concept of Box and Wilson

In 1951, Box and Wilson[2] introduced a new concept into experimental design by recognizing that, in many situations, one is not so much interested in testing the significance of factors associated with the process, as in simply determining the best operating conditions for the process. In other words, they recognized problems 1 and 2 above to be independent and that the solution of the latter might be obtained without first finding f. The class of experimental designs introduced has become known as *response surface designs* and the associated analyses yielding the optimum operational conditions is known as *response surface analysis.*

application

The behavior of any reaction is governed by laws which should be representable in mathematical form and thus it should be possible to determine the optimum conditions for the reaction by simply applying these laws. However, one often finds in practice that the underlying mechanism of the system is so complicated that the mathematical representation using theoretical considerations is essentially impossible. Or, if the mathematical representation is achieved, the resulting mathematical system is so complex that its optimization is impractical. It is under these conditions that the empirical approach is necessary.

When one is faced with need to use the empirical approach and is interested in determining the best operating conditions, say, for a chemical or industrial process, these response surface designs are appropriate. Thus we find the approach useful in petroleum engineering where, for example, one may be interested in minimizing the viscosity of a lube additive taking into account the several factors involved in the process. In the study of chemical reactions it is often important to maximize the yield of a product of a reaction for a given amount of one of the initial materials by changing the time, concentrations, amounts, temperatures, etc., involved in the reaction process. In fact, the whole field of process control centers about the location of such optimum operating conditions.

Many research problems also have a strong optimum response orientation. Medical research provides a good example. Thus, in cancer therapy research, one may inquire as to what values, associated with a given diagnosis, of the radiation treatment variables such as duration, intensity, orientation, etc., would maximize the effectiveness of the treatment; or in preoperative anesthesia medication, what values in the medication program such as amount, concentration and frequency of administration of the drug would minimize the effect on the maintenance of anesthesia during the operation, or, in clinical immunology research, the amount, number, and spacing of injections required so as to maximize the resulting antibody count. It is apparent that the general problem is to find the quantitative level of each of the so-called independent variables which will give an optimum response for the process or dependent variables.

For the convenience of the reader who may wish to review terminology, the following glossary (of terms used) is inserted: terminology

Test of significance (*statistical significance*). A procedure for determining the probability of an observed experimental result arising under an assumed hypothesis.

Factor space. The set of all possible combinations of controllable conditions involved in a process.

Factorial experiment. An experimental design that involves the running of the experiment so that each level being considered for a factor is paired with all possible combinations of levels of all the other factors.

Variance (σ^2). A measure of dispersion or variation. The average squared deviation of the individual observations from the mean or average of all the observations. The square of the standard deviation.

Second order terms. The terms of second degree. The squared terms such as X_1^2, or X_1X_2.

Unbiased estimates. Estimates of population characteristics (parameters) made from the sample or experimental observations in a manner such that in the long run the average of the estimates so obtained will equal the parameter value.

Multiple regression. The fitting of an equation (usually linear) involving the dependent variable (Y) as a function of several independent variables (X's) to a set of observed points involving these same variables.

Analysis of variance. The process of breaking down the total variation (variance) obtained in an experiment into independent sums attributable to the various components of the experiment including the experimental error so as to be able to make tests of significance of hypotheses associated with these components or factors.

Residual mean square. The average square of the discrepancy observed between experimental observations and the mathematical model used to fit such observations.

Standard error of estimate (*of coefficients*). A measure of the variation (standard deviation) associated with the estimates obtained by an experiment. Under normal conditions one expects about 68% of the estimates to be within one standard error of their true value and 95% within two standard errors.

Replication. The process of taking an additional observation (repeating the experiment) using the same values for the controllable factors as for the original observation.

method

The response, Y, is supposed dependent upon n variables X_i, which are capable of exact measurement and control. The form of the functional relationship $Y = f(X_1, X_2, \cdots, X_n)$ is unknown, and the problem is to find the combination of values of X_i which optimize the response within the region of the n-dimensional factor space where experimentation is feasible using as few experimental observations as possible. The number of observations required will, of course, depend upon the accuracy and precision of estimation desired. Where the problem is one of minimization, it can always be converted to one of maximization; for example, by considering the improvement as compared with some standard instead of the actual level achieved.

The technique assumes that the response function can be satisfactorily represented by a quadratic form in the area of interest, i.e., $Y = \sum_{i=0}^{n} \sum_{j\geq i}^{n} c_{ij}X_iX_j + e$ where Y is the property to be maximized, the X_i are the levels of the n independent variables ($X_0 = 1$), the c_{ij} are the unknown parameters to be estimated from the experiment and e is the residual or experimental error. The adequacy of the quadratic surface representation of the true response surface of the process being investigated depends on the use of a small sub-region of the factor space within which one restricts his determinations. In some experimental situations, such a small neighborhood within which the optimum point can be assumed to lie is already known to the experimenter from previous experience. However, if this is not the case, the procedure of locating optimum conditions involves two distinct phases. The first phase involves the location of the neighborhood, while the second is to determine within the neighborhood the optimum point.

location of neighborhood

The location of the neighborhood is accomplished by using what is called the *method of steepest ascent.* In this procedure, one assumes that the surface can be represented locally by a sloping plane. Starting at any point, P, the experimenter estimates the coefficients or slopes of the plane $Y = b_0 + b_1X_1 + b_2X_2 + \cdots b_nX_n$ by performing a suitably arranged set of trials in a small sub-region about P. From these observations, the coefficients are estimated and one then calculates the direction of steepest ascent or greatest slope up the plane. He then proceeds to a point, Q, in this direction, where new observations are made, the slopes are redetermined, and the process repeated. In this way, by a step-by-step procedure, points of higher and higher response are reached.

This procedure cannot, however, be used to actually reach the maximum response point since, as one goes farther up the surface, the slopes become more gradual and thus more difficult to estimate. The second-order terms also become relatively more important. The procedure generally followed is to compare the linear effects with the error variance and with the second-order effects, and if the linear model appears adequate, the path of steepest ascent is determined. At the point of diminishing returns, the new point is located around which the process is repeated.

The experimental design used during the first phase where one is seeking the path of steepest ascent from a given point on the surface is generally of a two-level factorial type, where the origin for each variable is taken at the initial point and the levels used are equidistant from it in either direction. Thus, in a three-variable situation, one would use a 2^3 factorial design, and the eight experimental points would be as shown in Table 1.

Table 1 Experimental points for a 2^3 factorial design

	Factor Level		
Point	X_1	X_2	X_3
1	+1	+1	+1
2	−1	+1	+1
3	+1	−1	+1
4	−1	−1	+1
5	+1	+1	−1
6	−1	+1	−1
7	+1	−1	−1
8	−1	−1	−1

The estimation of the b's from this type of design is straightforward. In fact, if σ^2 is the experimental error variance we have $b_i = \sum X_iY/\sum X_i^2$, and $V(b) = \sigma^2/\sum X_i^2$ (the variance of b).

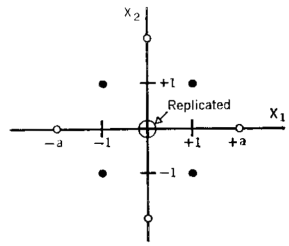
One thus has the essential ingredients needed to complete the first phase of the investigation.

location within neighborhood

In considering the second phase, we assume that we have identified a point P that is in the neighborhood of the optimum point. The experimental designs used at this stage of the problem are known as composite designs. There are two types of composite designs, central and non-central. The central composite designs consider the 2^n factorial designs and adds additional points with high and low levels for each variable as well as additional points at the center of the design.

The central composite design for $n = 2$ is shown in Figure 1. The 2^2 factorial points are given as solid points while the added points are open circles.

Figure 1 A two-dimensional central composite design



For the purpose of estimating the parameters of the quadratic form, the central composite design can be shown to be more efficient than the 3^n factorial design. As one might expect, this means that a savings in experimental points can be realized, since interest has been narrowed to estimating the optimum response point rather than to generally studying the nature of the mathematical model that explains the process under study.

As can be seen from Figure 1, one of the problems needing solution is to select the value of a to use when adding points to the factorial design. It is apparent that care is required in selecting a, since large values of a will tend to reduce the adequacy of the representation of the surface by the quadratic, while small values may fail to encompass the optimum point. One solution to the problem is to select $a = 2^{n/4}$, where the scale is determined by the levels used in the factorial. It should be noted that when several variables are involved in the system, it may be possible to choose a suitable fraction of the factorial points and still obtain unbiased estimates of the parameters.

The location of an optimum point usually requires a series of coordinated experiments, especially when one must first find the neighborhood of the optimum. If the process being studied has little or no time effect, so that one can combine results that are obtained at different time intervals, the series of experiments can often be developed into an organized sequential program. The non-central composite designs are useful if one uses such a sequential approach to his experimentation. The factorial portion and the central point are run first and, if the optimum is found to be close to the center being used in the factorial design, the additional points required for the central composite design are then used. If, however, the optimum response is nearer one of the other points the factorial portion is augmented to form a non-central composite design. Of course, if it is indicated that a new location should be sought through the use of the path of steepest ascent, then the sequence is as follows. The fitting of the quadratic surface, $Y = \sum_{i=0}^{n} \sum_{j \geq i}^{n} c_{ij} X_i X_j + e$, to the observations realized from the composite design can be obtained by standard multiple regression techniques. Following the estimation of the coefficients, one can perform an analysis of variance on the results to establish the significance of the several coefficients as well as the significance of the regression itself. If one has some prior information as to the value of σ^2, the experimental error, this information can be used in a comparison with the residual mean square associated with the regression analysis to provide a test of goodness of fit of the second-degree equation. If the fit is not satisfactory, one may change his neighborhood if this seems required, or increase the order of the regression equation.

When such a test has indicated that an adequate fit has been obtained the fact that an individual coefficient is or is not statistically significant is of no practical significance. What this means is that one might just as well retain the small coefficient in his future analyses since there appears to be no real good reason for making the hypothesis that one of the coefficients is actually zero in the population model.

When the second degree equation has been fitted, it is necessary to interpret it to see if one can, in fact, determine the coordinates of the optimum response point. Since the coefficients in general quadratic do not readily convey to the observer the nature of the surface being represented, one usually resorts to a canonical reduction of the equation so as to obtain the canonical form, $Y = B_0 + B_{11}X_1^2 + B_{22}X_2^2 + \cdots B_{nn}X_n^2$.

There are many types of surfaces that can be obtained through the use of the quadratic function. Under certain conditions, including those where all the B's are negative, there will be a point maximum in all the variables. Another situation, however, that may be encountered is where the maximum is in fact remote from the region of the design, but the surface is elongated along an axis which passes close to the design. This indicates that the previous experimentation has brought the experimenter not to a maximum but close to a rising ridge of the surface. No conclusion as to optimum conditions can be drawn in this latter case, but one can, from observation of the nature of the rising ridge, determine where additional experimentation should be carried out in attempting to locate the optimum point. In the case that the optimum point falls within the region of the experiment, its position can be obtained by differentiating the original quadratic with respect to the variables X_1, X_2, $\cdots$ X_n in turn and equating the results to zero. This will yield a set of linear equations which, when solved simultaneously, give the coordinates of the optimum point. It should be emphasized, however, that the nature of the surface should be critically examined through the use of the canonical transformations approach before one seeks these coordinates. In fact, as the dimension of the problem increases, making a careful examination becomes most important.

programming considerations

The mechanics of analyzing the data obtained from the sequence of observations made in following the approach outlined above can be readily adapted to digital computer programs. In fact, many of the procedures make use of techniques for which standard computer programs are already generally available. Thus in the initial phase, where one is interested in following the path of steepest ascent using a linear fit to the experimental data, multiple regression computer programs are applicable. These programs give not only the best estimates for the regression coefficients but also their standard errors as well as the standard error of estimate for the response variable. Through the use of transformations the significance of the quadratic terms in the surface can also be readily tested using the same computer program. This enables one to determine when to abandon the steepest ascent phase of the investigation.

In the calculation of the actual path of steepest ascent, the successive differentiation of the fitted linear relationship yields simultaneous linear equations whose solution can be obtained from standard programs for solving systems of simultaneous equations. In fact, even the determination of the possible steps up the path through the computation of coordinates of the points on the path can easily be programmed.

When one reaches the point of fitting the quadratic surface to the data obtained from a composite design, the determination of the coefficients of the surface, their standard errors and the standard error of estimate is also a multiple regression program application. The quadratic terms are simply treated as new linear variables in this case. The determination of the optimum point is again the solution of a set of simultaneous linear equations.

The other major computational task encountered in using the response surface analysis concept involves the transformation of the general quadratic equation to one in canonical form, that is, an equation involving only pure quadratic terms. To determine the coefficients of the canonical form one must solve the characteristic equation of the determinant of the coefficients of the original quadratic. Such solutions have been programmed for computers, and since the roots of the equation are the necessary coefficients, the computational activity has been essentially mechanized through the use of these standard computer programs.

selected papers

It is suggested that the reader intending to apply the Box-Wilson method consult the appended list of selected papers on specific applications in order to gain additional detailed insight on various aspects of the technique.

concluding remarks

One should note that the techniques associated with response surface analysis center around the use of sequential experimentation. Generally, each set of experiments in the sequence is performed within a given sub-region, and one must decide from the results obtained what to do next. Several alternatives present themselves:

1. There is insufficient reliability in the results obtained to draw any conclusions. In this case, replication of the experimental points may be used or improvement sought in the overall experimental design.
2. The first order effects are dominant in the region being studied. In this case, one may use the method of steepest ascent in order to determine the new center about which to perform the next set of experiments.
3. The first and second order effects are important in the surface. In this case, the original set of observational points should be

augmented so as to yield a composite design from which an optimum response analysis can be performed.

CITED REFERENCES AND FOOTNOTES

1. With relatively modest assumptions about the particular nature of *f*.
2. Box, G. E. P. and Wilson, K. B., *On the Experimental Attainment of Optimum Conditions*, J. Royal Statistical Society, Ser. B, v. 13 (1951), pp. 1–45.

PAPERS ON APPLICATIONS

- "Experience with Response Surface Design" by R. M. DeBaun and A. M. Schneider. American Cyanamid Co., Stanford, Connecticut. *Proceedings of Symposium on Design of Industrial Experiments*, Nov. 5–9, 1956, Institute of Statistics, University of North Carolina.
- "Experimental Designs for the Determination of Optimum Conditions." Dale E. Cooper, Continental Oil Company. *Application of Statistics and Computers to Fuel and Lubricant Research Problems*. Southwest Research Institute, San Antonio, Texas, March 1962.
- "Effects of Raw Material Ratios on Absorption of Whiteware Compositions." W. C. Hockler, W. W. Kriegel, R. J. Hoder. *Journal of American Ceramic Society*, Volume 39, No. 20.
- "An Investigation of Some of the Relationships between Copper, Iron and Molybdenum in the Growth and Nutrition of Lettuce." R. J. Hoder. *Proc. of American Soil Society*, Volume 21, No. 59.
- "Application of Statistical Procedures to a Study of the Flooding Capacity of a Pulse Column." Pike, F. P. *Chemical Engineering Technical Report*, North Carolina State College, Raleigh.
- "The Use of Box-Wilson Techniques in the Study of a Titania Pigment Process." B. S. Sanderson. *Statistical Methods in the Chemical Society, 1953*. American Society for Quality Control.
- "The Determination of Response Surfaces for Textile Resin Finishes." *Statistical Methods in the Chemistry Industry*, 1957. American Society of Quality Control.

BIBLIOGRAPHY

- Box, G. E. P., *The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples*, Biometrics, v. 10, (1954), pp. 16–60.
- Box, G. E. P. and Youle, P. V., *The Exploration and Exploitation of Response Surfaces: An Example of the Link Between the Fitted Surface and the Basic Mechanism of the System*, Biometrics, v. 11, (1955), pp. 287–323.
- Box, G. E. P. and Hunter, J. S., *Multi-Factor Experimental Designs for Exploring Response Surfaces*, Annals of Math. Statistics, v. 28, (1957), pp. 195–241.
- Box, G. E. P. and Draper, N. R., *A Basis for the Selection of a Response Surface Design*, Journal American Statistical Association, v. 54, (1959), pp. 622–654.
- Davies, O. L., *Design and Analysis of Industrial Experiments*, Hofner Publishing Company, New York, (1956), Chapter II.