1

INTRODUCTION

1.1 RESPONSE SURFACE METHODOLOGY

Response surface methodology (RSM) is a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes. It also has important applications in the design, development, and formulation of new products, as well as in the improvement of existing product designs.

The most extensive applications of RSM are in the industrial world, particularly in situations where several input variables potentially influence performance measures or quality characteristics of the product or process. These performance measures or quality characteristics are called the **response**. They are typically measured on a continuous scale, although attribute responses, ranks, and sensory responses are not unusual. Most real-world applications of RSM will involve more than one response. The input variables are sometimes called **independent variables**, and they are subject to the control of the engineer or scientist, at least for purposes of a test or an experiment.

Figure 1.1 shows graphically the relationship between the response variable yield (y) in a chemical process and the two process variables (or independent variables) reaction time (ξ_1) and reaction temperature (ξ_2). Note that for each value of ξ_1 and ξ_2 there is a corresponding value of yield y and that we may view these values of the response yield as a surface lying above the time–temperature plane, as in Fig. 1.1a. It is this graphical perspective of the problem environment that has led to the term **response surface methodology**. It is also convenient to view the response surface in the two-dimensional time–temperature plane, as in Fig. 1.1b. In this presentation we are looking down at the time–temperature plane and connecting all points that have the same yield to produce contour lines of constant response. This type of display is called a **contour plot**.

Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Fourth Edition. by Raymond H. Myers, Douglas C. Montgomery, and Christine M. Anderson-Cook.

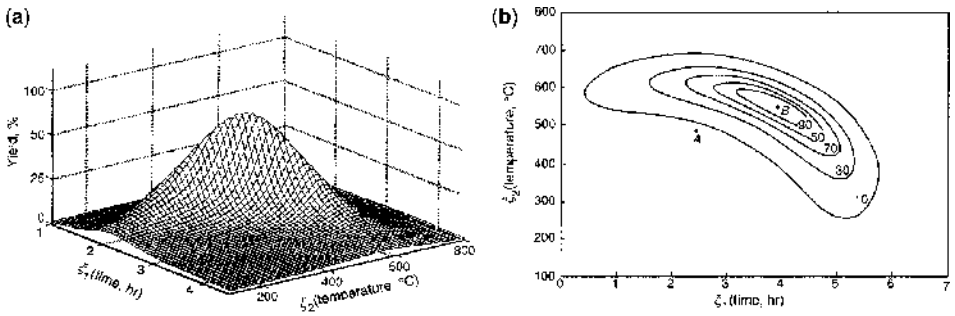

Figure 1.1 (**a**) A theoretical response surface showing the relationship between yield of a chemical process and the process variables reaction time (ξ_1) and reaction temperature (ξ_2). (**b**) A contour plot of the theoretical response surface.

Clearly, if we could easily construct the graphical displays in Fig. 1.1, optimization of this process would be very straightforward. By inspection of the plot, we note that yield is maximized in the vicinity of time $\xi_1 = 4$ hr and temperature $\xi_2 = 525°\text{C}$. Unfortunately, in most practical situations, the true response function in Fig. 1.1 is unknown. The field of response surface methodology consists of the experimental strategies for exploring the space of the process or independent variables (here the variables ξ_1 and ξ_2), empirical statistical modeling to develop an appropriate approximating relationship between the yield and the process variables, and optimization methods for finding the levels or values of the process variables ξ_1 and ξ_2 that produce desirable values of the responses (in this case that maximize yield).

1.1.1 Approximating Response Functions

In general, suppose that the scientist or engineer (whom we will refer to as the **experimenter**) is concerned with a product, process, or system involving a response y that depends on the controllable input variables $\xi_1, \xi_2, \ldots, \xi_k$. These input variables are also sometimes called factors, **independent variables**, or process variables. The actual relationship can be written

$$y = f(\xi_1, \xi_2, \ldots, \xi_k) + \varepsilon \tag{1.1}$$

where the form of the true response function f is unknown and perhaps very complicated, and ε is a term that represents other sources of variability not accounted for in f. Thus ε includes effects such as measurement error on the response, other sources of variation that are inherent in the process or system (background noise, or common/special cause variation in the language of statistical process control), the effect of other (possibly unknown) variables, and so on. We will treat ε as a **statistical error**, often assuming it to have a normal distribution with mean zero and variance σ^2. If the mean of ε is zero, then

$$\begin{aligned} E(y) \equiv \eta &= E[f(\xi_1, \xi_2, \ldots, \xi_k)] + E(\varepsilon) \\ &= f(\xi_1, \xi_2, \ldots, \xi_k) \end{aligned} \tag{1.2}$$

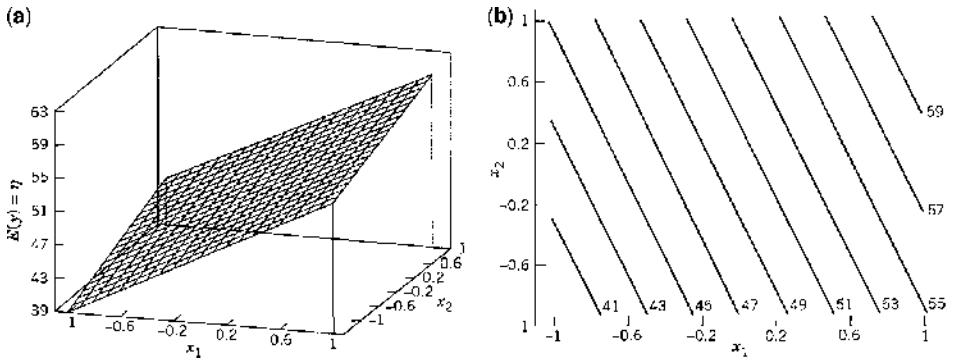

Figure 1.2 (**a**) Response surface for the first-order model $\eta = 50 + 8x_1 + 3x_2$. (**b**) Contour plot for the first-order model.

The variables $\xi_1, \xi_2, \ldots, \xi_k$ in Equation 1.2 are usually called the **natural variables**, because they are expressed in the natural units of measurement, such as degrees Celsius (°C), pounds per square inch (psi), or grams per liter for concentration. In much RSM work it is convenient to transform the natural variables to **coded variables** $x_1, x_2, \ldots, x_k$, which are usually defined to be dimensionless with mean zero and the same spread or standard deviation. In terms of the coded variables, the true response function (1.2) is now written as

$$\eta = f(x_1, x_2, \ldots, x_k) \tag{1.3}$$

Because the form of the true response function f is unknown, we must approximate it. In fact, successful use of RSM is critically dependent upon the experimenter's ability to develop a suitable approximation for f. Usually, a **low-order polynomial** in some relatively small region of the independent variable space is appropriate. In many cases, either a **first-order** or a **second-order** model is used. For the case of two independent variables, the first-order model in terms of the coded variables is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1.4}$$

Figure 1.2 shows the three-dimensional response surface and the two-dimensional contour plot for a particular case of the first-order model, namely,

$$\eta = 50 + 8x_1 + 3x_2$$

In three dimensions, the response surface for y is a plane lying above the x_1, x_2 space. The contour plot shows that the first-order model can be represented as parallel straight lines of constant response in the x_1, x_2 plane.

The first-order model is likely to be appropriate when the experimenter is interested in approximating the true response surface over a relatively small region of the independent variable space in a location where there is little curvature in f. For example, consider a small region around the point A in Fig. 1.1b; the first-order model would likely be appropriate here.

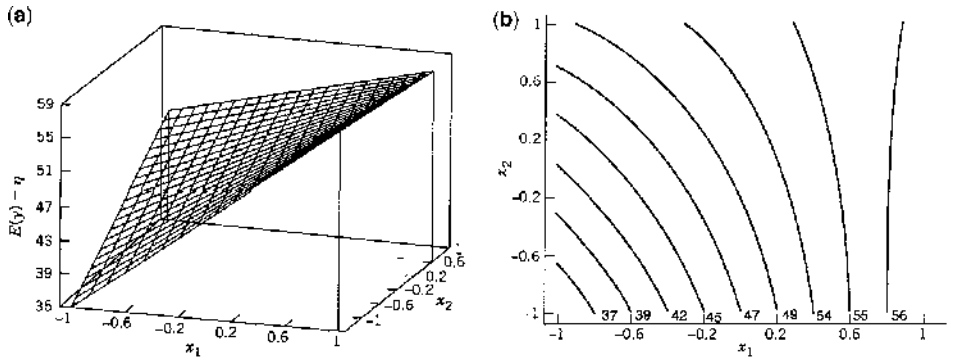

Figure 1.3 (**a**) Response surface for the first-order model with interaction $\eta = 50 + 8x_1 + 3x_2 - 4x_1x_2$. (**b**) Contour plot for the first-order model with interaction.

The form of the first-order model in Equation 1.4 is sometimes called a **main effects model**, because it includes only the main effects of the two variables x_1 and x_2. If there is an **interaction** between these variables, it can be added to the model easily as follows:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \tag{1.5}$$

This is the **first-order model with interaction**. Figure 1.3 shows the three-dimensional response surface and the contour plot for the special case

$$\eta = 50 + 8x_1 + 3x_2 - 4x_1x_2$$

Notice that adding the interaction term $-4x_1x_2$ introduces curvature into the response function. This leads to different rates of change of the response as x_1 is changed for different fixed values of x_2. Similarly, the rate of change in y across x_2 varies for different fixed values of x_1.

Often the curvature in the true response surface is strong enough that the first-order model (even with the interaction term included) is inadequate. A **second-order model** will likely be required in these situations. For the case of two variables, the second-order model is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \tag{1.6}$$

This model would likely be useful as an approximation to the true response surface in a relatively small region around the point B in Fig. 1.1b, where there is substantial curvature in the true response function f.

Figure 1.4 presents the response surface and contour plot for the special case of the second-order model

$$\eta = 50 + 8x_1 + 3x_2 - 7x_{11}^2 - 3x_{22}^2 - 4x_1x_2$$

Notice the mound-shaped response surface and elliptical contours generated by this model. Such a response surface could arise in approximating a response such as yield, where we would expect to be operating near a maximum point on the surface.

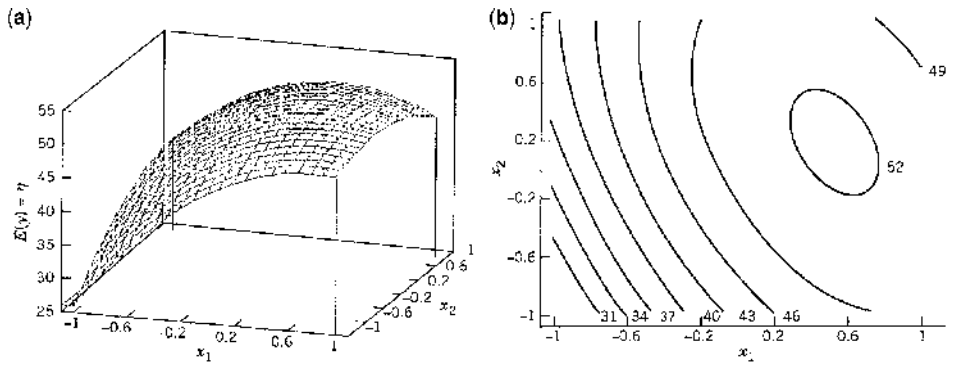

Figure 1.4 (**a**) Response surface for the second-order model $\eta = 50 + 8x_1 + 3x_2 - 7x_1^2 - 3x_2^2 - 4x_1x_2$. (**b**) Contour plot for the second-order model.

The second-order model is widely used in response surface methodology for several reasons. Among these are the following:

1. The second-order model is very **flexible**. It can take on a wide variety of functional forms, so it will often work well as an approximation to the true response surface. Figure 1.5 shows several different response surfaces and contour plots that can be generated by a second-order model.

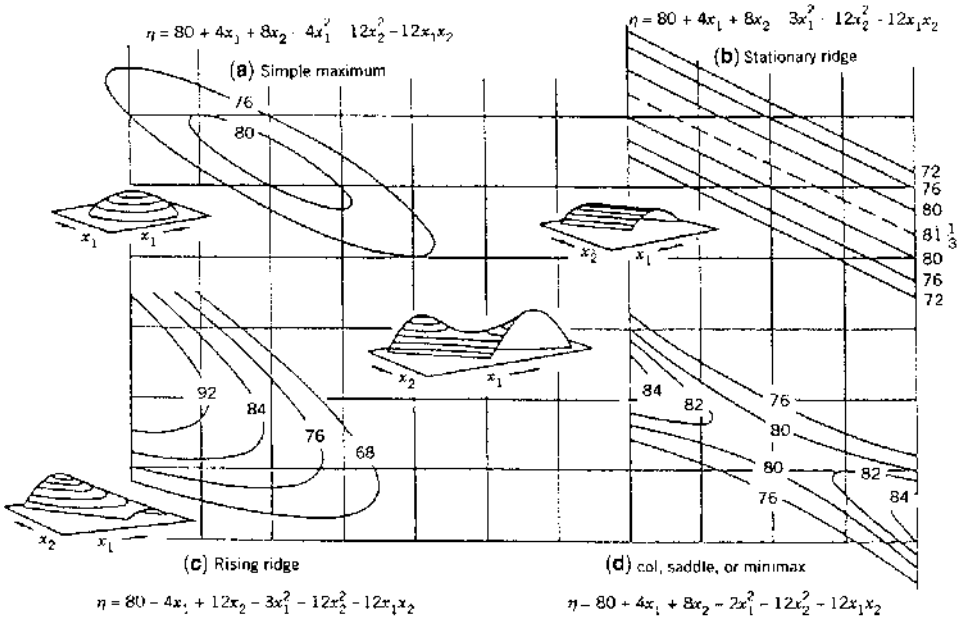

Figure 1.5 Some examples of types of surfaces defined by the second-order model in two variables x_1 and x_2. (Adapted with permission from *Empirical Model Building and Response Surfaces*, G. E. P. Box and N. R. Draper, John Wiley & Sons, New York, 1987.)

2. It is **easy** to **estimate the parameters** (the β's) in the second-order model. The method of least squares, which is presented in Chapter 2, can be used for this purpose.
3. There is considerable **practical experience** indicating that second-order models work well in solving real response surface problems.

In general, the first-order model is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{1.7}$$

and the second-order model is

$$\eta = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + \sum_{i<j=2}^{k} \sum \beta_{ij} x_i x_j \tag{1.8}$$

In some situations, approximating polynomials of order higher than two are used. The general motivation for a polynomial approximation for the true response function f is based on the **Taylor series expansion** around the point $x_{10}, x_{20}, \ldots, x_{k0}$. For example, the first-order model is developed from the first-order Taylor series expansion

$$\begin{aligned} f \cong f(x_{10}, x_{20}, \ldots, x_{k0}) &+ \left.\frac{\partial f}{\partial x_1}\right|_{\mathbf{x}=\mathbf{x}_0} (x_1 - x_{10}) \\ &+ \left.\frac{\partial f}{\partial x_2}\right|_{\mathbf{x}=\mathbf{x}_0} (x_2 - x_{20}) + \cdots + \left.\frac{\partial f}{\partial x_k}\right|_{\mathbf{x}=\mathbf{x}_0} (x_k - x_{k0}) \end{aligned} \tag{1.9}$$

where $\mathbf{x}$ refers to the vector of independent variables and $\mathbf{x}_0$ is the vector of independent variables at the specific point $x_{10}, x_{20}, \ldots, x_{k0}$. In Equation 1.9 we have only included the first-order terms in the expansion, so if we let $\beta_0 = f(x_{10}, x_{20}, \ldots, x_{k0})$, $\beta_1 = (\partial f/\partial x_1)\big|_{\mathbf{x}=\mathbf{x}_0}, \ldots, \beta_k = (\partial f/\partial x_k)\big|_{\mathbf{x}=\mathbf{x}_0}$, we have the first-order approximating model in Equation 1.7. If we were to include second-order terms in Equation 1.9, this would lead to the second-order approximating model in Equation 1.8.

Finally, note that there is a close connection between RSM and **linear regression analysis**. For example, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

The β's are a set of unknown parameters. To estimate the values of these parameters, we must collect data on the system we are studying. Regression analysis is a branch of statistical model building that uses these data to estimate the β's. Because, in general, polynomial models are linear functions of the unknown β's, we refer to the technique as **linear regression analysis**. We will also see that it is very important to plan the data collection phase of a response surface study carefully. In fact, special types of experimental designs, called **response surface designs**, are valuable in this regard. A substantial part of this book is devoted to response surface designs. Note that analyses and designs need to be carefully matched. If we are planning to analyze data from our planned experiment using a first order model, then the design that we select should be well suited for this analysis. Similarly, if we anticipate curvature similar to what can be modeled with a second-order model, then a different design should be selected.

Good response surface designs have been constructed to perform well based on a particular assumed model, but also have been structured so that they are able to evaluate the assumptions of the model being analyzed to determine if the experimenter's initial impressions of the system under study match the true underlying relationship which produced the data to be analyzed. Hence the experimenter should think carefully about the goals of a particular experiment and what the anticipated analysis will involve before selecting the design for data collection.

1.1.2 The Sequential Nature of RSM

Most applications of RSM are **sequential** in nature. That is, at first some ideas are generated concerning which factors or variables are likely to be important in the response surface study. This usually leads to an experiment designed to investigate these factors with a view toward verifying the role of the factors in influencing the response and eliminating the unimportant ones. This type of experiment is usually called a **screening experiment**. Often at the outset of a response surface study there is a rather long list of variables that could be important in explaining the response. The objective of factor screening is to reduce this list of candidate variables to a relative few so that subsequent experiments will be more efficient and require fewer runs or tests. We refer to a screening experiment as **phase zero** of a response surface study. Since interest in a screening experiment lies in understanding the gross behavior of the system and how factors are related to the response, a first-order model is commonly selected. The class of response surface designs which are used for screening experiments are well suited for gaining understanding about the main effects from different independent variables and comparing their relative contributions to changes in the response values. Since this represents an early stage in the planned sequence of experiments, the goal is to determine which of the factors are more influential on the response while using as small a fraction of the total experimental budget as possible. You should never undertake a response surface analysis until a screening experiment has been performed to identify the important factors.

Once the important independent variables are identified, **phase one** of the response surface study begins. In this phase, the experimenter's objective is to determine where the collected data lie relative to an ideal response. Often, there are two possible outcomes with the current levels or settings of the independent variables resulting in a value of the response that is near the optimum (such as the point B in Fig. 1.1b), or the process is operating in some other region that is (possibly) remote from the optimum (such as the point A in Fig. 1.1b). If the current settings or levels of the independent variables are not consistent with optimum performance, then the experimenter must determine a set of adjustments to the process variables that will move the process toward the optimum. This phase of response surface methodology makes considerable use of the first-order model and an optimization technique called the **method of steepest ascent**. These techniques will be discussed and illustrated in Chapter 5.

Phase two of a response surface study begins when the process is near the optimum. At this point the experimenter usually wants a model that will accurately approximate the true response function within a relatively small region around the optimum. Because the true response surface usually exhibits curvature near the optimum (refer to Fig. 1.1), a second-order model (or very occasionally some higher-order polynomial) will be used. Once an appropriate approximating model has been obtained, this model may be analyzed

to determine the optimum conditions for the process. Chapter 6 will present techniques for the analysis of the second-order model and the determination of optimum conditions.

Response surface designs for modeling the response near the optimum are again selected to match the anticipated analysis. Often, the plan is to characterize the relationship between the response and the key independent variables using the second-order model of the form in Equation 1.8. Designs are constructed to be able to estimate the response for input factor combinations around the expected optimum, where curvature in the relationship is common. Since this stage of experimentation is focused on determining a best set of input values for which the process to operate, a generous portion of the experimental budget is generally reserved for this portion of the process.

A final stage of experimentation, which generally does not require sophisticated response surface designs or a large portion of the experimental budget, is a **confirmatory experiment**. This data collection is generally simple and small, but is designed to confirm that the identified optimum that was obtained in phase two can be achieved by setting the independent variables at the designated settings.

This sequential experimental process is usually performed within some region of the independent variable space called the **operability region**. For the chemical process illustrated in Fig. 1.1, the operability region is $0\text{ hr} < \xi_1 \leq 7\text{ hr}$ and $100°\text{C} \leq \xi_2 \leq 800°\text{C}$. Suppose we are currently operating at the levels $\xi_1 = 2.5$ hr and $\xi_2 = 500°$C, shown as point A in Fig. 1.6. Now it is unlikely that we would want to explore the entire region of operability with a single experiment. Instead, we usually define a smaller **region of interest** or **region of experimentation** around the point A within the larger region of operability. Typically, this region of experimentation is either a cuboidal region, as shown around the point A in Fig. 1.6, or a spherical region, as shown around point B. The choice of response surface design matches the specified region of experimentation.

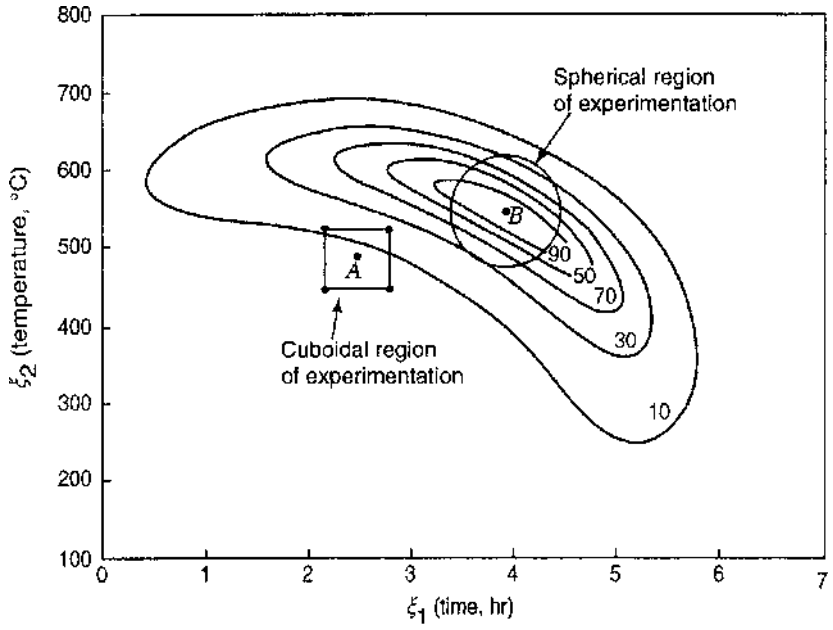

Figure 1.6 The region of operability and the region of experimentation.

The sequential nature of response surface methodology allows the experimenter to learn about the process or system under study as the investigation proceeds. This ensures that over the course of the RSM application the experimenter will learn the answers to questions such as (1) the location of the region of the optimum, (2) the type of approximating function required, (3) the proper choice of experimental designs, (4) how much replication is necessary, and (5) whether or not transformations on the responses or any of the process variables are required. Because the nature of a response surface study has multiple stages with different goals, there are several improtant aspects that need to be managed throughout the process. First, many studies have budget constraints that will dicate how much and what data can be collected. It is important to plan for all of the stages of the study and to allow for adequate resources to be available to effectively answer the important questions in each phase.

Second, since the knowledge gained in early phases of the study help to determine what subsequent experiments will study, it is important to plan how the different phases will connect to each other, and what information can be leveraged from early phases. Thirdly, the selection of a model for the analysis of the data from each phase is based on current understanding of the underlying process. It is important to think of the sequence of experiments as a mechanism for not having to make too many assumptions at any stage, Running a large complicated experiment that has many untested assumptions can lead to costly errors and wasting of resources. Hence, a series of smaller experiments can verify some assumptions early in the sequence and can allow the experimenter to proceed in later stages with greater confidence.

Lastly, we again mention the connection between the choice of experiment and the planned analysis. Before jumping in to collect data, the goals of each phase should be clearly defined, and the nature of the response surface design selected should reflect the goals and the planned analysis. Since there are often surprises when collecting and analyzing data, it is helpful to consider what could go wrong with the experiment and to have a plan for how to deal with some of these surprises. A substantial portion of this book—Chapters 3, 4, 8, and 9—is devoted to designed experiments useful in RSM.

1.1.3 Objectives and Typical Applications of RSM

Response surface methodology is useful in the solution of many types of industrial problems. Generally, these problems fall into three categories:

1. *Mapping a Response Surface over a Particular Region of Interest*. Consider the chemical process in Fig. 1.1b. Normally, this process would operate at a particular setting of reaction time and reaction temperature. However, some changes to these normal operating levels might occasionally be necessary, perhaps to produce a product that meets other specific customer requirements. If the true unknown response function has been approximated over a region around the current operating conditions with a suitable fitted response surface (say a second-order surface), then the process engineer can predict in advance the changes in yield that will result from any readjustments to the input variables, namely, time and temperature.
2. *Optimization of the Response*. In the industrial world, a very important problem is determining the conditions that optimize the process. In the chemical process of Fig. 1.1b, this implies determining the levels of time and temperature that result in maximum yield. An RSM study that began near point *A* in Fig. 1.1b would eventually

lead the experimenter to the region near point *B*. A second-order model could then be used to approximate the yield response in a narrow region around point *B*, and from examination of this approximating response surface the optimum levels or condition for time and temperature could be chosen.

3. *Selection of Operating Conditions to Achieve Specifications or Customer Requirements*. In most response surface problems there are several responses that must be simultaneously considered. For example, in the chemical process of Fig. 1.1, suppose that in addition to yield, there are two other responses: cost and concentration. We would like to maintain yield above 70%, while simultaneously keeping the cost below \$34/pound; however, the customer has imposed specifications for concentration such that this important physical property must be 65±3 g/liter.

One way that we could solve this problem is to obtain response surfaces for all three responses—yield, cost, and concentration—and then superimpose the contours for these responses in the time–temperature plane, as illustrated in Fig. 1.7. In this figure we have shown the contours for yield = 70%, cost = \$34/pound, concentration = 62 g/liter, and concentration = 68 g/liter. The unshaded region in this figure represents the region containing operating conditions that simultaneously satisfy all requirements on the process.

In practice, complex process optimization problems such as this can often be solved by superimposing appropriate response surface contours. However, it is not unusual to encounter problems with more than two process variables and more complex response requirements to satisfy. In such problems, other optimization methods that are more effective than overlaying contour plots will be necessary, and can often not only identify a region which satisfies the minimal customer requirements, but also find an optimal combination of input variables to achieve ideal performance. We will discuss methodology for solving these types of problems in Chapter 7.

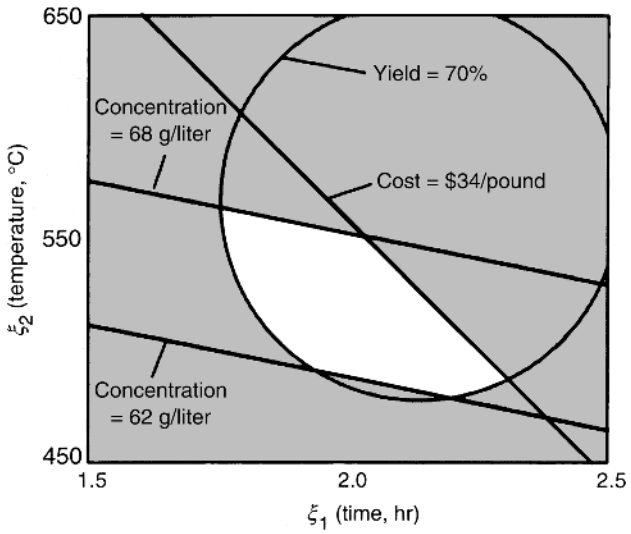

Figure 1.7 The unshaded region showing the conditions for which yield ≥70%, cost ≤\$34/pound, and 62 g/liter ≤ concentration ≤ 68 g/liter.

1.1.4 RSM and the Philosophy of Quality Improvement

During the last few decades, industrial organizations in the United States and Europe have become keenly interested in quality and process improvement. Statistical methods, including statistical process control (SPC) and design of experiments, play a key role in this activity. Quality improvement is most effective when it occurs early in the product and process development cycle. It is very difficult, expensive, and inefficient to manufacture a poorly designed product. Industries such as semiconductors and electronics, aerospace, automotive, biotechnology and pharmaceuticals, medical devices, chemical, and process industries are all examples where experimental design methodology has resulted in shorter design and development time for new products, as well as products that are easier to manufacture, have higher reliability, have enhanced field performance, and meet or exceed customer requirements.

RSM is an important branch of experimental design in this regard. RSM is a critical technology in developing new processes, optimizing their performance, and improving the design and/or formulation of new products. It is often an important **concurrent engineering tool**, in that product design, process development, quality, manufacturing engineering, and operations personnel often work together in a team environment to apply RSM. The objectives of quality improvement, including reduction of variability and improved product and process performance, can often be accomplished directly using RSM.

1.2 PRODUCT DESIGN AND FORMULATION (MIXTURE PROBLEMS)

Many product design and development activities involve formulation problems, in which two or more ingredients are mixed together. For example, suppose we are developing a new household cleaning product. This product is formulated by mixing several chemical surfactants together. The product engineer or scientist would like to find an appropriate blend of the ingredients so that the grease-cutting capability of the cleaner is good, and so that it generates an appropriate level of foam when in use. In this situation the response variables—namely, grease-cutting ability and amount of foam—depend on the percentages or proportions of the individual chemical surfactants (the ingredients) that are present in the product formulation.

There are many industrial problems where the response variables of interest in the product are a function of the proportions of the different ingredients used in its formulation. This is a special type of response surface problem called a **mixture problem**.

While we traditionally think of mixture problems in the product design or formulation environment, they occur in many other settings. Consider plasma etching of silicon wafers, a common manufacturing process in the semiconductor industry. Etching is usually accomplished by introducing a blend of gases inside a chamber containing the wafers. The measured responses include the etch rate, the uniformity of the etch, and the selectivity (a measure of the relative etch rates of the different materials on the wafer). All of these responses are a function of the proportions of the different ingredients blended together in the etching chamber.

There are special response surface design techniques and model-building methods for mixture problems. These techniques are discussed in Chapters 12 and 13.

1.3 ROBUST DESIGN AND PROCESS ROBUSTNESS STUDIES

It is well known that variation in key performance characteristics can result in poor product and process quality. During the 1980s, considerable attention was given to this problem, and methodology was developed for using experimental design, specifically for the following:

1. For designing products or processes so that they are robust to environment conditions.
2. For designing or developing products so that they are robust to component variation.
3. For minimizing variability in the output response of a product around a target value.

By **robust**, we mean that the product or process performs consistently on target and is relatively insensitive to factors that are difficult to control.

Professor Genichi Taguchi used the term **robust parameter design** (or RPD) to describe his approach to this important class of industrial problems. Essentially, robust parameter design methodology strives to reduce product or process variation by choosing levels of controllable factors (or parameters) that make the system insensitive (or robust) to changes in a set of uncontrollable factors that represent most of the sources of variability. Taguchi referred to these uncontrollable factors as **noise factors**. These are the environmental factors such as humidity levels, changes in raw material properties, how the customer will use the product, product aging, and component variability referred to in 1 and 2 above. We usually assume that these noise factors are uncontrollable in the field, but can be controlled during product or process development for purposes of a designed experiment.

Considerable attention has been focused on the methodology advocated by Taguchi, and a number of flaws in his approach have been discovered. However, there are many useful concepts in his philosophy, and it is relatively easy to incorporate these within the framework of response surface methodology. In Chapter 11 we will present the response surface approach to robust design and process robustness studies.

1.4 USEFUL REFERENCES ON RSM

The origin of RSM is the seminal paper by Box and Wilson (1951). They also describe the application of RSM to chemical processes. This paper had a profound impact on industrial applications of experimental design, and was the motivation of much of the research in the field. Many of the key research and applications papers are cited in this book.

There have also been five review papers published on RSM: Hill and Hunter (1966), Mead and Pike (1975), Myers et al. (1989), Myers et al. (2004) and Anderson-Cook et al. (2009a). The paper by Myers (1999) on future directions in RSM offers a view of research needs in the field. There are also two other full-length books on the subject: Box and Draper (1987) and Khuri and Cornell (1996). A second edition of the Box and Draper book was published in 2007 with a slightly different title [Box and Draper (2007)]. An edited volume by Khuri (2006) considers some specialized RSM topics. The monograph by Myers (1976) was the first book devoted exclusively to RSM.