UNIVERSITY OF WISCONSIN
Madison, Wisconsin

Technical Report No. 33

July, 1964

"An Introduction to Response Surface Methodology"

by

G. E. P. Box

*Originally prepared for International Encyclopedia of the Social Sciences; reproduced by permission of the Crowell-Collier Co.
Response surface methodology is a statistical technique for the design and analysis of experiments that seeks to relate an average response to the values of quantitative variables that affect response. For example, response in a chemical investigation might be yield of sulphuric acid, and the quantitative variables affecting yield might be pressure and temperature of the reaction.

In a psychological experiment an investigator might want to find out how a test score achieved by certain subjects depended upon duration of the period during which they studied
the relevant material and the delay between study and test. In mathematical language this investigator is interested in the presumed functional relationship $\eta = f(\xi_1, \xi_2)$ which expresses the response score ($\eta$) as a function of the two variables duration ($\xi_1$) and delay ($\xi_2$). If repeated experiments were made in each of which he attempted to hold these variables at the same fixed levels, the measured response would nevertheless vary because of measurement errors, observational errors, and variability in the experimental material. We conceive $\eta$ therefore as the mean response at particular conditions and we say that $y$, the response actually observed in a particular experiment differs from $\eta$ because of an (all inclusive) error $e$. Thus $y = \eta + e$ and we can write a mathematical model relating the observed response to the levels of $k$ variables in the form

$$y = f(\xi_1, \ldots, \xi_k) + e$$

(1)

The appropriate investigational strategy depends heavily on the state of ignorance concerning the functional form $f$. At one extreme the investigator may not know even which variables $\xi$ to include and must make a preliminary screening investigation. At the other extreme the true functional form may actually be known or can be deduced from a mechanistic theory.
Response surface methods are appropriate in the intermediate situation. The important variables are known but the true functional form is neither known nor is it easily deducible.

The general procedure is to approximate \( f \) locally by suitable functions, such as polynomials, which act as "mathematical french curves". This enables us to carry out a series of experimental trials where the location for the next set of trials depends on the preceding observations.

**Geometric Representation of Response Relationships**

The three curves of figure la, showing a hypothetical relationship associating percent recall and initial learning period for three different periods of delay, are shown in figure lb as sections of a response surface. This surface is represented by its response contours in figure lc. Figure ld shows how a third variable may be accommodated by the use of three dimensional contour surfaces.

**Local Graduation**

It is usually most convenient to work with coded variables such as

\[ x_1 = (\xi_1 - \xi_1^0) / S_1, \quad x_2 = (\xi_2 - \xi_2^0) / S_2 \]

in which \( \xi_1^0, \xi_2^0 \) are the coordinates of the center of a region of current interest and \( S_1 \) and \( S_2 \) are convenient scale factors.

We shall use the notation \( \hat{y} \) to mean the calculated value of the response obtained by fitting an approximating function by the method of least squares. In a region like \( R_1 \) in figure lc an adequate approximation can be obtained by
fitting the first degree polynomial

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 \]  \hspace{1cm} (2)

The response contours of such a fitted plane are, of course, equally spaced parallel straight lines. In a region like \( \mathbb{R}_2 \) fair approximation might be achieved by fitting a second degree polynomial

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + b_{12} x_1 x_2 \]  \hspace{1cm} (3)

Flexibility of functions like those in equations (2) and (3) is greatly increased if we allow the possibility that \( y, x_1, \) and \( x_2, \) are suitably transformed values of the response and of the variables. For example, it might be appropriate to analyze log score rather than score itself.

Ways of choosing suitable transformations are described in references [5] and [6].
Uses of Response Surface Methodology

A special pattern of points at which observations are to be made is called an experimental design. In figure 10 are shown a first order design in \( R_1 \), suitable for fitting and checking a first degree polynomial, and a second order design in \( R_2 \), suitable for fitting and checking a second degree polynomial. Response surface methodology has been applied

a) To provide a description of how the response is affected by a number of variables over some already chosen region of interest

b) To study and exploit multiple response relationships and constrained extrema

In drug therapy, for example, the true situation might be as depicted in figure 2. First degree approximating functions fitted to each of the three responses \( \eta_1 \) therapeutic effect, \( \eta_2 \) nausea, and \( \eta_3 \) toxicity could approximately locate the point P where maximum therapeutic effect is obtained with nausea and toxicity maintained at the acceptable limits \( \eta_2 = 5 \), \( \eta_3 = 30 \).
c) To locate and explore the neighborhood of maximal or minimal response.

Because problems in (c) often subsume those in (a) and (b) we consider only this application in more detail.

**Location and Exploration of a Maximal Region**

Various tactics have been proposed to deal with this problem. However, because the appropriateness of a particular tactic usually depends upon factors which are initially unknown an adaptive strategy of multiple iteration must be employed. That is, we must put ourselves in a position to learn more about each of a number of uncertainties as we proceed and to modify tactics accordingly. It is doubtful if an adaptive strategy could be found appropriate to every conceivable response function. We illustrate one such procedure which has worked well in chemical applications and ought to be applicable in some other areas.

**An Example**

In this example iteration occurs in A) the amount of replication (to achieve sufficient accuracy), B) the location of the region of interest, C) the scaling of the variables, D) the transformation in which the variables are considered, E) the necessary degree of complexity of approximating functions and of the corresponding design. The letters A), B), etc. are used parenthetically to indicate the particular type of iteration which is being furthered at any stage. Suppose that unknown to the experimenter the true dependence of percentage yield on temperature and concentration is as shown in figure 3a and the
experimental error standard deviation is 1.2%.

A First Degree Approximation

Suppose that five initial duplicate runs made in random order at points labelled 1, 2, 3, 4, 5 in figure 3b yield the results \( y_1^* = 24 \), \( y_4^* = 42 \), \( y_1 = 24.5 \), \( y_2 = 38 \), \( y_3 = 42 \), \( y_4^* = 42 \), \( y_4 = 42 \), \( y_5 = 50 \), \( y_5^* = 53 \).

The average yields at the five points are then \( \bar{y}_1 = 24.5 \), \( \bar{y}_2 = 39 \), \( \bar{y}_3 = 42 \), \( \bar{y}_4 = 41.5 \), \( \bar{y}_5 = 51.5 \). At this stage it is convenient to work with the coded

variables \( x_1 = \frac{\text{Temp} - 70}{10} \) and \( x_2 = \frac{\text{Conc} - 42.4}{2.5} \).

Using standard least squares theory the coefficients \( b_0 \), \( b_1 \), and \( b_2 \) of equation (2) are then easily estimated, for example \( b_1 = \frac{1}{4} \left\{ -\bar{y}_1 + \bar{y}_2 - \bar{y}_4 + \bar{y}_5 \right\} = 5.9 \), and the locally best fitting plane is \( \hat{y} = 39.9 + 5.9 x_1 + 6.9 x_2 \). \( (4) \)

The differences in the duplicate runs provide an estimate \( s = 1.5 \) of \( \sigma \) having five degrees of freedom. The standard errors of \( b_0 \), \( b_1 \) and \( b_2 \) are then easily calculated as \( \pm 0.5 \), and no further replication (A) appears necessary to obtain adequate estimation of \( y \).

Checking the Fit

To check the appropriateness of the first degree equation it would be sensible to look at the size of second order effects. For reason of experimental economy a first order design usually contains points at too few distinct levels to allow separate estimation of all second order terms in (3). The design may be chosen, however, so as to allow estimates of "specimen" second order coefficients or combinations thereof to be examined. In the present case we can calculate \( b_{12} = \frac{1}{4} (\bar{y}_1 - \bar{y}_2 - \bar{y}_3 + \bar{y}_4) = -0.9 \pm 0.5 \) and
\((b_{11} + b_{22}) = \frac{1}{4} (\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4) - \bar{y}_5 = -2.6 \pm 1.2\). Some inadequacy of the first degree equation is indicated, therefore, but this is tentatively ignored because of the dominant magnitude of \(b_1\) and \(b_2\).

**Steepest Ascent**

It now is logical to explore \((B)\) higher temperatures and concentrations. The points 6, 7 and 8 are along a steepest ascent path obtained by changes proportional to \(b_1 \times S_2 = 5.9 \times 10^6 = 59^\circ\) in temperature and \(b_2 \times S_2 = 6.9 \times 2.5\% = 17.25\%\) in concentration. Let us suppose that \(y_6 = 59, y_7 = 63\) and \(y_8 = 50\). Graphical interpolation indicates that the highest yield on this path is between runs 6 and 7 and this is chosen \((B)\) as the center of the new region to be explored.

The path calculated as above is at right angles to contours of the fitted plane when 10° units of temperature and 2.5% units of concentration are represented by the same distances. That the experimenter currently regards these units as appropriate is implied by his choice of levels in the design.

**Scaling Correction**

To correct unsuitable scaling \((C)\) we can adopt the rule that if a variable produces an effect which is small compared with that produced by the other variables, the center level for that variable is moved away from the calculated path and a larger change is made for this variable in the next set of runs. No change of relative scaling is indicated here but we would normally expect our progress up the surface to be accompanied by reduction in the sizes of \(b_1\) and \(b_2\). Also our checks have already indicated that second order effects can scarcely be estimated with adequate accuracy in the present scaling. Thus,
we should employ wider ranges in both variables in a second design.

A Second Degree Approximation

A widened first order design at the new location might give \( y_9 = 50 \)
\( y_{10} = 66 \), \( y_{11} = 66 \), \( y_{12} = 63 \), \( y_{13} = 52 \), as in figure 3c.

The best fitting plane is

\[
\hat{y} = 59.4 + 1.3 x_1 - 0.3 x_2 \quad \text{with} \quad x_1 = \frac{\text{Temp} - 90}{15}, \quad x_2 = \frac{\text{Conc} - 18.75}{3.75},
\]

and the estimated standard error of the coefficients is about \( \pm 0.8 \) and
\( (b_{11} + b_{22}) = -8.25 \pm 1.7 \). It is clear without this time duplicating the design
that the first order terms no longer dominate and no worthwhile further pro-
gress can be made by ascent methods. To make possible the fitting and
checking of a second degree polynomial (E) five additional observations
might be taken, say \( y_{14} = 54 \), \( y_{15} = 54 \), \( y_{16} = 57 \), \( y_{17} = 65 \), \( y_{18} = 55 \). The
last ten observations now form a second order design. A second degree
equation fitted to these observations gives

\[
\hat{y} = 65.50 + 1.16 x_1 + 0.05 x_2 - 4.31 x_1^2 - 4.81 x_2^2 - 6.75 x_1 x_2
\]

(5)

The design allows us to check the adequacy (E) of the second degree
equation by providing estimates of certain "specimen" combinations of third
order terms

\[
(3b_{111} + b_{122}) = \frac{1}{2} \{y_9 - y_{10} + y_{12} - y_{13} - \sqrt{2y_{15}} - \sqrt{2y_{16}}\} = -0.19 \pm 1.03
\]

\[
(3b_{222} + b_{112}) = \frac{1}{2} \{y_9 + y_{10} - y_{12} - y_{13} - \sqrt{2y_{14}} - \sqrt{2y_{19}}\} = -0.60 \pm 1.03
\]

The standard errors of the linear coefficients \( b_1 \) and \( b_2 \) quadratic coeffi-
cients \( b_{11} \) and \( b_{22} \) and of the interaction coefficient \( b_{12} \) are respectively
\( \pm 0.52, \pm 0.62, \pm 0.73 \).
Before attempting to interpret equation (5) we need some reassurance (A) that the change in response it predicts is large compared with the standard error of that prediction. For a design requiring N observations and an approximating equation containing p constants the average variance of the N calculated responses \( \hat{y} \) is 
\[
\frac{p}{N} s^2 = \frac{6}{12} \times 2.1 = 1.1 \text{ for this example.}
\]
The square root (1.0 for this example) gives an "average" standard error for \( \hat{y} \). This may be compared with the range of the predicted \( \hat{y} \)'s which is 17.08, the highest predicted value being \( \hat{y}_{11} = 65.50 \) and the lowest \( \hat{y}_9 = 48.42 \).

A more precise indication of adequacy may be obtained by an application of the Analysis of Variance. A discussion of this is outside of the scope of the present account. It is to be noted, however, that bare statistical significance of the regression would not insure that the response surface was estimated with sufficient accuracy for the interpretation discussed below.

**Interpretation**

Once adequate fit and precision have been obtained a contour plot of the equation over the region of the experimental design is helpful in interpretation. Especially where there are more than two variables, interpretation is further facilitated by writing the second degree equation in canonical form (D). In most cases, this means that the center of the quadratic system is chosen as a new origin and a rotation of axes is performed to eliminate cross-product terms.

Figure 3d shows how in a final group of experiments the new canonical axes and scales could be used to position the design. The design points
are so chosen in figure 3d they roughly follow a contour and make a rather precise final fitting possible.

It might be asked: why not simply use the 20 or so experimental points to cover the region shown in figure 3a with some suitable grid in the first place? The answer is that we do not know initially that the region of interest will be in the area covered by that diagram. The "content" of the space to be explored rapidly goes up as the number of dimensions is increased.

Suitable Designs

From the above discussion it will be clear that the arrangements of experimental points suitable for response surface study should satisfy a number of requirements. A response surface design should: (i) allow \( \hat{y}(x) \) to be estimated throughout the region of interest \( R \) (ii) ensure that \( \hat{y}(x) \) is as "close" as possible to \( \eta(x) \) (iii) give good detectability of lack of fit (iv) allow transformations to be fitted (v) allow experiments to be performed in blocks (vi) allow designs of increasing order to be built up sequentially (vii) provide an internal estimate of error (viii) be insensitive to wild observations (ix) require a minimum number of experimental points (x) provide patterning of data allowing ready visual appreciation (xi) ensure simplicity of calculation (xii) behave well when errors occur in settings of the \( x \)'s.

A variety of designs have been developed, many of which have remarkably good overall behavior with respect to these requirements.
BIBLIOGRAPHY


Figure (1c)

delay $\xi_2$ (weeks)

learning period $\xi_1$ (hours)
Figure 2  Dependence of three responses on combined dosages of drugs.
Figure 3(c) Second order design with fitted contours

Figure 3(d) Final design in canonical variables $x_1$ and $x_2$
Figure 3(a) True yield contours

Figure 3(b) First order design with steepest ascent path