Using Statistical Experimental Design to Optimize GC Operation

Douglas K. Ludlow
The University of North Dakota

Introduction

Statistical experimental design is useful to determine the optimum operating conditions of real processes and has applications for quality control and improvement. A laboratory assignment has been developed which uses a gas chromatography experiment to give quantitative results which the students use to apply statistical skills without being impeded by complex equipment or experimental methods. The experiment has been used for several years in the undergraduate chemical engineering laboratories at the University of North Dakota [1]. In a classroom setting typical "experimental results" can be analyzed without necessarily running the experiment. One of the unique aspects of the experiment is that there is a trade-off between the two most significant variables, forcing students to compromise in the selection of optimum conditions. Such compromises are typical in many real-world industrial situations.

Data Analysis Techniques

The problem presented assumes that the students have been introduced to statistical experimental strategies (or designs) and response surface analysis. Experimental designs such as the Central Composite design[2,3] (based on the 2" factorial design) or the Box-Behnken design[2-4] are used to collect data in a systematic way so that a mathematical model to the response surface can be determined. The Box-Behnken design has the advantage over the Central Composite Design in that fewer total experimental runs are needed and that the independent variables or factors are varied between three equally spaced values. This design collects enough data so that a mathematical model of the response surface can be determined.

A simple response surface model is the polynomial:

$$f = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 + b_{11}X_1^2 + b_{22}X_2^2 + b_{33}X_3^2 \quad [1]$$

Where $X_1, X_2,$ and X_3 are the independent variables (or factors) varied in the experiments and b_0, b_1, etc. are coefficients or effects determined by regression of the data. This simple polynomial model has been found to be quite adequate for the majority of practical problems. The use of an experimental design assures that enough data is collected to assess the statistical significance of each term in the model so that the "best" model of the data can be determined. By "best" we mean the simplest model that adequately describes the response surface. The final model (with only statistically significant effects) can then be used to draw whatever conclusions are warranted about the system under investigation. We encourage the use of spreadsheets for this analysis since the use of statistical packages often makes the analysis too "canned" and inhibits learning of all the steps involved.

Task Assignment

The students are asked to optimize the operating conditions of a GC so that the analysis of a two-component liquid mixture can be performed in the minimum amount of time with the desired level of accuracy (resolution). A problem that occasionally occurs in GC analyses is that of peak resolution. Due to similarities of the physical properties of some substances, their GC peaks will overlap. This often leads to inaccurate or unusable results.

Typical procedures to improve peak resolution include using a smaller sample size, operating the GC at a lower temperature, "or using a lower carrier gas flow rate. The last two procedures also lead to longer times for analysis since the sample will take a longer time to elute from the GC column.

-The output of the GC is processed with an integrator that gives the results of elution time, area under the peak (A), and the area-to-height ratio *(A/H)* for each compound eluting from the GC. The *A/His* essentially the width of the peak at half of its height and corresponds to the time that the bulk of the component is actually eluting from the GC. (If you assume that the peak is approximated by a triangle, then the area is $(B \times H)/2$, and *A/His* the half width of the peak). One method to quantify the resolution of peaks is by calculating the following:

$$Y_1 = (\text{Time between peaks})/(A/H_{\text{peak1}} + A/H_{\text{peak2}}) \quad [2]$$

Higher values of Y, indicate better resolution of the peaks. To ensure good resolution with no distortion of the peaks due to overlap, Y, should have a value of at least 1.5. Optimal operating conditions will also include the shortest operation time which gives an adequate value of Y,. The second function that the students optimize (minimize) is the time for the last peak to pass through the GC, or

$$Y_2 = (\text{Time for second peak}) \quad [3]$$

The students are told to use a Box-Behnken statistical design which gives all the information to fit the response surface of three variables (factors) using only 15 experiments (which includes three replicates at the average conditions) The three operating variables are the sample size, the GC oven temperature, and the carrier gas flow rate. Each variable is set at three equally spaced levels. For the regression analysis the variables are coded using

$$X_i = (\text{Factor Value}_i - \text{Center})(\text{High Value-Center}) \quad [4]$$

so that the largest value of a variable is one and the lowest is negative one. There are several statistical reasons why the variables should be coded, the major one being to minimize interdependency of the coefficients in the quadratic equation. It also puts all factors on the same scale, so that the most important coefficient has the largest absolute value.

Table 1 gives some actual experimental results for a Box-Behnken design. The variables are listed in coded form and the two responses have been calculated from the A/H ratio and elution times for each experiment. From the experimental data the optimum operating conditions (minimize Y_2 for conditions where Y, is at least 1.5) can be determined.

Solution

Typically the students analysis the data in Table 1 using a spreadsheet since they are familiar with it. Multiple regression analyses are completed for both Y, and Y_2 to fit the general quadratic model (Equation 1) with the coded independent variables . The multiple regression package of the spreadsheet gives all of the coefficients of the quadratic model with their corresponding standard errors.

Next the best response surface model is found by eliminating arty nonsignificant terms. First, the t-statistic for each coefficient is calculated by dividing the coefficient (determined from multiple regression) by its standard deviation. If the t-statistic is less than the critical value from tables for 95% confidence (approximately 2), then the term is not significantly different from zero and can be dropped. The left-handed columns of Table 2 give the regression analysis for the Y, values given in Table 1. As shown, for this set of data, the coefficient b_{23} for the interaction effect between variables X_2 and X_3 is not shown to be significantly different than zero (its t-statistic is less than two). The data are then regressed again with a model that eliminates the X_2X_3 interaction, and the results are given in the right-hand columns of Table 2. In this case, all of the remaining coefficients are shown to be statistically significant, and the response surface model is completed.

Finally, the response surface model is then plotted as contour plots using a graphical software package. In this case, since X_1 (sample size) had the least effect (as noted by the smallest coefficients), plots are made of the responses versus X_2 and X_3. Several plots of the response surface can be generated for both Y_1 and Y_2. Figures 1 and 2 show two such plots generated for the data in Table 1. By overlaying the Y, and Y_2 contour plots (Figure 3),

the point can be found where Y_2 is minimized subject to the constraint that Y, is 1.5 or greater. It should be noticed from the contour plots that this experiment demonstrates vividly the compromises that must often be made in the real world between various objectives. In this case, a compromise must be made between speed and accuracy (resolution). Resolutions of more than 3 could be obtained, but the price is doubling or tripling of the analysis time.

Conclusion

This experiment, which was developed to reinforce the concepts of experimental design and statistical analysis, has been successfully used in a laboratory situation. The experimental results can also be analyzed in a classroom setting using the data. The multiple regression analysis and the determination of the best response surface model by the elimination of all nonsignificant coefficients can be completed entirely using a statistical software package. Another reason for its successful use is that the effects of the variables are not easily predicted which is often the case in real-world industrial optimization problems. The results are interesting since they predict a curvature in the response surface and the optimal operation conditions are not on one of the boundaries or limits of operating conditions. The task of finding conditions that are a good compromise for two responses is a very realistic situation.

References

1) Ludlow, D. K., Schulz, K. H., Erjavec, J. "Teaching Statistical Experimental Design Using a Laboratory Experiment" *Journal of Engineerirzg Education 84(4)351-359(1995).*

2) Lawson, J. and J. Erjavec, *Basic Experimental Strategies and Data Analysis,* BYU Press, Provo, Utah, 1992.

3) Box, G. E. P., W.G. Hunter, and J.S. Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building,* John Wiley & Sons, New York, 1978.

4) Box, G. E. P., Behnken, D. W., "Some New Three Level Designs for the Study of Quantitative Variables, " *Technometrics 2* 455-475 (1960).

Douglas K. Ludlow

Douglas K. Ludlow is an Associate Professor and Chair of the Department of Chemical Engineering at the University of North Dakota. Dr. Ludlow is active in ASEE, and is currently serving as the Program Chair of the NEE Division. He is also active in AIChE, where he is currently the Student Chapter Advisor at UND and Chair of the National Student Paper Competition held each year at the AIChE National Meeting. Address: Department of Chemical Engineering, University of North Dakota, Box 7101, Grand Forks, ND 58202-7101.
Phone: 701-777-3091, Fax: 701-777-4838, E-mail: Doug_Ludlow@mail.und.nodak.edu

Box-Behnken Experimental Design for Three Factors with Typical Experimental Responses.

TABLE 1

Run	Coded Sample Size	Coded Gas Flow	Coded Temp	Resolution Y_1	Run Time Y_2
1	-1	-1	0	2.81	5.95
2	+1	-1	0	1.91	5.53
3	-1	+1	0	1.49	1.98
4	+1	+1	0	1.25	1.89
5	-1	0	-1	3.11	5.82
6	+1	0	-1	2.30	5.34
7	-1	0	+1	1.04	1.63
8	+1	0	+1	0.865	1.51
9	0	-1	-1	3.09	11.6
10	0	+1	-1	2.25	3.96
11	0	-1	+1	1.44	3.83
12	0	+1	+1	0.771	1.15
13	0	0	0	1.59	2.67
14	0	0	0	1.60	2.70
15	0	0	0	1.61	2.68

Typical Regression Analysis to Determine Response Surface Model (Analysis for Y_1, Resolution of Peaks)

TABLE 2

First Regression	
Constant	1.596
Std Error of Y Est.	0.10016
R Squared	0.99375
No. of Observations	15
Degrees of Freedom	5

Coefficients			
Coef.	Value	Std. Err.	t
b_1	-0.2644	0.03541	7.466
b_2	-0.4372	0.03541	12.35
b_3	-0.8296	0.03541	23.43
b_{12}	0.1618	0.05008	3.230
b_{13}	0.1580	0.05008	3.155
b_{23}	0.04325	0.05008	0.8636*
b_{11}	0.1040	0.05213	1.995
b_{22}	0.1628	0.05213	3.122
b_{33}	0.1265	0.05213	2.427

Final Regression	
Constant	1.598
Std Error of Y Est.	0.09802
R Squared	0.99282
No. of Observations	15
Degrees of Freedom	6

Coefficients			
Coef.	Value	Std. Err.	t
b_1	-0.2644	0.03465	7.629
b_2	-0.4373	0.03465	12.62
b_3	-0.8296	0.03465	23.94
b_{12}	0.1618	0.04901	3.300
b_{13}	0.1580	0.04901	3.224
b_{23}			
b_{11}	0.1040	0.05101	2.039
b_{22}	0.1628	0.05101	3.191
b_{33}	0.1265	0.05101	2.480

* Small t-value indicates that this coefficient is not significantly different from zero and the interaction term is eliminated from the response surface model.

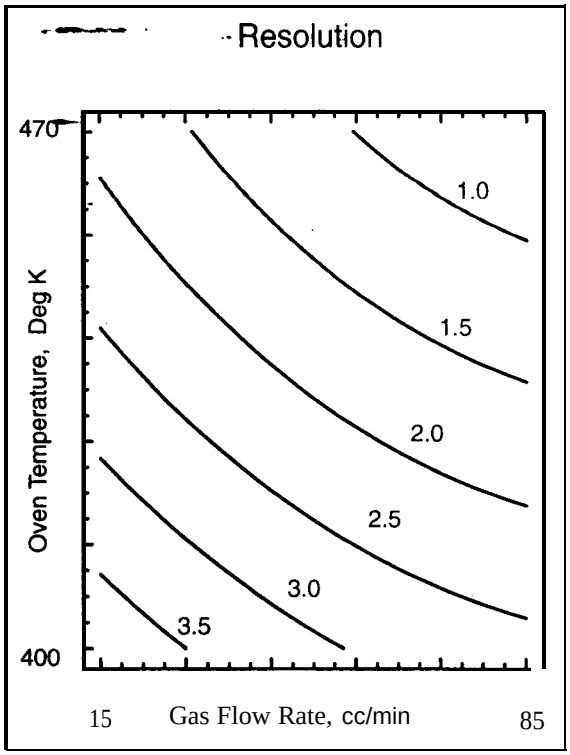

Figure 1. Contour plot of Y_1 (Resolution of Peaks) as a function of X_2 (Oven Temperature) and X_3 (Gas Flow Rate) at the average value of X_1 =1 (Sample Size). Plot shows regions on response surface map where acceptable resolution ($Y_1 \geq 1.5$) are obtainable.

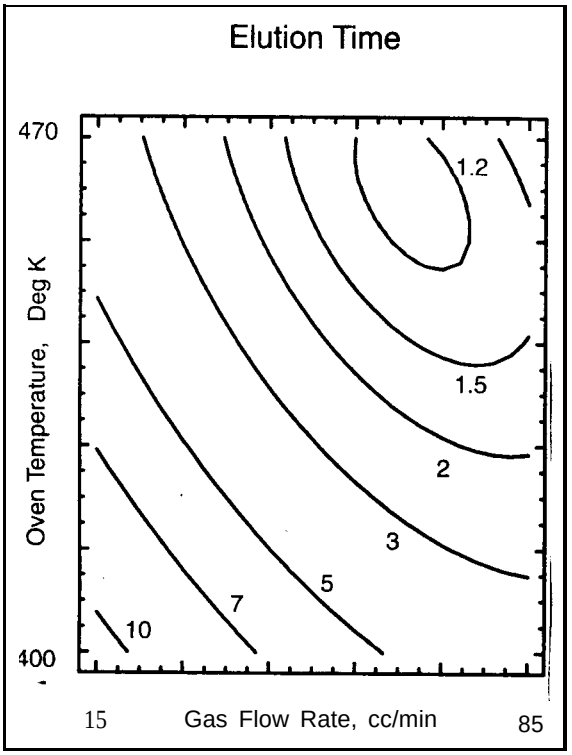

Figure 2. Contour plot of Y_2 (Analysis Time) as a function of X_2 (Oven Temperature) and X_3 (Gas Flow Rate) at the average value of X_1 =1 (Sample Size). Plot shows operating conditions that minimizes the total analysis time.

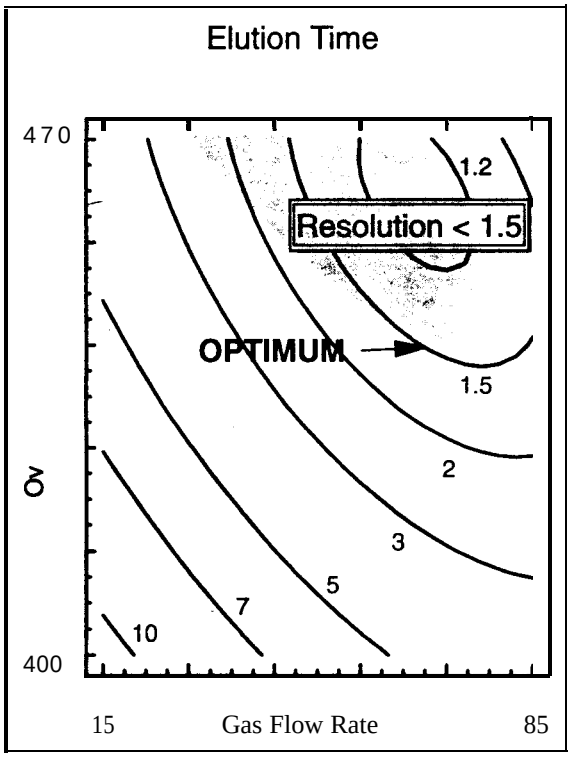

Figure 3. Contour plot of Y_2 (Analysis Time) as a function of X_2 (Oven Temperature) and X_3 (Gas Flow Rate) at the average value of X, =1 (Sample Size) with the resolution constraint (Y,) overlaid. This plot shows the need to compromise needs to be made between accuracy (resolution) and speed of analysis.