Efficient Discovery of Nonlinear Dependencies in a Combinatorial Catalyst Data Set

James N. Cawse*

GE Global Research, 1 Research Circle, Niskayuna, New York 12309

Manfred Baerns and Martin Holena

ACA-Berlin, Institut für Angewandte Chemie Berlin-Adlershof e.V.,
Postfach 96 11 56, 12474 Berlin, Germany

Received August 5, 2003

Exploration of a complex catalyst system using Genetic Algorithm methods and combinatorial experimentation efficiently removes noncontributing elements and generates data that can be used to model the remaining system. In particular the combined methods effectively navigate and optimize systems with highly nonlinear dependencies (3-way and higher interactions).

INTRODUCTION

Combinatorial experimentation in the materials development arena can be viewed as a search of a relatively high dimensional space for synergistic interactions of the components and processing variables. In previous work we have noted "in many of the targets studied, the main effects and frequently the two-way interactions of the major process variables have been thoroughly studied. New and substantially enhanced properties will be found in high-order interactions, such as... synergistic interactions in a catalyst formulation".[1] In this note we would like to report the observation of such a synergy, discovered by the use of combinatorial experimentation using Genetic Algorithm (GA) methodology. We further show that the GA efficiently led to a parsimonious system, which could be easily modeled using standard regression methodology.

METHODOLOGY

The system studied was a propane-propene oxidation catalyst that has already been reported by Wolf et al.,[2] with additional points added based on the observations of the trends in the GA. Statistical analysis of the data was performed using Minitab[3] and Design-Expert.[4]

Wolf et al. used genetic algorithm techniques to explore an eight-component catalyst system consisting of mixtures of B, Fe, Ga, La, Mg, Mn, Mo, and V. Using it they located a region of improved product yield. In the process, the genetic algorithm's semirandom search of the experimental space has given us a valuable set of data to explore using more conventional statistical techniques. The data set contains 328 samples. Each sample consists of a mixture of some of the elements and a resultant propene yield.

An experimental set of this size can be expected to contain data that is not useful, along with the useful data. In a statistical modeling effort, it is quite possible to have *too much* data. In particular, data that are taken far from the area of interest will tend to add noise and confusion without increasing knowledge. Therefore we will engage in judicious subsetting of the data to extract the most useful information.

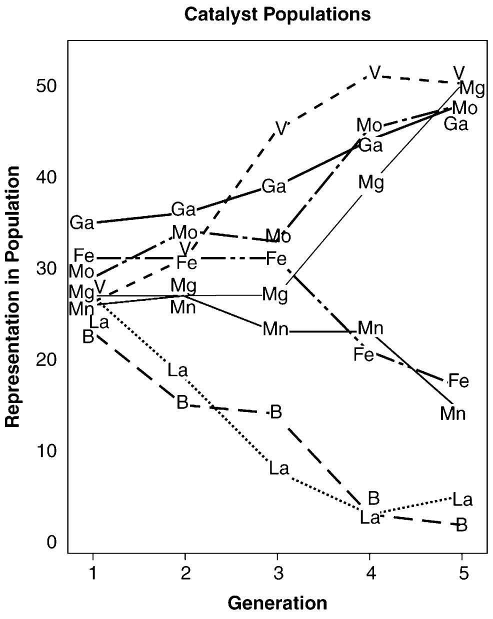

Figure 1. Representation of the eight candidate elements in each generation of the genetic algorithm.

The genetic algorithm[5] starts with an initial population of samples. After determination of the outcome of each sample, the best samples are selected as "parents", and one or more genetic operators such as mutation and recombination are applied. The resulting "children" become the next population for the experiment. This process is repeated until either the goal is reached or no more improvement is seen. In this experiment, the first generation was a set of 56 samples. For each catalyst sample of the first generation, four elements were randomly selected from the 8 candidate elements, and their concentrations set randomly. In the subsequent samples, the number of elements in each sample varied as the selection process attempted to find the best propene yield. After five generations, an additional set of 48 catalysts was selected

* Corresponding author phone: (518)387-6095; fax: (518)387-7611; e-mail: cawse@crd.ge.com.

10.1021/ci034171+ CCC: $27.50
Published on Web 11/22/2003

Table 1. Model Terms in Standard Mixture Designs[3]

this model type	fits these terms	and models this type of blending
linear (first-order)	linear	additive
quadratic (second-order)	linear and quadratic	additive nonlinear synergistic binary or additive nonlinear antagonistic binary
special cubic (third-order)	linear, quadratic and special cubic	additive nonlinear synergistic ternary nonlinear antagonistic ternary
full cubic (third-order)	linear, quadratic, special cubic, and full cubic	additive nonlinear synergistic binary nonlinear antagonistic binary nonlinear synergistic ternary nonlinear antagonistic ternary

Table 2. Terms Provided in Standard Software for a Multicomponent Mixture Model[a]

this model	fits these terms
linear	X1, X2, X3, X4, X5 (note that there is no constant term)
quadratic	linear terms plus X1*X2, X1*X3, X1*X4, X1*X5, X2*X3, X2*X4, X2*X5, X3*X4, X3*X5, X4*X5
special cubic	quadratic terms plus X1*X2*X3, X1*X2*X4, X1*X2*X5, X1*X3*X4, X1*X3*X5, X1*X4*X5, X2*X3*X4, X2*X3*X5, X2*X4*X5, X3*X4*X5
full cubic	special cubic terms plus X1*X2*(X1-X2), X1*X3*(X1−X3), X1*X4*(X1−X4), X1*X5*(X1−X5), X2*X3*(X2−X3), X2*X4*(X2−X4), X2*X5*(X2−X5), X3*X4*(X3−X4), X3*X5*(X3−X5), X4*X5*(X4−X5)

[a] A 5-component model is given here as an example.[3]

Table 3. ANOVA for the Five Generations of the GA Process (280 Samples)[a]

source	DF	seq SS	adj SS	adj MS	F-ratio	p-value
regression	119	0.14704	0.14704	0.00124	11.80	0.000
linear	7	0.08982	0.00062	0.00009	0.85	0.546
quadratic	28	0.03777	0.00688	0.00025	2.35	0.000
special cubic	56	0.01352	0.00881	0.00016	1.50	0.026
full cubic	28	0.00593	0.00593	0.00021	2.02	0.004
residual error	160	0.01675	0.01675	0.00011		
lack-of-fit	141	0.01490	0.01490	0.00011	1.09	0.442
pure error	19	0.00185	0.00185	0.00010		
total	279	0.16379				

[a] S = 0.01, R-Sq = 90%, R-Sq(adj) = 82%.

based on the experimenter's conclusions from the data at hand.

An examination of the catalyst populations for each generation is immediately instructive (Figure 1). Two elements (B and La) are immediately selected against and rapidly disappear from the population. Two others (Fe and Mn) follow more slowly but are steadily decreasing by the fourth generation. It is still not clear whether they might act as low-concentration promoters. The other four elements (Ga, Mg, V, and Mo) are strongly represented throughout the five generations. This information is valuable as we begin statistical modeling of the data.

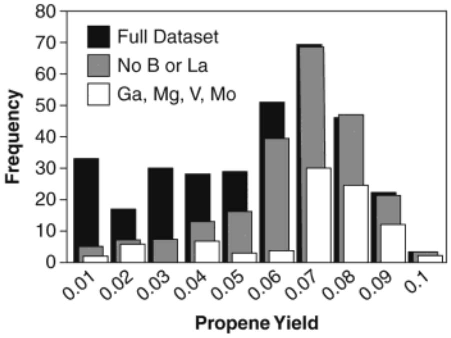

Figure 2. Histogram of the propene yield for the full data set and its subsets.

Table 4. ANOVA of 177-Point Data Set with No B or La[a]

source	DF	seq SS	adj SS	adj MS	F-ratio	p-value
regression	40	0.03911	0.03911	0.00098	9.75	0.00
linear	5	0.01080	0.00297	0.00059	5.92	0.00
quadratic	15	0.02039	0.00823	0.00055	5.47	0.00
special cubic	20	0.00792	0.00792	0.00040	3.95	0.00
residual error	136	0.01364	0.01364	0.00010		
lack-of-fit	128	0.01287	0.01287	0.00010	1.05	0.53
pure error	8	0.00077	0.00077	0.00010		
total	176	0.052745				

[a] S = 0.010, R-Sq = 74%, R-Sq(adj) = 66%.

Since the eight elements of the catalyst formulation form a true mixture, with the proportions of the elements in each sample adding up to 1.0, these data must be analyzed with mixture design tools.[6] Mathematical models for mixture experiments are different from conventional regression models because of the constraint of the components adding to a constant. In essence, the eight components are not all independent; if seven are specified, the eighth can be calculated. This makes the following changes in the model equations:

The regression equation for a mixture design is usually formulated as a Sheffé polynomial that has no constant term. This can be seen to be obvious if one considers a typical regression equation

$$Y = C_0 + C_1x_1 + C_2x_2 + ... C_nx_n$$

In a normal situation, $Y = C_0$ if all the x's are zero. However, in a mixture experiment, all the x's cannot be zero so $Y = C_0$ is not meaningful.

The Sheffé polynomial for a mixture design does not have any squared terms. Curvature is represented by interaction terms.

The classes of mixture designs that are analyzed by standard statistical software are given in Tables 1 and 2.

RESULTS OF THE STATISTICAL ANALYSIS

Initial regression of the whole data set shows that there are very significant effects (Table 3), given the very favorable p-values for the regression as a whole and the quadratic and special cubic models in particular.

Following the lead given by the genetic algorithm, we remove all samples containing either B or La; the number of samples in the actual GA decreases to 177. A histogram (Figure 2) of the data shows that a very large fraction of the low yield samples has been removed, but little information has been lost about the high yield samples.

In this data set the statistical analysis showed that the full cubic model was no longer significant. Removal of those

Table 5. Significant Terms (p<0.100) of 177 Sample Model (No B or La), Sorted by p-Value

source	sum of squares	DF	coefficient estimate	mean square	F-value	prob > F
model	0.047832	40		0.001196	9.763	<0.0001
linear mixture	0.01584	5		0.003168	25.87	<0.0001
Mg*V	0.001654	1	0.324	0.001654	13.5	0.0003
Mg*Mn*V	0.001286	1	−1.61	0.001286	10.5	0.0015
Mg*Mn	0.000833	1	−1.2	0.000833	6.802	0.0101
Mn*V	0.00067	1	0.498	0.00067	5.467	0.0208
Fe*Ga*V	0.00064	1	1.398	0.00064	5.223	0.0238
Ga*Mo*V	0.000637	1	1.722	0.000637	5.197	0.0242
Mg*Mo*V	0.000633	1	1.104	0.000633	5.168	0.0246
Fe*Mo*V	0.000423	1	1.511	0.000423	3.454	0.0652
Fe*Mn*V	0.000395	1	1.083	0.000395	3.228	0.0746
Mg*Mn*Mo	0.000374	1	1.083	0.000374	3.052	0.0829

Table 6. ANOVA for the Ga/Mg/V/Mo Model (65 Samples)[a]

source	DF	seq SS	adj SS	adj MS	F-ratio	p-value
regression	13	0.019677	0.019677	0.001514	13.42	0.000
linear	3	0.007436	0.002791	0.00093	8.25	0.000
quadratic	6	0.011462	0.002834	0.000472	4.19	0.002
special cubic	4	0.000779	0.000779	0.000195	1.73	0.158
residual error	51	0.005753	0.005753	0.000113		
lack-of-fit	48	0.005134	0.005134	0.000107	0.52	0.863
pure error	3	0.000619	0.000619	0.000206		
total	64	0.025431				

[a] S = 0.01062, R−Sq = 77.4%, R−Sq(adj) = 71.6%.

Table 7. Terms for the G/Mg/V/Mo Model, Sorted by p-Value[a]

term	coeff	StDev	T	P
Ga	0.0906	0.02457	3.688	0.001
Mg*V	0.2706	0.08547	3.166	0.003
Mg*Mo*V	1.1375	0.62636	1.816	0.075
Ga*Mo*V	2.0965	1.3577	1.544	0.129
Ga*Mo	−0.5524	0.45982	−1.201	0.235
Mo	0.2571	0.25049	1.026	0.31
Mo*V	−0.5129	0.59338	−0.864	0.391
Ga*Mg*V	0.3598	0.43568	0.826	0.413
Mg*Mo	−0.2761	0.33758	−0.818	0.417
Mg	−0.0066	0.0144	−0.456	0.65
Ga*V	0.0328	0.10509	0.312	0.756
Ga*Mg	−0.028	0.09916	−0.282	0.779
V	−0.0043	0.02125	−0.202	0.84
Ga*Mg*Mo	0.0923	0.5707	0.162	0.872

[a] The most significant terms are given in boldface.

Table 8. ANOVA for the Ga/Mg/V/Mo Model (95 Samples)[a]

source	DF	seq SS	adj SS	adj MS	F-ratio	p-value
regression	13	0.02824	0.02824	0.002173	16.21	0.000
linear	3	0.00845	0.00700	0.002334	17.41	0.000
quadratic	6	0.01845	0.00424	0.000707	5.28	0.000
special cubic	4	0.00134	0.00134	0.000335	2.5	0.049
residual error	81	0.01086	0.01086	0.000134		
lack-of-fit	74	0.01010	0.01010	0.000137	1.26	0.405
pure error	7	0.00075	0.00075	0.000108		
total	94	0.03910				

[a] S = 0.11, R-squared =72%, adj R-squared = 67%.

Table 9. Terms for the G/Mg/V/Mo Model Sorted by p-Value (95 Samples)[a]

source	sum of squares	DF	mean square	F-value	prob > F
model	0.028245	13	0.0021727	16.205	<0.0001
linear mixture	0.023781	3	0.0079271	59.125	<0.0001
Mg*V	0.002500	1	0.0025005	18.650	<0.0001
Ga*Mo*V	0.000759	1	0.000759	5.661	0.0197
Mg*Mo*V	0.000712	1	0.0007119	5.310	0.0238
Ga*Mo	0.000146	1	0.0001461	1.090	0.2996
Ga*V	0.000124	1	0.0001239	0.924	0.3392
Mo*V	8.010E−05	1	8.01E−05	0.597	0.4418
Ga*Mg	7.524E−05	1	7.524E−05	0.561	0.4559
Ga*Mg*Mo	4.027E−05	1	4.027E−05	0.300	0.5852
Ga*Mg*V	2.448E−05	1	2.448E−05	0.183	0.6703
Mg*Mo	2.273E−06	1	2.273E−06	0.017	0.8967
residual	0.010860	81	0.0001341		
lack of fit	0.010101	74	0.0001365	1.259	0.4055
pure error	0.000759	7	0.0001084		
cor total	0.039105	94			

[a]The most significant terms are given in boldface.

model terms gave a simpler model in which all of the levels of the model were extremely significant (Table 4).

Examination of the detailed terms of this model (Table 5) reveals several interesting things: (1) After the linear term in Ga, seven of the eight next significant terms are interactions with V. (2) Five of the eleven most significant terms (p<0.100) are 3-way interactions. (3) Mn is present in four of the most significant terms even though Figure 1 shows that it is strongly deselected by the fourth generation. This makes sense given that the coefficients for the two most significant of those terms are negative, suggesting that it might act as a catalyst poison.

If we further remove the samples with Fe and Mn components, the analysis gives a weak indication of a special cubic design (Table 6) with at least one nearly significant 3-way interaction (Table 7).

Fortunately for this modeling effort, the experimenter chose to add 48 more samples, based on the observations of the trends in the data. Most of those samples were in the Ga/Mg/V plane, but a few were also in the Ga/Mg/V/Mo space. If we add those samples to the above data, we get Table 8, in which the special cubic model is now fully significant. The detailed terms (Table 9) now show two

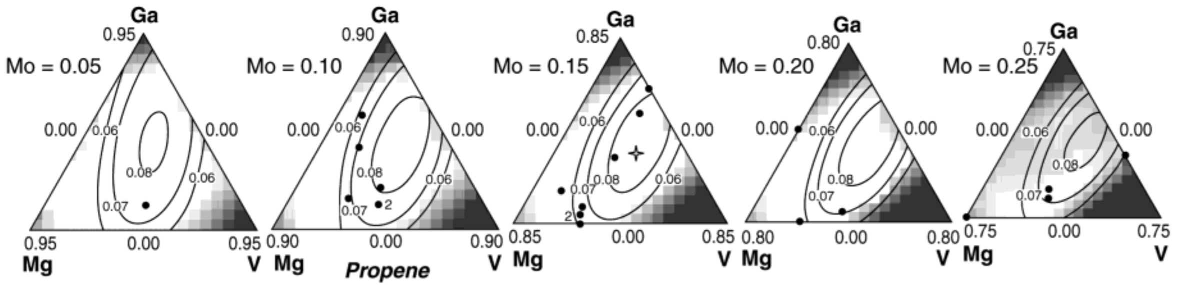

Figure 3. Contour maps of the Ga/Mg/V/Mo system for Mo = 0.05 to 0.25. Contour lines show predicted propene yield. Dots show experimental points; cross shows approximate location of highest predicted yield or 8.7%. Darkened areas indicate regions in which the model predicts poorly because of lack of data.

significant 3-factor interactions.

Optimization of this model gives the highest propene yields when Mo is approximately 0.15. A four-component mixture of this type yields a tetrahedron; a set of slices across that tetrahedron gives contour maps showing the region of highest yield (Figure 3). This region, bordered by the yield = 0.08 contour line, is an ellipsoid suspended in the middle of the tetrahedral experimental region between Mo = 0.05 and Mo = 0.25. The maximum predicted propene yield is 8.7% at$Ga_{0.37}Mg_{0.16}Mo_{0.16}V_{0.32}$. The standard error of prediction is 1.2% so the 95% prediction interval is [6.2%, 11.0%]. The predicted maximum is essentially identical to the maximum experimental points; there are five measured points larger than 8.7%, with the largest 9.1%. The RMS error of the experimental system, from the data available, is 1.0%.

Although the system clearly exhibits 4-component synergy (nonlinear blending behavior), the model has no 4-way interaction terms. The two 3-way interactions are sufficient for the prediction equation to model the 4-component ellipsoid.

CONCLUSION

These results show that a GA is a very effective tool for exploring a complex system when a high throughput experimental process is available to generate the required data. It efficiently removes noncontributing factors and factors with negative impacts on the outcome and generates data that can be used to model the remaining system in terms of the high value components.

The model of this system has confirmed our claim that the optima of new catalyst systems will be the result of high order synergies and that combinatorial methodology is an appropriate tool for locating those interactions and optima.

ACKNOWLEDGMENT

Thanks to Prof. John Cornell (University Florida) and Pat Whitcomb (Stat-Ease) for valuable discussions on interpretation of the mixture designs.

REFERENCES AND NOTES

(1) *Experimental Design for Combinatorial and High Throughput Materials Development*; Cawse, J. N., Ed.; John Wiley and Sons: New York, 2002; p 16.

(2) Wolf, D.; Buyevskaya, O. V.; Baerns, M. An evolutionary approach in the combinatorial selection and optimization of catalytic materials. *Appl. Catal. A* **2000**, *200*, 63−77.

(3) Minitab Inc.; 12 ed.: State College, PA, 1999.

(4) Stat-Ease; 6 ed.; Stat-Ease, Inc: Minneapolis, MN, 2000.

(5) Wolf, D.; Baerns, M. In *Experimental Design for Combinatorial and High Throughput Materials Development*; Cawse, J. N., Ed.; John Wiley and Sons: New York, 2002; pp 147−162.

(6) Cornell, J. A. *Experiments With Mixtures: Designs, Models, and the Analysis of Mixture Data*, 3rd ed.; John Wiley and Sons: New York, 2002.

CI034171+