Genetic Function Approximation Experimental Design (GFAXD): A New Method for Experimental Design

Thomas R. Kowar

G. D. Searle/Monsanto Life Sciences, 4901 Searle Parkway, Skokie, Illinois 60077

Received April 6, 1998

The application of Roger's Genetic Function Approximation (GFA) algorithm to experimental design data produces a population of model equations that contains the same regression equation as derived using Statistical Experimental Design analysis. GFA is able to identify this regression equation from fewer experiments than required by Statistical Experimental Design without the loss of information about higher order effects. A proposal for a novel method for the design, conduct, and analysis of experiments, Genetic Function Approximation Experimental Design (GFAXD), is presented. An example application of the GFAXD method to a theoretical 12-variable experimental design problem demonstrates the potential of this new method to significantly increase the productivity of designed experimentation.

INTRODUCTION

Statistical experimental design is a powerful method for efficiently extracting information from a group of related experiments. These experiments simulate the operation of some process under specific sets of conditions. The process may be any series of operations or combination of components that produces an outcome such as the yield of a chemical reaction, the dissolution rate of a pharmaceutical dosage form, or the thermal conductivity of a metallurgical formulation.

Also known as design of experiments (DOE), the experimental design method seeks to identify those independent variables $\{x_i\}$ that govern the outcome of the experiment and to mathematically model the relationship between independent and dependent variables $\{y_j\}$ of the process under study. The mathematical model is used to increase the understanding of the process and to predict the performance of the process under conditions not yet examined by experiment. Experimental design methods can evaluate either categorical or continuous variables. However, the method assumes that continuous variables function over the entire variable ranges evaluated by the study.

There are numerous papers and monographs that address various aspects of statistical experimental design.[1−4] Additionally, there are commercial computer programs that facilitate the design and analysis of statistically designed experimental programs.[5−10]

Experimental design methodology is used extensively in industrial research and process development for the characterization and validation of production processes. The experimental design tool allows the experimentalist to accumulate information about a process rapidly and efficiently while minimizing experimentation costs. The competitive nature of industry, however, demands further improvements in efficiency via reduction of process development time cycles and costs.

In this paper we report on a study of the use of the Genetic Function Approximation (GFA) algorithm[11] to accomplish objectives similar to those of statistical experimental design. We present a novel proposal on how more efficient experimental programs may be conducted and demonstrate how analysis of those experiments using the GFA algorithm allows the experimentalist to increase the quality and quantity of information derived. This new method is known as Genetic Function Approximation Experimental Design (GFAXD).

METHODS

Computation. Experimental design factor effect calculations and analysis of variance (ANOVA) calculations were conducted using the Design-Ease software package.[5] GFA calculations were conducted using the Cerius[2] QSAR+ software package.[12]

Experimental Design Procedure. The form of the mathematical relationship between independent and dependent variables derived from an experimental design study is a linear polynomial equation as shown in eq 1

$$y_j = a_0 + \sum a_i x_i \tag{1}$$

Typically, experimental designs specify that the experiments be conducted with the independent variables set at various combinations of high and low values. This type of design is called a two-level factorial design. The number of experiments needed for an *n*-variable two-level factorial design study is 2^n. Therefore, study of a process with four independent variables *A*, *B*, *C*, and *D* is symbolically designated as a 2^4 design and it requires 16 experiments.

Such a 16-experiment study is fully powered to determine the effects of each of the four independent variables (factors), each of the six two-factor interactions, each of the four three-factor interactions, and the one four-factor interaction. These 15 factors and the required single degree of freedom for the overall mean are equivalent in number to the 16 experimental points, thereby demonstrating the efficiency of the experimental design method. A summary of the factors and

S0095-2338(98)00205-4 CCC: $15.00
Published on Web 08/01/1998

Table 1. Factors and Multifactor Interactions of a 2^4 Design

one-factor	*A*	*B*	*C*	*D*
two-factor	*AB*	*AC*	*AD*	*BC*
	BD	*CD*		
three-factor	*ABC*	*ABD*	*ACD*	*BCD*
four-factor	*ABCD*			

multifactor interactions involved in this design is shown in Table 1. Each of the 15 factors is considered to be a potential independent variable $\{x_i\}$ for constructing the mathematical relationship expressed by eq 1.

The same variable space of a 2^4 design may also be studied in fewer than 16 experiments using a fractional factorial design. However, fractional factorial designs provide less information regarding the higher order two-, three-, and four-factor interactions.

The effect of an individual factor is calculated as the difference between the averaged values of the dependent variables when the individual factors are set at the high and low levels. For example, the effect of temperature on chemical reaction yield is determined by the difference between the average yield of the experiments in which the temperature was set at the high level and the average yield of the experiments in which the temperature was set at the low level.

The individual factor effects are plotted on normal probability paper. Those factors that do not govern the outcome of the experiment exhibit effects of the same order of magnitude as the experimental error and together these factor effects constitute a straight line centered at zero on the abscissa of the normal probability plot. This result follows from one of the central tenets of statistics that states that experimental errors are often randomly distributed and centered on a mean of zero. An example of a normal probability plot is shown in Figure 1.

Those factors that do influence the outcome of the experiment exhibit effects that are substantially larger than the experimental error and do not lie on the no-effect line of the normal probability plot. These factors $\{x_i\}$ are considered to be significant and are used as terms in an equation of the type shown in eq 1. Least-squares regression techniques are used to calculate the coefficients $\{a_i\}$ of the equation using the experimental data.

The coefficient-fitted linear polynomial equation constitutes the mathematical model of the process under study. The validity of the mathematical model is then tested by conducting carefully selected experiments.

Experimental Design Example. Box, Hunter, and Hunter[1] give an example implementation of a 2^4 design and this same problem is used as a demonstration example in the commercial experimental design computer program Design-Ease.[5] The experimental data from this example are summarized in Table 2. In this example, the yield of a chemical reaction is studied as a function of four independent variables or factors, *A* (catalyst charge), *B* (temperature), *C* (pressure), and *D* (concentration). The dependent variable is the yield of the chemical reaction and it is designated *Y*.

The calculated individual factor effects are presented on the normal probability plot shown in Figure 1. This normal probability plot shows that the factors *A*, *B*, *D*, and the two-factor interaction *BD* exhibit effects that are significantly greater than the experimental error.

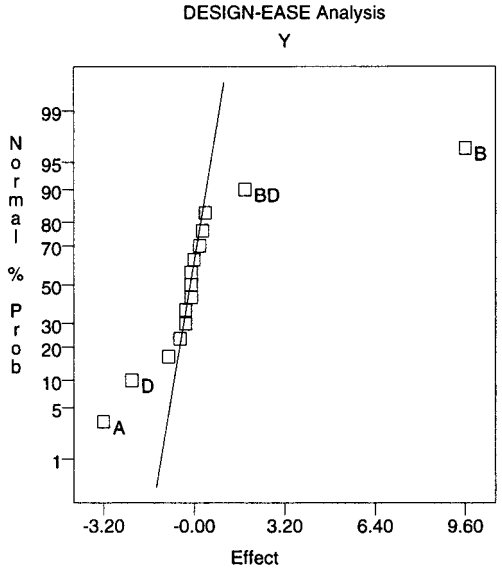

Figure 1. Example 2^4 design normal probability plot.

Table 2. Example 2^4 Design Data from Box, Hunter, and Hunter

A	*B*	*C*	*D*	*Y*
10	220	50	10	71
15	220	50	10	61
10	240	50	10	90
15	240	50	10	82
10	220	80	10	68
15	220	80	10	61
10	240	80	10	87
15	240	80	10	80
10	220	50	12	61
15	220	50	12	50
10	240	50	12	89
15	240	50	12	83
10	220	80	12	59
15	220	80	12	51
10	240	80	12	85
15	240	80	12	78

The least-squares regression of factors *A*, *B*, *D*, and *BD* against the experimental data is done as part of the ANOVA calculation and eq 2 is the derived regression equation. The multiple correlation coefficient r^2 and the *F* ratio metrics for this equation are excellent.

$$Y = 415.75 - 1.6\text{“}A\text{”} - 1.275\text{“}B\text{”} - 54.5\text{“}D\text{”} + 0.225\text{“}BD\text{”}$$

$$r^2 = 0.99 \qquad F = 194.76 \tag{2}$$

Equation 2 serves as the mathematical model that relates the independent and dependent variables. A plot of *Y* as a function *A*, *B*, *D*, and *BD* in 5-space would provide a response surface for the reaction yield in terms of these factors. Given the difficulty of visualizing such a plot, often, some variables are held constant and yield is plotted in 3-space using two of the independent variables.

The reaction conditions for optimizing or minimizing the yield may be obtained by putting the mathematical model into a spreadsheet and using the spreadsheet solver or a genetic algorithm solver[13] to adjust the values of $A-D$ to attain the maximum or minimum values of *Y*. Application of this technique to eq 2 provides the results shown in Table 3. The maximum achievable yield for this reaction is 88.75% according to the model as expressed in eq 2. The actual

Table 3. Spreadsheet Solver Derived Conditions for Maximum and Minimum Yield Using Equation 2

A	*B*	*C*	*D*	*Y*
10	240	50	10	88.75
15	220	50	12	51.25

yield under these experimental conditions was observed to be 90% according to the experimental data in Table 2.

Experimental design leaves the discrimination of factor effects to the discretion of the experimentalist. Because the no-effect line is never defined perfectly, one must judge if a factor effect is sufficiently close to the no-effect line to be considered part of the line or if it is sufficiently far from the no-effect line to be considered significant.

Commercial software packages allow the user to select those factors that are to be included in the ANOVA calculation. Figure 2 shows the normal probability plot when variable *C* is included in the analysis.

$$Y = 420.625 - 1.6\text{“}A\text{”} - 1.275\text{“}B\text{”} - 0.075\text{“}C\text{”} - 54.5\text{“}D\text{”} + 0.225\text{“}BD\text{”}$$

$$r^2 = 0.99 \qquad F = 296.77 \tag{3}$$

The regression equation resulting from inclusion of variable *C* in the ANOVA calculation is shown in eq 3 and it serves as a slightly better model of the data than does eq 2. The need to judge which factors to include in the ANOVA calculation means, in effect, that experimental design may present a number of possible mathematical models in the form of regression equations to characterize the data set. The derivation of multiple models from experimental design studies is an important concept that relates directly to the output of the GFAXD method discussed later.

Experimental Design Limitations. The major issues to address while planning the study of a process using experimental design are:

- to understand which independent variables to study
- to determine what ranges the independent variables should be allowed to assume.

Because the number of experiments needed for an *n*-variable two-level factorial design study is 2^n, the required number of experiments grows quickly. Experiments are expensive, and clearly some critical choices must be made for the economy of the experimental function. A disadvantage of experimental design is that the selection of which variables to include in a study may be incorrect and, consequently, important independent variables may be excluded from the design.

In chemistry as well as most other disciplines, the starting point for an experimental design is not always known but rather it must be defined by conducting exploratory experiments. The exploratory experiments help to define reasonable operating conditions for the process and also serve to identify those regions of variable space that are not likely to be productive. The number of exploratory experiments varies with the complexity of the process and also with the experimentalist's knowledge of the process.

Often, the conditions under which the exploratory experiments are conducted do not coincide with the conditions suggested by the experimental design. The facility of calculation of factor effects is compromised by exploratory experiment factor values that lack orthogonal propriety.

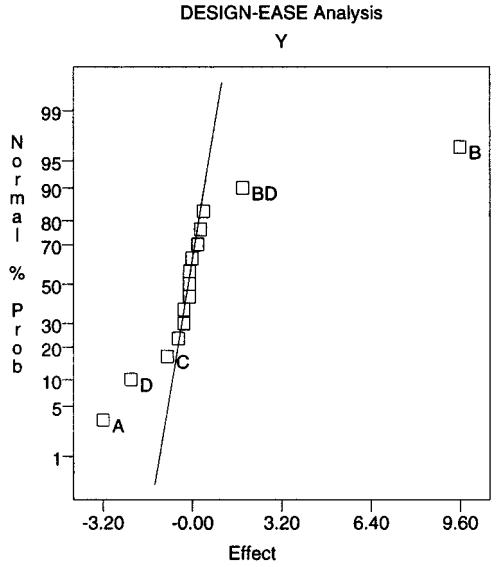

Figure 2. Example 2^4 design normal probability plot including variable C.

Therefore, the use of the data from the exploratory experiments together with the experimental design data to derive the mathematical model is difficult and not well defined. Consequently, an appreciable portion of the information contained in the exploratory experiments may be lost. This loss of information possibly may be justified by a concern that the relationship between the independent and dependent variables is not operational across the entire variable space connecting the designed experiments and, at least some of, the exploratory experiments.

Process development efficiency would be increased considerably if some or all of the exploratory experimental data could be used in conjunction with, or as a replacement for, the experimental design data. Further efficiency gains would be realized if operational discontinuities in the independent variable ranges could be identified and quantitated. The use of GFA as embodied in the GFAXD method offers the potential for accomplishing these efficiency improvements.

GFA Algorithm Methodology. The genetic algorithm, originated by Dr. John Holland, is a mathematical technique that is based on the natural process of evolution.[14] Genetic algorithms are being applied to solve a variety of problems in science and engineering.[15] Studies of the use of genetic algorithms in experimental design have been reported.[13,16] Additionally, the application of genetic algorithms to industrial optimization problems is growing.[17−19]

An elegant application of genetic algorithms to problems related to experimental design is embodied in the GFA algorithm[11] developed by Dr. David Rogers. The GFA algorithm generates a population of models that are used to represent a set of experimental data. The population of models is optimized by evolution using a genetic algorithm.[20]

Rogers and Hopfinger[21] demonstrated the usefulness of GFA for defining Quantitative Structure−Activity Relationships (QSAR) and Quantitative Structure−Property Relationships (QSPR). QSAR and QSPR methods quantitate the relationship between molecular structure descriptors and biological activity or physical property characteristics of molecules, respectively. The molecular descriptors may be viewed as a set of independent variables and the biological

activity or physical property viewed as the dependent variable.

In QSAR and QSPR, the number of molecular descriptors that can be calculated from a series of molecular structures far exceeds the number of available compounds with biological or physical property measurements. Therefore, a method is needed to test subsets of the molecular descriptors against the experimental data with the goal of eventually finding which of the molecular descriptors can be quantitatively related to the experimental data. Rogers and Hopfinger used GFA to accomplish the testing of subsets of the molecular descriptors for the determination of the mathematical models.

The GFA algorithm generates an initial population of putative equations by the random selection of molecular descriptors (factors). The form of the equations is that of a linear polynomial and is exactly the same type of equation as used in experimental design and as shown above in eq 1. The length of the equations is determined by the number of molecular descriptors selected and is allowed to vary. Each equation is fit to the experimental data using linear least-squares regression techniques, and the equations are ranked according to quality using Friedman's lack of fit (LOF) measure,[22] which is a penalized least-squares error statistic (eq 4). LOF is the least-squares error (LSE) of the model adjusted by a term comprised of the number of factors (f) in the model, a user-defined parameter (d), the total number of parameters (p) used in the model (factors and data points used), and the number of experimental data points (n).

$$\mathrm{LOF} = \mathrm{LSE}/A^2 \tag{4}$$

where $A = (1 - [f + dp]/n)$.

The GFA algorithm accomplishes the breeding of the best equations and the elimination of the poorer equations using a genetic algorithm. The genetic algorithm cuts and separates individual equations and then recombines the fragments to form new equations. Additionally, mutation of the equations is implemented stochastically by the genetic algorithm. The genetic algorithm uses the equation LOF to select equations for breeding and survival. Use of the LOF measure drives the population toward more parsimonious, simple models and avoids over-fitting of the data. As generations of equations are bred and mutated, the population evolves to a series of ever-increasing quality equations.

The output of a GFA calculation is a population of regression equations that model the relationship between the independent and dependent variables. Each equation is constituted by a different combination of factors. The relationship that represents the underlying physical model of the process is embodied in many of these equations. Other equations in the population, however, are artifacts that relate the independent and dependent variables mathematically but do not have any physical basis. The experimentalist must interpret the population of equations and select the best models based on domain-specific knowledge.

The GFA method does not require the assumption that the relationship between independent and dependent variables operates over the entire variable range. GFA circumvents the need for this assumption by using spline-based terms for the construction of its regression equations. A spline term is designated as $\langle X - a \rangle$ where X is an independent variable and a is a constant called the knot of the spline. The value of the term $\langle X - a \rangle$ is equal to zero if the quantity $(X - a)$ is zero or negative and equal to the quantity $(X - a)$ if that quantity is positive. The use of spline-based terms allows discontinuous variable ranges to be examined and provides an effective means of partitioning a variable range.

An example of the productive use of a spline term would be to characterize the performance of a hydrogenation catalyst that is poisoned by a contaminant in the chemical reaction mixture. The first 20% of the catalyst charge might be rendered ineffective by the catalyst poison. A study of the effect of the amount of catalyst on the hydrogenation reaction yield could be modeled well with a spline term but could not be modeled properly with a simple linear term.

RESULTS AND DISCUSSION

The Use of GFA in Experimental Design. QSPR (QSAR) and experimental design methods seek similar objectives:

- isolate the independent variables that govern the outcome of a process from those which do not affect the outcome;
- derive linear polynomial regression equations that constitute robust mathematical models relating the governing independent variables to the dependent variables;
- use the mathematical models to predict future, optimized experimental outcomes.

The successful use of the GFA algorithm in QSPR (QSAR) suggested it might be productively applied to experimental design. Investigation of the use of GFA for accomplishing experimental design objectives was initiated in an attempt to make process development more efficient. In particular, we wanted to develop a method that allows the extraction of more of the information contained in exploratory experimental data.

The appropriate starting point for this study was the application of the GFA algorithm to the 2^4 design example of Box, Hunter, and Hunter. Our initial goal was to determine if the same model, eq 2, could be determined using GFA. The data was analyzed using the GFA software default values. The evolution of the 100-equation population was allowed to proceed for 10 000 generations. A summary of the first 20 equations produced by the GFA analysis of the example data is given in Table 4.

Equation 2 was identified as the 18th best equation in the population of 100 equations produced by the GFA algorithm using the LOF measure to rank the equations. This equation was presented with a multiple correlation coefficient of $r^2 = 0.99$ and a Fisher ratio of $F = 194.76$. These statistical metrics for the equation are identical to those produced by the ANOVA calculations and this is the expected outcome because both methods employ the same technique for determining the regression equation coefficients.

The second model, eq 3, was identified as the fourth best equation in the population and also was presented with a multiple correlation coefficient and a Fisher ratio that were identical to those produced by the ANOVA calculations.

An important concept to emphasize is that the **experimental design regression equations are usually members of the GFA population of equations.** The GFA method provides the same information as experimental design and may provide additional information as well.

Table 4. Summary of GFA Equations for Example 2^4 Design

equation		r^2	F
4-1	$Y = 415.75 - 1.6$"A" $- 54.5$"D" $+ 0.226975$"BD" $- 1.275$"B" $- 3e^{-05}$"BCD"	0.993750	318.072139
4-2	$Y = 415.75 - 1.6$"A" $- 54.5$"D" $+ 0.225$"BD" $- 1.25333$"B" $- 0.000333$"BC"	0.993645	312.719101
4-3	$Y = 415.75 - 1.6$"A" $+ 0.225$"BD" $- 54.056$"D" $- 1.275$"B" $- 0.006831$"CD"	0.993392	300.676705
4-4	$Y = 420.625 - 0.075$"C" $- 1.6$"A" $- 54.5$"D" $+ 0.225$"BD" $- 1.275$"B"	0.993306	296.773333
4-5	$Y = 395.75 - 0.144262$"AD" $- 52.6967$"D" $+ 0.226975$"BD" $- 1.275$"B" $- 3e^{-05}$"BCD"	0.993002	283.806629
4-6	$Y = 395.75 - 0.000333$"BC" $+ 0.225$"BD" $- 0.144262$"AD" $- 52.6967$"D" $- 1.25333$"B"	0.992896	279.530730
4-7	$Y = 395.75 - 0.144262$"AD" $+ 0.225$"BD" $- 52.2527$"D" $- 1.275$"B" $- 0.006831$"CD"	0.992643	269.855211
4-8	$Y = 395.75 - 0.006906$"AB" $- 54.5$"D" $+ 0.226975$"BD" $- 1.18868$"B" $- 3e^{-05}$"BCD"	0.992590	267.909115
4-9	$Y = 400.625 - 0.075$"C" $+ 0.225$"BD" $- 0.144262$"AD" $- 52.6967$"D" $- 1.275$"B"	0.992557	266.702182
4-10	$Y = 395.75 - 1.16701$"B" $- 0.006906$"AB" $- 54.5$"D" $- 0.000333$"BC" $+ 0.225$"BD"	0.992484	264.092490
4-11	$Y = 127.567 - 1.6$"A" $- 28.5163$"D" $+ 0.112027$"BD" $- 0.000339$"BC"	0.986292	197.867224
4-12	$Y = 411.341 - 2.4e^{-05}$"ABC" $- 1.24728$"A" $- 54.5$"D" $+ 0.225$"BD" $- 1.25583$"B"	0.992246	255.941831
4-13	$Y = 395.75 - 0.006906$"AB" $- 54.056$"D" $+ 0.225$"BD" $- 1.18868$"B" $- 0.006831$"CD"	0.992231	255.432704
4-14	$Y = 122.5 - 3e^{-05}$"BCD" $- 1.6$"A" $- 28.0594$"D" $+ 0.112016$"BD"	0.986140	195.662103
4-15	$Y = 411.38 - 1.25043$"A" $+ 0.225$"BD" $- 54.1028$"D" $- 1.275$"B" $- 0.000489$"ACD"	0.992203	254.510126
4-16	$Y = 122.459 - 0.006942$"AB" $+ 0.117865$"BD" $- 29.859$"D"	0.978302	180.348267
4-17	$Y = 127.519 + 0.119849$"BD" $- 0.006942$"AB" $- 30.3152$"D" $- 0.000339$"BC"	0.986111	195.254112
4-18	$Y = 415.75 - 54.5$"D" $+ 0.225$"BD" $- 1.275$"B" $- 1.6$"A"	0.986076	194.756410
4-19	$Y = 400.625 - 1.18868$"B" $+ 0.225$"BD" $- 0.006906$"AB" $- 54.5$"D" $- 0.075$"C"	0.992145	252.603610
4-20	$Y = 415.75 - 54.5$"D" $+ 0.225$"BD" $- 0.005385$"AC" $- 1.275$"B" $- 1.25$"A"	0.992132	252.191972

There may be some discomfort with the fact that there are regression equations that appear to represent the data in Table 1 better than do either eq 2 or eq 3. This result raises a fundamental question when using the GFA algorithm for experimental design. How does one separate the mathematical artifacts from the equations that represent the underlying physical model of the process? A simple method to separate the mathematical artifacts from the equations representing the underlying physical model is to rerun the GFA determination of the regression equations. Because the GFA process is stochastic, each run may give a different set of equations. The true, underlying model equations are expected to appear and to be highly ranked in all populations given the assumption that the genetic algorithm has had the opportunity to evolve sufficiently. The distribution of the mathematical artifact equations is likely to change on iterative GFA runs. Further, the genetic algorithm parameters may also be changed between runs to present opportunities for the development of different populations of equations.

A qualitative evaluation of the equation population in Table 4 suggests that those factors that appear frequently may constitute the significant factors of the underlying physical model. An analysis of the frequency of appearance of factors in the population of equations is revealing in that the same factors, *A*, *B*, *D*, and *BD* are appearing throughout. In fact, *D* and *BD* appear in all 20 equations, whereas *B* appears in 16 equations and *A* appears in 10 equations. Factor *C* appears in only three of the first 20 equations, suggesting that eq 4-18 (eq 2) may more likely be an equation that represents the underlying physical model than equation 4-4 (eq 3).

It is also revealing to note the similarity of the values of the coefficients of these repeating factors. This similarity suggests that many of the equations are simply variations on a theme. The theme expressed is either an equation that represents the underlying physical model of the process or it represents a competing mathematical artifact. For example, in Table 4, eq 4-1 is nearly identical to eq 4-18 (eq 2) except that the first equation contains a *BCD* term. This situation is similar to the experimental design example where eq 3 was a variation on the theme expressed in eq 2. Thus, by analogy to experimental design, the *BCD* term in eq 4-1 could be viewed as lying close to the no-effect line of a normal probability plot.

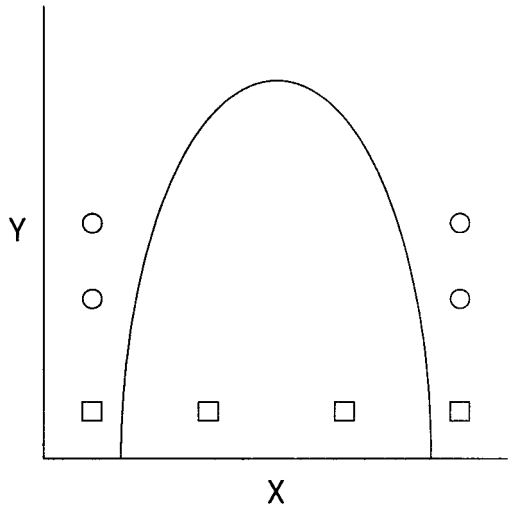

Figure 3. One dimension experimental design strategies.

Further examination of the equations in Table 4 shows that there are two types of equation themes expressed in the data set. In the first theme, the equations begin with a constant that lies between 395 and 420 (all equations except 4-11, 4-14, 4-16, and 4-17). Each equation expressing this theme contains minimally the terms *B*, *D*, and *BD*, and each of these terms is associated with approximately the same coefficient in all the equations in the family of equations expressing this theme. The second equation theme is characterized by equations beginning with a constant between 122 and 127 (eqs 4-11, 4-14, 4-16, and 4-17). The equations in this family contain minimally the terms *D* and *BD* and, again, each of these terms is associated with approximately the same coefficient in all the equations. However, the coefficients in this family are different from those of the family expressing the first theme.

One could "average" the equations in each equation theme family as proposed by Rogers.[23] Each constant or coefficient would be summed over all members of the family and then divided by the number of members of the family to derive average values for the constant and the coefficients. The averaged equations could be used to determine the theme

that best expresses the underlying physical model of the data. Averaged equations from several families could be compared regarding their ability to predict the outcome of future experiments. The results of carefully chosen experiments would allow discrimination amongst the averaged equations and would lead to the selection of the best equation theme to represent the underlying physical model of the process.

Truncated versions of the 2^4 design example data set were also evaluated. GFA analysis of the first 12 and the first 14 experiments using 20 000 generations provided approximations to eq 2 as the first equation in both populations. The 12-experiment test gave eq 5 and the 14-experiment test afforded eq 6. These approximate equations correctly identify the factors that govern the outcome of the experiments. These tests reveal the true power of the GFA method. GFA did not uncover an approximation to eq 2 when only the first 10 experiments were analyzed.

$$Y = 524.917 - 1.63333\text{“}A\text{”} - 1.775\text{“}B\text{”} - 65.375\text{“D”} + 0.275\text{“}BD\text{”}$$

$$r^2 = 0.990844 \qquad F = 189.390698 \qquad (5)$$

$$Y = 539.857 - 66.875\text{“}D\text{”} + 0.28125\text{“}BD\text{”} - 1.62857\text{“}A\text{”} - 1.8375\text{“}B\text{”}$$

$$r^2 = 0.992957 \qquad F = 317.196955 \qquad (6)$$

In this particular example, GFA was able to identify approximate versions of the underlying regression equation from fewer experiments than experimental design without loss of information about higher order effects.

GFAXD: A New Method for the Design, Conduct, and Analysis of Groups of Experiments. We have shown that GFA analysis of experimental results gives the same regression equations as the experimental design analysis. We now focus on the way in which experiments are designed and propose a novel approach on how to develop more efficient experimental designs.

Ideally, during the early phase of process development, as many variables as possible should be studied while conducting the exploratory experiments. Further, the values of those variables that are not being purposefully controlled during the experiment also should be observed and recorded. In such a scenario, one may accumulate information for many variables in a relatively few number of experiments.

It is appropriate to evaluate how simple variable relationships are studied. Consider the two-variable model of a process shown in Figure 3. The dependent variable *Y* is a function of a single independent variable *X*. A classical experimental design study of this simple case would suggest a strategy that evaluates a low and a high value of the independent variable as depicted by the circles in Figure 3. This strategy would result in a conclusion that there is no relationship between the independent and dependent variables in this example. However, if resources were available to conduct four experiments, then most experimentalists would study the entire range of the variable space by doing experiments as depicted by the squares in Figure 3 and not by duplicating the high and low values as might be suggested by traditional experimental design. This strategy would reveal some aspect of the relationship between the independent and dependent variables. The strategy to cover the entire range of a variable is intuitive and should be extended to studies where many variables are being evaluated simultaneously.

Table 5. Synthetic Rendition of Example 2^4 Design Data Set

A	*B*	*C*	*D*	*Y*
14.82	226.03	60.23	10.41	65.92
13.91	238.88	63.74	10.55	80.99
12.31	222.27	62.40	11.09	62.87
14.72	224.36	68.32	11.26	60.89
13.02	232.11	72.34	11.61	72.56
14.39	220.69	54.77	10.25	61.69
14.61	222.31	59.02	11.31	58.26
10.93	231.74	76.23	10.16	78.83
13.86	233.28	51.09	11.75	72.50
10.32	226.30	52.01	10.09	74.56
10.79	228.68	64.23	11.11	73.07
14.11	233.33	63.94	11.35	72.97
12.29	236.70	77.36	10.80	80.87
11.89	226.06	78.83	10.23	71.30
12.75	236.26	74.37	10.20	80.44
11.47	239.45	65.77	11.43	84.97

The GFAXD method proposes that experimental designs be structured such that variable ranges are set as wide as possible in the context of physical constraints and domain knowledge. Further, the variable settings are to be generated randomly and are to cover the entire variable range. Stochastic coverage of the entire, practical variable range provides an unbiased sampling of the unknown variable space. It also maximizes the opportunity for variables to interact in interesting ways during experimentation and provides an opportunity to partition the variable space, if necessary, by use of the GFA spline facility.

Because one random number is as good as any other random number, incidental values that approximate the target values for the variable settings can be used without penalty. This feature of the GFAXD method should substantially facilitate the actual conduct of experiments. Random generation of variable settings may be modified if domain knowledge mandates the selection of particular variable values.

The essential features of the GFAXD method are:

- all potential independent variables, both controlled and noncontrolled, are studied;
- the entire practical range of each controlled independent variable is studied;
- target values for controlled independent variables are generated randomly;
- accurate approximations to target values are determined for controlled independent variables;
- accurate incidental values are determined for noncontrolled independent variables;
- the GFA algorithm is used to analyze the results of the experiments;
- the experimentalist evaluates the GFA population of models and uses scientific reason and validation experiments to select the best process models.

Application of GFAXD to Experimental Design Example. The GFAXD experimental design strategy was applied to the example 2^4 design problem that we have been considering. We created a synthetic spreadsheet rendition of the example data using randomly generated values for the independent variables and using eq 2 to define the yield. For example, *A* (catalyst charge) was maintained between

Table 6. Synthetic 12-Variable Problem Design Data

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*Y*
0.05	0.31	0.85	0.78	0.58	0.59	0.19	0.75	0.73	0.45	0.63	0.12	43.20
0.25	0.98	0.08	0.86	0.92	0.79	0.61	0.32	0.27	0.83	0.93	0.46	71.18
0.20	1.00	0.18	0.67	0.76	0.44	0.49	0.97	0.82	0.47	0.89	0.95	50.86
0.23	0.58	0.64	0.07	0.52	0.84	0.08	0.73	0.54	0.57	0.85	0.26	18.91
0.57	0.30	0.16	0.30	0.41	0.89	0.27	0.85	0.37	0.91	0.62	0.82	34.22
0.83	0.78	0.99	0.05	0.41	0.30	0.09	0.59	0.35	0.13	0.59	0.28	14.19
0.10	0.27	0.97	0.15	0.31	0.29	0.16	0.87	0.41	0.46	0.21	0.14	20.88
0.99	0.94	0.53	0.75	0.30	0.23	0.88	0.68	0.43	0.37	0.23	0.09	66.62
0.32	0.92	0.15	0.37	0.64	0.56	0.33	0.56	0.96	0.43	0.34	0.85	31.53
0.10	0.77	0.46	0.45	0.25	0.00	0.74	0.62	0.62	0.36	0.11	0.80	49.30
0.81	0.11	0.62	0.11	0.51	0.19	0.56	0.86	0.73	0.21	0.23	0.89	31.30
0.91	0.28	0.36	0.87	0.56	0.28	0.78	0.65	0.41	0.99	0.93	0.07	83.66
0.34	0.67	0.91	0.76	0.06	0.57	0.11	0.84	0.32	0.05	0.90	0.97	34.51
0.21	0.22	0.32	0.05	0.25	0.37	0.92	0.63	0.26	0.65	0.30	0.45	56.07
0.81	0.03	0.76	0.23	0.74	0.94	0.22	0.83	0.00	0.05	0.16	0.37	17.31
0.31	0.04	0.16	0.76	0.36	0.00	0.65	0.94	0.37	0.42	0.22	0.71	59.70

Table 7. Equation Set I of GFAXD Analysis of 12-Variable Problem

equation		r^2	*F*
7-1	$Y = 10.175 + 38.835\langle$"*D*" $- 0.23\rangle + 29.2929$"*G*" $+ 9.20109$"*J*" $+ 21.3923$"*GJ*"	0.999854	18888.717318
7-2	$Y = 10.0557 + 38.7575\langle$"*D*" $- 0.23\rangle + 29.7823$"*G*" $+ 20.3488$"*GJ*" $- 0.985681$"*JL*" $+ 10.0333$"*J*"	0.999903	20630.735001
7-3	$Y = 9.52469 + 33.453$"*G*" $+ 12.4538$"*J*" $+ 39.2353\langle$"*D*" $- 0.23\rangle + 16.4917\langle$"*GJ*" $- 0.14\rangle$	0.999849	18219.354372
7-4	$Y = 13.2583 + 29.2979\langle$"*G*" $- 0.09\rangle + 38.8802\langle$"*D*" $- 0.23\rangle + 21.5772$"*GJ*" $+ 9\ .04928\langle$"*J*" $- 0.05\rangle$	0.999871	21351.123231
7-5	$Y = 9.65577 + 39.1522\langle$"*D*" $- 0.23\rangle + 16.7702\langle$"*GJ*" $- 0.12\rangle + 33.0211$"*G*" $+ 12.1414$"*J*"	0.999812	14642.886731

10−15 and each value of *A* was determined using a random number generator. Likewise, values for *B*, *C*, and *D* were generated randomly and the higher order interactions and the yields were calculated. This data set is summarized in Table 5.

The data in Table 5 were evaluated using GFA. The GFA analysis was allowed to proceed to 20 000 generations while employing the GFA default parameters. We evaluated the first nine experiments as well as the total 16-experiment data set. In these examples, GFAXD was able to discern the underlying equation of the example 2^4 design problem. Equation 7 was derived from the 16-experiment data set and eq 8 was derived from the 9-experiment data set.

$$Y = 415.169 - 1.27241\text{"}B\text{"} - 1.59995\text{"}A\text{"} + 0\text{"}BC\text{"} + 0.224759\text{"}BD\text{"} - 54.4457\text{"}D\text{"}$$

$$r^2 = 1.000000 \qquad F = 14723406.373182 \qquad (7)$$

$$Y = 414.928 - 54.4258\text{"}D\text{"} - 1.59856\text{"}A\text{"} - 1.27146\text{"}B\text{"} + 0.224673\text{"}BD\text{"}$$

$$r^2 = 1.000000 \qquad F = 10882222.396607 \qquad (8)$$

These findings demonstrate that GFA analysis of experiments based on random independent variable settings results in the proper identification of the underlying equation of the data set.

The GFA algorithm allows one to examine all the possible combinations of variables. This feature is a modest achievement when only four variables are involved but it becomes a very significant achievement if a larger number of independent variables are being studied.

Application of GFAXD to 12-Variable Problem. The need to limit the number of variables studied in traditional experimental design to minimize the required number of experiments represents a significant compromise. For example, 4096 experiments are required to study 12 variables in a full factorial design. This number of experiments is prohibitive, particularly when the duration of a single experiment may be measured in hours or days. Fractional factorial designs would attenuate this large number of experiments, but even a 2^{12-5} design would still require 128 experiments.

Consider the theoretical study of 12 independent variables during the course of conducting 16 exploratory experiments. These 12 factors give rise to higher order interaction variables constructed as products of 2−12 factors. There are 66 two-factor interactions for 12 variables. Disregarding three-factor and higher order interactions for the sake of simplicity, the example under consideration presents a total of 78 independent variables to evaluate and only 16 experimental data points with which to do the evaluation.

We created a spreadsheet model of this hypothetical problem. The 12 single-factor independent variable ranges were normalized between 0.0 and 1.0 and were varied randomly across each entire variable range. The calculation of the two-factor interactions was direct and the yield for each experiment was calculated from underlying eq 9. The independent single-factor variables and the dependent variable are summarized in Table 6.

$$Y = 10 + 10\text{"}J\text{"} + 20\text{"}GJ\text{"} + 30\text{"}G\text{"} + 40\ \langle\text{"}D\text{"} - 0.25\rangle \qquad (9)$$

The task for the GFAXD method was to divine which of the 78 independent variables contributes to the experimental outcome, the yield, and to quantify the way in which the independent variables contribute to the yield via the underlying equation. Clearly this is a challenging problem that can only be solved by a highly efficient method.

The data in Table 6 were analyzed by GFA using the software default values. Linear and spline models were examined in a population of 100 equations that was allowed to evolve for 50 000 generations. Tables 7-9 summarize the first five equations from the populations of each of three duplicate analyses of the data. Each of the equations in

Table 8. Equation Set II of GFAXD Analysis of 12-Variable Problem

equation		r^2	F
8-1	$Y = 37.6418 + 10.1308\text{“}J\text{”} + 27.2947\langle\text{“}D\text{”} - 0.23\rangle - 30.0459\langle 0.92 - \text{“}G\text{”}\rangle + 13.007\langle\text{“}D\text{”} - 0.3\rangle + 19.6453\text{“}GJ\text{”}$	0.999982	111639.402005
8-2	$Y = 10.175 + 9.20109\text{“}J\text{”} + 29.2929\text{“}G\text{”} + 21.3923\text{“}GJ\text{”} + 38.835\langle\text{“}D\text{”} - 0.23\rangle$	0.999854	18888.717318
8-3	$Y = 12.8058 + 29.2979\langle\text{“}G\text{”} - 0.09\rangle + 21.5772\text{“}GJ\text{”} + 38.8802\langle\text{“}D\text{”} - 0.23\rangle + 9.04928\text{“}J\text{”}$	0.999871	21351.123321
8-4	$Y = 13.1636 + 29.6684\langle\text{“}G\text{”} - 0.11\rangle + 21.3411\text{“}GJ\text{”} + 39.1102\langle\text{“}D\text{”} - 0.23\rangle + 9.10197\text{“}J\text{”}$	0.999818	15133.364887
8-5	$Y = 9.7678 + 10.5718\text{“}J\text{”} + 31.1697\text{“}G\text{”} + 4.84409\langle\text{“}GJ\text{”} - 0.25\rangle + 39.101\langle\text{“}D\text{”} - 0.23\rangle + 14.9218\text{“}GJ\text{”}$	0.999906	21189.020048

Table 9. Equation Set III of GFAXD Analysis of 12-Variable Problem

equation		r^2	F
9-1	$Y = 9.99958 + 27.2947\langle\text{“}D\text{”} - 0.23\rangle + 13.007\langle\text{“}D\text{”} - 0.3\rangle + 10.1308\text{“}J\text{”} + 19.6453\text{“}GJ\text{”} + 30.0459\text{“}G\text{”}$	0.999982	111639.402005
9-2	$Y = 10.0659 + 33.167\langle\text{“}D\text{”} - 0.23\rangle + 9.79111\text{“}J\text{”} + 19.8798\text{“}GJ\text{”} + 30.058\text{“}G\text{”} + 7.42715\langle\text{“}D\text{”} - 0.37\rangle$	0.999976	83521.809824
9-3	$Y = 10.175 + 29.2929\text{“}G\text{”} + 21.3923\text{“}GJ\text{”} + 9.20109\text{“}J\text{”} + 38.835\langle\text{“}D\text{”} - 0.23\rangle$	0.999854	18888.717318
9-4	$Y = 10.4776 + 9.48115\text{“}J\text{”} + 20.6562\text{“}GJ\text{”} + 29.5642\text{“}G\text{”} - 2.94869\text{“}D\text{”} + 42.4051\langle\text{“}D\text{”} - 0.23\rangle$	0.999906	21241.472833
9-5	$Y = 10.0615 + 19.9728\text{“}GJ\text{”} + 30.1052\text{“}G\text{”} + 5.39348\langle\text{“}D\text{”} - 0.45\rangle + 9.62457\text{“}J\text{”} + 35.5191\langle\text{“}D\text{”} - 0.23\rangle$	0.999937	31726.242243

Tables 7−9 contain the variables *G*, *J*, and *GJ*, and each also contains a spline term involving variable *D*. The appearance of these common terms suggests their likely importance in the underlying equation. Almost all the equations in these tables appear to be variations on a single theme. The statistical metrics for all equations are exceptional.

Each of the three populations represented in Tables 7−9 is constituted by different equations. When this occurs, it provides an excellent opportunity to differentiate the underlying equation from the competing equations. There is a single equation (its constant term is 10.175) that appears in each population and its repeated appearance in different populations is sufficient evidence to separate it from the competing equations and to associate it with the underlying equation. Clearly this repeating equation is not identical to eq 9. The reason for this anomaly is that the GFA algorithm uses experimental values for the knot, *a*, in the spline terms $\langle X - a\rangle$ of the model. Consequently, the emergence of eq 9 from the GFA analysis would be exact only if, by chance, factor *D* assumed a value of exactly 0.25 in one of the experiments. The closest value of *D* to 0.25 in the experimental data set is 0.23, and this is the best value to be used in the spline term.

The GFAXD derivation of a close approximation to the underlying equation of the 12-variable problem from a data set containing only 16 experiments is a very significant accomplishment. Recall that 4096 experiments are required to study 12 variables in a full factorial experimental design! However, we sought to derive additional efficiency from the use of the GFAXD method by analyzing portions of the data set.

Analysis of the first 12 experiments in the data set failed to provide an approximation to eq 9 using a 100-equation population and a 50 000 generation evolution. The analysis was repeated using 100 000 generations, and the first equation of the resulting population, as shown in eq 10, was found to be a reasonable approximation of eq 9. This equation is the best that can be derived from this data because 0.3 is the value of *D* closest to the value of 0.25 required for the proper spline knot in eq 9.

$$Y = 9.14126 + 15.2885\text{“}GJ\text{”} + 32.7447\text{“}G\text{”} + 13.0186\text{“}J\text{”} + 42.5718\langle\text{“}D\text{”} - 0.3\rangle$$

$$r^2 = 0.999293 \qquad F = 2475.044270 \qquad (10)$$

When the last 12 experiments in the data set were evaluated with GFAXD using 50 000 generations, the first equation of the population was eq 11 which is slightly more accurate than eq 10 because of the availability of the value of 0.23 for use as the spline knot.

$$Y = 10.2164 + 29.2992\text{“}G\text{”} + 22.1124\text{“}GJ\text{”} + 8.69068\text{“}J\text{”} + 38.573\langle\text{“}D\text{”} - 0.23\rangle$$

$$r^2 = 0.999875 \qquad F = 13970.813300 \qquad (11)$$

We were unable to find eq 9 using the first 10 experiments of the data set even when the evolution was allowed to proceed 100 000 generations. However, the last 10 experiments provided eq 12 as the first member of the population when the analysis was done using 50 000 generations.

$$Y = 10.1946 + 38.5686\langle\text{“}D - 0.23\rangle + 21.9149\text{“}GJ\text{”} + 29.3144\text{“}G\text{”} + 8.87141\text{“}J\text{”}$$

$$r^2 = 0.999848 \qquad F = 8238.038731 \qquad (12)$$

Analysis of the last nine experiments of the data set failed to reveal an approximation to eq 9 in either 50 000 or 100 000 generations. This appeared to be the limit of the method while conducting the analysis using the GFA default parameters.

These results show that the GFAXD method is capable of finding a very good approximation to the underlying equation of the 12-variable problem using only 10 experimental data points. This remarkable achievement was dependent on which 10 experiments were used in the evaluation because of the need for a data point close to the value of the spline knot in the underlying equation.

CONCLUSIONS

Experimental design methods are used productively in industrial research and development for developing models of processes. There are several limitations of the experimental design method. First, the number of required experiments grows exponentially as the number of variables under study increases. Also, continuous variables are assumed to operate over the entire variable range and this may not always be a valid assumption. Finally, it is difficult to integrate results from experimental design studies with results from exploratory experiments.

Roger's GFA algorithm is a computational technique that can efficiently extract information from any set of experimental data. GFA utilizes a genetic algorithm to develop and evolve a population of high quality least-squares-regression-based equations that relate independent and dependent variables. GFA does not require continuous variables to operate over the entire variable range. The GFA algorithm has been applied successfully to experimental design data to extract process models.

GFAXD, based on the GFA algorithm, is a novel method for the design, conduct, and analysis of groups of experiments with the objective of identifying process models. The GFAXD method proposes that the maximum amount of information can be derived from a set of experiments when they are designed such that the independent variables, both controlled and noncontrolled, are allowed to assume values randomly distributed over the entire variable range and that the data from such experiments be analyzed using the GFA algorithm. The experimentalist uses scientific reason and validation experiments to select the best process models from the GFA population of model equations.

The use of GFAXD allows the experimentalist to study many more factors than would be practical using traditional experimental design. The method also provides for the acquisition of information about partitioning of the variable space and it accommodates the use of incidental experimental variable settings. GFAXD leverages the instincts, skills, and experience of the experimentalist by providing a means for extracting more information from exploratory experiments. GFAXD offers the potential to increase the productivity of process development experimentation by significantly increasing the amount of information that can be derived from each set of experiments. The present study was intended to demonstrate the utility of the GFAXD method. Additional studies of this novel experimental design method are needed. The results of a study of the parameterization of the GFAXD method will be presented in a future report.

ACKNOWLEDGMENT

The author thanks Dr. David Rogers, Dr. Thompson Doman, Mr. Robert Dillard, and Mr. Joseph Wieczorek for their helpful discussions during the course of this study and during the preparation of this manuscript.

REFERENCES AND NOTES

(1) Box, George E. P.; Hunter, William G.; Hunter, J. Stuart *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*; John Wiley & Sons: New York, 1978; pp 324−334.

(2) Hockman, K. K.; Berengut, D. Design of Experiments. *Chem. Eng. (N. Y.)* **1995**, *102(11)*, 142−143, 145, 147.

(3) Hunter, J. S. Design of Experiments. In *Encyclopedia of Chemical Process and Design*; McKetta, John J.; Cuningham, William A., Eds.; Marcel Dekker: New York, 1982; Chapter 15 and references cited therein.

(4) Sutton, N. Variables Search: A Simple Technique for Spotting the Key Parameters. *Chem. Eng. (N. Y.)* **1997**, *104(8)*, 106−109.

(5) Design-Ease®; Stat-Ease, Inc.: Minneapolis, MN.

(6) DOEpack; PQ Systems, Inc.: Dayton, OH.

(7) MODDE; UMETRI AB: Umeå, Sweden.

(8) ECHIP; ECHIP, Inc.: Hockessin, DE.

(9) DOE-PC IV; Quality America, Inc.: Tucson, AZ.

(10) Factorial-Design; Statistical Programs: Houston, TX.

(11) Rogers, D. Development of the Genetic Function Approximation Algorithm. In *Proceedings of the 6th International Conference on Geneic Algorithms*; Eshelman, L. J., Ed.; Pittsburg, PA, Morgan-Kaufmann: San Mateo, 1995.

(12) Cerius^2 QSAR+; Molecular Simulations, Inc.: San Diego, CA.

(13) Mehta, R. K.; Dieudonne, V.; Yoon, R. H. Optimization of a Chemical Coal Cleaning Process via Simple Genetic Algorithm. In *Hydrometall. Proc. Milton E. Wadsworth Int. Symp., 4th*; Hiskey, J. Brent; Warren, Garry W.; Eds.; Soc. Min., Metall. Explor.: Littleton, CO, 1993; p 629−43.

(14) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan: Ann Arbor, MI, 1975.

(15) Holland, John H. Genetic Algorithms. *Sci. Am.* **1992**, *267(1)*, 66−72.

(16) Reeves, Colin R.; Wright, Christine, C.; Genetic Algorithms versus Experimental Methods: A Case Study. In *Proceedings of the Seventh International Conference on Genetic Algorithms*; Bäck, Thomas, Ed.; East Lansing, MI, Morgan-Kaufmann: San Francisco, 1997; pp 214−220.

(17) Deb, Kalyanmoy; Saxena, Vikas Car Suspension Design for Comfort Using Genetic Algorithms. In *Proceedings of the Seventh International Conference on Genetic Algorithms*; Bäck, Thomas, Ed.; East Lansing, MI, Morgan-Kaufmann: San Francisco, 1997; pp 553−560.

(18) Cunha, A. Gaspar; Oliveira, Pedro; Covas, Jose. Use of Genetic Algorithms in Multicriteria Optimization to Solve Industrial Problems. In *Proceedings of the Seventh International Conference on Genetic Algorithms*; Bäck, Thomas, Ed.; East Lansing, MI, Morgan-Kaufmann: San Francisco, 1997; pp 682−688.

(19) Upreti, Simant R.; Deb, Kalyanmoy. Optimal Design of an Ammonia Synthesis Reactor Using Genetic Algorithms. *Comput. Chem. Eng.* **1996**, *21(1)*, 87−92.

(20) Goldberg, David E. *Genetic Algorithms in Search, Optimization & Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(21) Rogers, David; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(22) Friedman, J. Multivariate Adaptive Regression Splines. In *Technical Report No. 102*, Laboratory for Computational Statistics, Department of Statistics, Stanford University: Stanford, CA, Nov 1988 (revised Aug 1990).

(23) Rogers, David. Evolutionary Statistics: Using a Genetic Algorithm and Model Reduction to Isolate Alternate Statistical Hypotheses of Experimental Data. In *Proceedings of the Seventh International Conference on Genetic Algorithms*; Bäck, Thomas, Ed.; East Lansing, MI, Morgan-Kaufmann: San Francisco, 1997; pp 553−560.

CI980205F