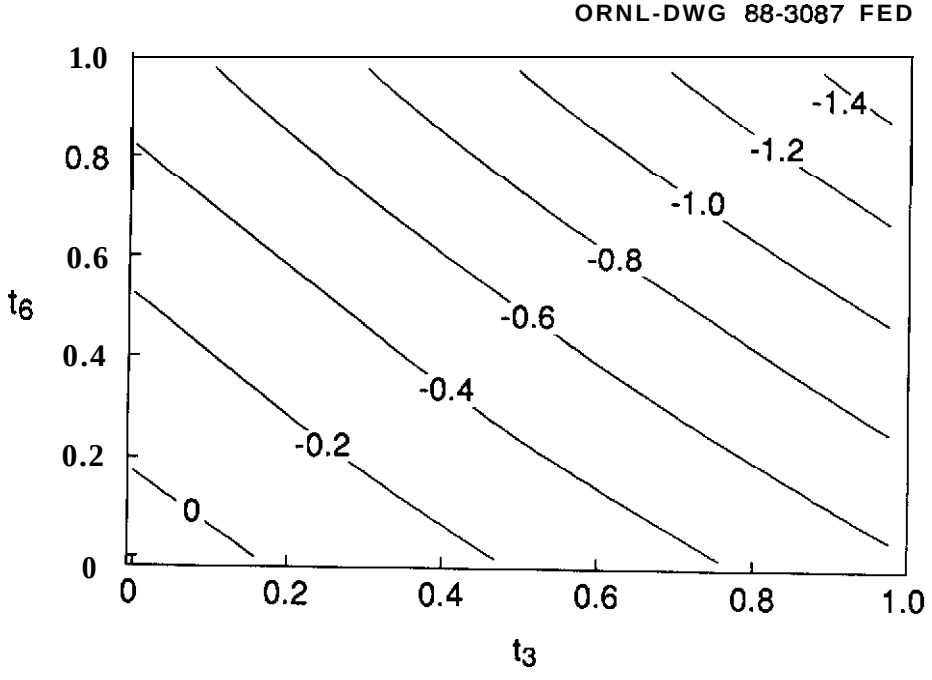


Figure 13. Contours of constant $\hat{y}(t_3, t_6)$, with $t_1 = t_2 = t_4 = t_5 = 0.5$, after running the circuit simulation **model at 32 sites.**

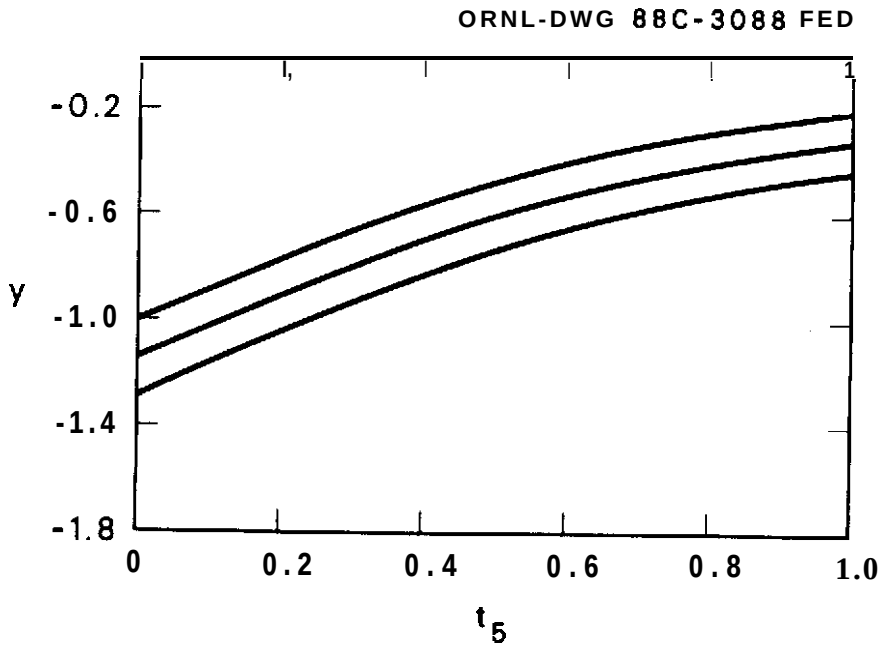


Figure 14 **Predictive mean $\hat{y}(t_5)$ and 95% predictive probability bounds $(\hat{y}(t_5) \pm 1.96\sigma_{t|D})$, where $t_1 = t_2 = t_3 = t_4 = t_6 = 0.5$, after running the circuit simulation model at 32 sites.**

6. Acknowledgements

This work has benefitted fmm our many conversations with Professor Jerry Sacks of the University of Illinois, Professor William Welch of the University of Waterloo, and Professor Henry Wynn of City University, London. This collaboration was started, and continues to be nurtured, by a series of workshops on efficient data collection, funded by a National Science Foundation grant (NSF DMS 86-098 19).

We are also grateful to the Seed Money Program at the Oak Ridge National Laboratory for funding much of the initial work that led to the preparation of this paper, and to the Oak Ridge Science and Engineering Research Semester Program, funded by the U.S. Department of Energy through Oak Ridge Associated Universities, for supporting Carla Currin's work.

Our thanks also to Professor Sung MO Kang, Tat-Kwan Yu, and Robert Buck of the University of Illinois for providing and running the circuit simulation code used in the example of Section 5.4.

7. References

[1] Baker, F. D. and Bargmann, R. E. (1985), "Orthogonal Central Composite Designs of me Third Order in the Evaluation of Sensitivity and Plant Growth Simulation Models," *Journal of the American Statistical Association* 80, 574-579.

[2] Borth, D. M. (1975), "A Total Entropy Criterion for the Dual Problem of Model Discrimination and Parameter Estimation," *Journal of the Royal Statistical Society, Series B* 37, 77-87.

[3] BOX, G.E.P. and Hill, W.J. (1967), "Discrimination Among Mechanistic Models," *Technometrics* 9, 57-70.

[4] BOX, Hunter, and Hunter (1978). *Statistics for Experimenters*, Wiley, New York.

[5] Box, G. E. P. and Meyer, R. D. (1986), "An Analysis for Unreplicated Fractional Factorials," *Technometrics 28,* 11-18.

[6] Cressie, N. (1986), "Kriging Nonstationary Data," *Journal of the American Statistical Association* 81, 625-634.

[7] Galil, Z. and Kiefer, J. (1980), "D-Optimum Weighing Designs," *Annals of Statistics 8,* 1293-1306.

[8] Geisser, S. and Eddy, W. F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association* 74, 153-160; correction 15,765.

[9] Kimeldorf, G. S. and Wahba, G. (1970), "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines," *Ann. Math. Statist.* 41, 495-502.

[10] Kitanidis, P. K. (1986), "Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis," *Water Resources Research* 22, 499-507.

[11] Lindley, D. V. (1956). "On a Measure of the Information Provided by an Experiment," *Ann. Math. Statist.* 27, 986-1005.

[12] Mitchell, T. J. (1974). "An Algorithm for the Construction of 'D-Optimal' Experimental Designs," *Technometrics* 16, 203-210.

[13] Mitchell, T. J.. Morris, M. D. and Ylvisaker. D. (1988), "Existence of Smoothed Stationary Processes on an Interval," unpublished manuscript.

[14] Mitchell, T. J. and Scott, D. S. (1987). "A Computer Program for the Design of Group Testing Experiments," *Communications in Statistics - Theory and Methods* 16, 2943-2955.

[15] Parzen, E. (1962), *Stochastic* Processes, Holden-Day, San Francisco.

[16] Ripley, B. (1981). *Spatial Statistics,* Wiley, New York.

[17] Sacks, J. and Schiller, S. (1987). "Spatial Designs," *Fourth Purdue Symposium on Statistical Decision Theory and Related Topics,* ed. S.S. Gupta, Academic Press.

[18] Sachs, J., Schiller, S.B., and Welch, W.J. (1987), "Designs for Computer Experiments," Technical Report #1, Department of Statistics, University of Illinois.

[19] Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal* 27, 379-423, 623-656.

[20] Shewry, M. C. and Wynn, H. P. (1986). "Maximum Entropy Sampling," Technical Report No. 2. The Statistical Laboratory, City University, London.

[21] Welch, W. J., Yu, T.-K., Kang, S. M., and Sacks, J. (1988). "Computer Experiments for Quality Control by Parameter Design," Technical Report #4, Department of Statistics, University of Illinois.

[22] Ylvisaker, D. (1975). "Designs on Random Fields," *A Survey of Statistical Design and Linear Models,* J.N. Srivastava, ed., North-Holland, Amsterdam.

[23] Ylvisaker, D. (1987). "Prediction and Design," *Ann. Statist.* 15, 1-19.

ORNG6498

INTERNAL DISTRIBUTION

1. S. M. Bartell
2. C. K. Bayne
3. B. Butler
4-8. c. Currin
9. B. Dory
10. D. J. Downing
11. J. B. Drake
12. R. H. Gardner
13. L. J. Gray
14-15. R. F. Harbison/ Mathematical Sciences Library
16-20. J. K. Ingersoll
21-25. F. C. Maienschein
26. J. McGrory
27-31. T. J. Mitchell
32-36. M. D. Morns
37. E. Oblow
38-42. R. C. Ward
43. R. Wood
44. B. Worley
45. A. Zucker
46. J. J. Doming (Consultant)
47. R. M. Haralick (Consultant)
48. Central Research Library
49. K-25 Plant Library
50. ORNL Patent Office
51. Y-12 Technical Library/ Document Reference Section
52-53. Laboratory Records - RC
54. Laboratory Records Department

EXTERNAL DISTRIBUTION

55. Dr. Donald M. Austin, ER-7, Applied Mathematical Sciences, Scientific Computing Staff, Office of Energy Research, Office G-437 Germantown, Washington, D.C. 20545

56. Dr. Kathryn Chaloner, Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota 55108

57. Dr. Norman Draper, Department of Statistics, University of Wisconsin, 1210 West Dayton Street, Madison, Wisconsin 53706

58. Dr. William DuMouchel, BBN Software Products Corporation, 10 Fawcett Street, Cambridge, Massachusetts 02238

59. Dr. Robert Easterling, Sandia National Laboratories, Div. 7223, Albuquerque, New Mexico 87185

60. Dr. Walter Federer, Biometrics Unit, Plant Breeding & Biometry, Cornell University, 337 Warren Hall, Ithaca, New York 14853

61. Dr. William Fellner. E.I. du Pont de Nemours and Company, P.O. Box 6090, Newark, Delaware 19714-6090

62. Dr. Jerome Friedman, Department of Statistics, Stanford University, Stanford, California 94305

63. Dr. Alan George, Vice President, Academic and Provost, Needles Hall, Waterloo, Ontario, CANADA N2L 3G1

64. Dr. Prem Goel, Statistics Department, Ohio State University, 1958 Neil Avenue, Columbus. Ohio 43210

65. Dr. Ronald Iman, Sandia National Laboratory, Div. 6415, Albuquerque, New Mexico 87185

66. Dr. Mark Johnson, Operations Research Department, Georgia Institute of Technology, Atlanta, Georgia 30332

67. Dr. Robert Lamer, Army Research Office, Research Triangle Park, North Carolina 277092211

68. Dr. L. M. Moore, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, New Mexico 87545

69. Professor W. Mueller, I.I.A.S.A., Laxenburg. A-2361 AUSTRIA

70. Dr. Christopher Nachtsheim, School of Management, University of Minnesota, Minneapolis, Minnesota 55455

71. R. Nagtegaal, Volvo Car B.V., Postbox 101.5, Hehnond 5700 MC, THE NETHERLANDS

72. Dr. Art Owen, Stanford University, Department of Statistics, Sequoia Hall, Stanford, California 943054065

73. Dr. Nelson Pacheco, The MITRE Corporation, 1259 Lake Plaza Drive, Colorado Springs, Colorado 80906

74. Dr. Ronald Peierls, Applied Mathematics Department, Brookhaven National Laboratory, Upton, New York 11973

75. Professor Derek Pike, Department of Applied Statistics, University of Reading, P.O. Box 217, Reading, Berkshire RG62AN, ENGLAND

76. Prof.Dr. Friedrich Pukelsheim, Institut für Mathematik, Universität Augsburg, Memminger Strasse 6. D-8900 Augsburg, GERMANY

77. Professor Baldev Raj, School of Business and Economics, Wilfrid Laurier University, Waterloo, Ontario, CANADA N2L 3C5

78. Dr. Jerome Sacks, Department of Statistics, University of Illinois, 101 Illini Hall, 725 South Wright Street, Champaign. Illinois 61820

79. Susamah B. Schiller, National Bureau of Standards, Gaithersburg, Maryland 20899

80. Dr. L. R. Shenton, Office of Computing and Information Service, Boyd Graduate Studies Building, University of Georgia, Athens, Georgia 30602

81. Dr. AM Shoemaker, Quality Assurance Center, AT&T Bell Laboratories, Room 2K-501, Crawfords Comer Road, Holmdel, New Jersey 07733

82. Dr. N. D. Singpurwalla, School of Engineering and Applied Science, Staughton Hall, 707 22nd Street, N.W., Washington, D.C. 20052

83. Dr. Alan Solomon, 90 Eshel Street, Omer 84965, ISRAEL

84. Dr. Robert A. Stokes, Energy Systems Department, Pacific Northwest Laboratory, P.O. Box 999, Richland, Washington 99352

85. Dr. Ray A. Waller, S-l. Statistics, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, New Mexico 87545

86. Dr. William Welch, Department of Statistics, University of Waterloo, Waterloo, Ontario. CANADA N2L 3G1

87. Major Brian Woodruff, AFOSR/NM, Bolling AFB, Washington, D.C. 20332

88. Professor Henry Wynn, Department of Mathematics, The City University, Northhampton Square, London, EC1V OHB, ENGLAND

89-93. D. Ylvisaker, Department of Mathematics, Universtiy of California at Los Angeles, Los Angeles, California 90024

94. Dr. Mark A. Youngren, U.S. Army Concept Analysis Agency, 8120 Woodmont Avenue, Bethesda, Maryland 20814-2797

95. Office of Assistant Manager for Energy Research and Development, Department of Energy, Oak Ridge Operations Office, P.O. Box 2001, Oak Ridge, Tennessee 37831-8600

96-105. Technical Information Center