CHAPTER 3 *Response surface methodology*

3.1 Introduction

Response surface methodology (RSM) is a collection of mathematical and statistical techniques for empirical model building. By careful design of *experiments*, the objective is to optimize a *response* (output variable) which is influenced by several *independent variables* (input variables). An experiment is a series of tests, called *runs*, in which changes are made in the input variables in order to identify the reasons for changes in the output response.

Originally, RSM was developed to model experimental responses (Box and Draper, 1987), and then migrated into the modelling of numerical experiments. The difference is in the type of error generated by the response. In physical experiments, inaccuracy can be due, for example, to measurement errors while, in computer experiments, numerical noise is a result of incomplete convergence of iterative processes, round-off errors or the discrete representation of continuous physical phenomena (Giunta et al., 1996; van Campen et al., 1990, Toropov et al., 1996). In RSM, the errors are assumed to be random.

The application of RSM to design optimization is aimed at reducing the cost of expensive analysis methods (e.g. finite element method or CFD analysis) and their associated numerical noise. The problem can be approximated as described in Chapter 2 with smooth functions that improve the convergence of the optimization process because they reduce the effects of noise and they allow for the use of derivative-based algorithms. Venter et al. (1996) have discussed the advantages of using RSM for design optimization applications.

For example, in the case of the optimization of the calcination of Roman cement described in Section 6.3, the engineer wants to find the levels of temperature (x_1) and time (x_2) that maximize the early age strength (y) of the cement. The early age strength is a function of the levels of temperature and time, as follows:

$$y = f(x_1, x_2) + \varepsilon \tag{3.1}$$

where ε represents the noise or error observed in the response y. The surface represented by $f(x_1, x_2)$ is called a *response surface*.

The response can be represented graphically, either in the three-dimensional space or as *contour plots* that help visualize the shape of the response surface. Contours are curves of constant response drawn in the x_i, x_j plane keeping all other variables fixed. Each contour corresponds to a particular height of the response surface, as shown in Figure 3.1.

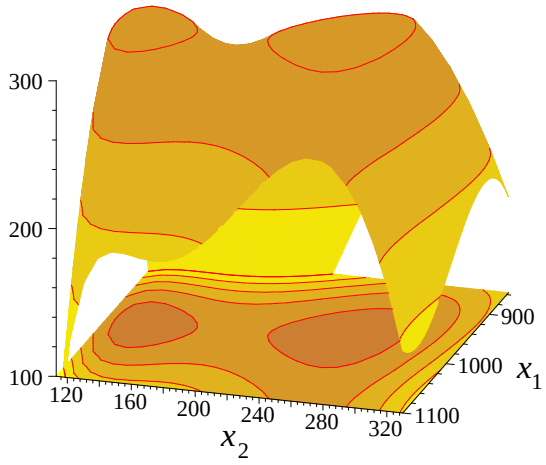

Figure 3.1 Three-dimensional response surface and the corresponding contour plot for the early age strength of Roman cement where x_1 is the calcination temperature (°C) and x_2 is the residence time (mins).

This chapter reviews the two basic concepts in RSM, first the choice of the approximate model and, second, the plan of experiments where the response has to be evaluated.

3.2 Approximate model function

Generally, the structure of the relationship between the response and the independent variables is unknown. The first step in RSM is to find a suitable approximation to the true relationship. The most common forms are low-order polynomials (first or second-order).

In this thesis a new approach using genetic programming is suggested. The advantage is that the structure of the approximation is not assumed in advance, but is given as part of the solution, thus leading to a function structure of the best possible quality. In addition, the complexity of the function is not limited to a polynomial but can be generalised with the inclusion of any mathematical operator (e.g.

trigonometric functions), depending on the engineering understanding of the problem. The regression coefficients included in the approximation model are called the *tuning parameters* and are estimated by minimizing the sum of squares of the errors (Box and Draper, 1987):

$$G(\boldsymbol{a}) \;=\; \sum_{p=1}^{P} \left\{ w_p \left(F_p - \tilde{F}_p(\boldsymbol{a}) \right)^2 \right\} \;\rightarrow\; \min \tag{3.2}$$

where w_p is a weight coefficient that characterizes the relative contribution of the information of the original function at the point p, p=1,...,P.

The construction of response surface models is an iterative process. Once an approximate model is obtained, the goodness-of-fit determines if the solution is satisfactory. If this is not the case, the approximation process is restarted and further experiments are made or the GP model is evolved with different parameters, as explained in Chapter 4.

To reduce the number of analyses in computer simulations, sensitivity data may be used in the model fitting, although this information is not always available at low cost. If in addition to the values of the original function $F_p = F(x_p)$ their first order derivatives at point p $F_{p,i} = \dfrac{\partial}{\partial x_i} F_p$ (i=1,…,N, p=1,…,P) are known, the problem (3.2) is replaced by the following one (Toropov et al., 1993):

$$G(\boldsymbol{a}) = \sum_{p=1}^{P} \left\{ w_p \left(\left(F_p - \tilde{F}_p(\boldsymbol{a}) \right)^2 + \gamma \frac{\sum_{i=1}^{N} \left(F_{p,i} - \tilde{F}_p(\boldsymbol{a})_{,i} \right)^2}{\sum_{i=1}^{N} {F_{p,i}}^2} \right) \right\} \rightarrow \min \quad (3.3)$$

where $\gamma > 0$ is the parameter characterizing a degree of inequality of the contribution of the response and the sensitivity data. In this thesis, γ is taken as 0.5, following recommendations by Toropov et al. (1993).

Van Keulen et al. (2000) have presented a methodology for the construction of responses using both function values and derivatives on a weighted least-squares formulation. The authors conclude that the use of derivatives provides better accuracy and requires a reduced number of data.

3.3 Design of experiments

An important aspect of RSM is the *design of experiments* (Box and Draper, 1987), usually abbreviated as DoE. These strategies were originally developed for the model fitting of physical experiments, but can also be applied to numerical experiments. The objective of DoE is the selection of the points where the response should be evaluated.

Most of the criteria for optimal design of experiments are associated with the mathematical model of the process. Generally, these mathematical models are polynomials with an unknown structure, so the corresponding experiments are designed only for every particular problem. The choice of the design of experiments

can have a large influence on the accuracy of the approximation and the cost of constructing the response surface.

In a traditional DoE, *screening experiments* are performed in the early stages of the process, when it is likely that many of the design variables initially considered have little or no effect on the response. The purpose is to identify the design variables that have large effects for further investigation. Genetic Programming has shown good screening properties (Gilbert et al., 1998), as will be demonstrated in Section 6.2, which suggests that both the selection of the relevant design variables and the identification of the model can be carried out at the same time.

A detailed description of the design of experiments theory can be found in Box and Draper (1987), Myers and Montgomery (1995) and Montgomery (1997), among many others. Schoofs (1987) has reviewed the application of experimental design to structural optimization, Unal et al. (1996) discussed the use of several designs for response surface methodology and multidisciplinary design optimization and Simpson et al. (1997) presented a complete review of the use of statistics in design.

As introduced in Section 3.1, a particular combination of runs defines an *experimental design*. The possible settings of each independent variable in the *N*-dimensional space are called *levels*. A comparison of different methodologies is given in the next section.

3.3.1 Full factorial design

To construct an approximation model that can capture interactions between *N* design variables, a full factorial approach (Montgomery, 1997) may be necessary to

investigate all possible combinations. A *factorial* experiment is an experimental strategy in which design variables are varied together, instead of one at a time.

The lower and upper bounds of each of *N* design variables in the optimization problem needs to be defined. The allowable range is then discretized at different levels. If each of the variables is defined at only the lower and upper bounds (two levels), the experimental design is called 2^N full factorial. Similarly, if the midpoints are included, the design is called 3^N full factorial and shown in Figure 3.2.

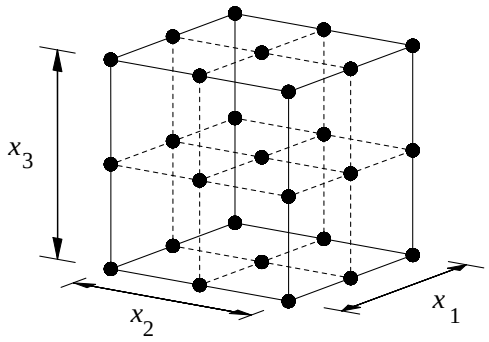

Figure 3.2 A 3^3 full factorial design (27 points)

Factorial designs can be used for fitting second-order models. A second-order model can significantly improve the optimization process when a first-order model suffers lack of fit due to interaction between variables and surface curvature. A general second-order model is defined as

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{\substack{i=1 \\ i<j}}^{n} a_{ij} x_i x_j \qquad (3.4)$$

where x_i and x_j are the design variables and a are the tuning parameters.

The construction of a quadratic response surface model in *N* variables requires the study at three levels so that the tuning parameters can be estimated. Therefore, at

least (N+1) (N+2) / 2 function evaluations are necessary. Generally, for a large number of variables, the number of experiments grows exponentially (3^N for a full factorial) and becomes impractical. A full factorial design typically is used for five or fewer variables.

If the number of design variables becomes large, a fraction of a full factorial design can be used at the cost of estimating only a few combinations between variables. This is called *fractional factorial design* and is usually used for screening important design variables. For a 3^N factorial design, a $\left(\frac{1}{3}\right)^p$ fraction can be constructed, resulting in 3^{N-p} points. For example, for p=1 in a 3^3 design, the result is a one-third fraction, often called 3^{3-1} design, as shown in Figure 3.3 (Montgomery, 1997).

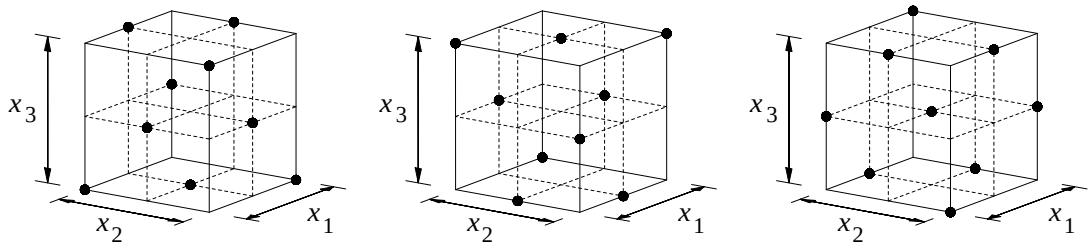

Figure 3.3 Three one-third fractions of the 3^3 design

3.3.2 Central composite design

A second-order model can be constructed efficiently with central composite designs (CCD) (Montgomery, 1997). CCD are first-order (2^N) designs augmented by additional centre and axial points to allow estimation of the tuning parameters of a second-order model. Figure 3.4 shows a CCD for 3 design variables.

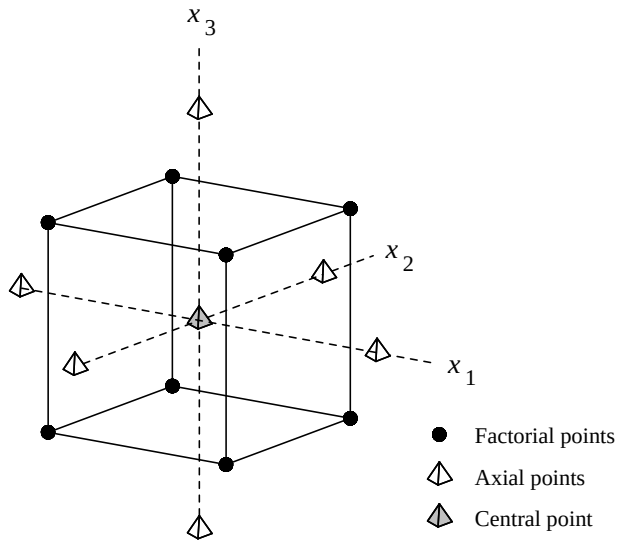

Figure 3.4 Central composite design for 3 design variables at 2 levels

In Figure 3.4, the design involves 2^N factorial points, $2N$ axial points and 1 central point. CCD presents an alternative to 3^N designs in the construction of second-order models because the number of experiments is reduced as compared to a full factorial design (15 in the case of CCD compared to 27 for a full-factorial design). CCD have been used by Eschenauer and Mistree (1997) for the multiobjective design of a flywheel.

In the case of problems with a large number of designs variables, the experiments may be time-consuming even with the use of CCD.

3.3.3 D-optimal designs

The D-optimality criterion enables a more efficient construction of a quadratic model (Myers and Montgomery, 1995). The objective is to select P design points from a larger set of candidate points.

Equation (3.4) can be expressed in matrix notation as:

$$Y = X * B + e \tag{3.5}$$

where Y is a vector of observations, e is a vector of errors, X is the matrix of the values of the design variables at plan points and B is the vector of tuning parameters. B can be estimated using the least-squares method as:

$$B = (X^T * X)^{-1} X^T Y \tag{3.6}$$

The D-optimality criterion states that the best set of points in the experiment maximizes the determinant $|X^T X|$. "D" stands for the determinant of the $X^T X$ matrix associated with the model. From a statistical point of view, a D-optimal design leads to response surface models for which the maximum variance of the predicted responses is minimized. This means that the points of the experiment will minimize the error in the estimated coefficients of the response model.

The advantages of this method are the possibility to use irregular shapes and the possibility to include extra design points. Generally, D-optimality is one of the most used criteria in computer-generated design of experiments.

Several applications are described in Giunta et al. (1996) for the wing design of a high-speed civil transport and Unal et. al. (1996) for a multidisciplinary design optimization study of a launch vehicle. Haftka and Scott (1996) have reviewed the use of D-optimality criteria for the optimization of experimental designs.

3.3.4 Taguchi's contribution to experimental design

Taguchi's methods (Montgomery, 1997) study the parameter space based on the fractional factorial arrays from DoE, called *orthogonal arrays*. Taguchi argues that it

is not necessary to consider the interaction between two design variables explicitly, so he developed a system of tabulated designs which reduce the number of experiments as compared to a full factorial design. An advantage is the ability to handle discrete variables. A disadvantage is that Taguchi ignores parameter interactions.

3.3.5 Latin hypercube design

Latin hypercube design (McKay et al., 1979) can be viewed as an *N*-dimensional extension of the traditional Latin square design (Montgomery, 1997). On each level of every design variable only one point is placed. There are the same number of levels as runs and the levels are assigned randomly to runs.

This method ensures that every variable is represented, no matter if the response is dominated by only a few ones. Another advantage is that the number of points to be analyzed can be directly defined.

An example of the use of such plans can be found in Schoofs et al. (1997).

3.3.6 Audze-Eglais' approach

Audze and Eglais (1977) suggested a non-traditional criterion for elaboration of plans of experiments which, similar to the Latin hypercube design, is not dependent on the mathematical model of the problem under consideration. The input data for the elaboration of the plan only include the number of factors *N* (number of design variables) and the number of experiments *K*. The main principles in this approach are as follows:

1. The number of levels of factors (same for each factor) is equal to the number of experiments and for each level there is only one experiment. This is similar to the Latin hypercube design.

2. The points of experiments are distributed as uniformly as possible in the domain of variables. There is a physical analogy with the minimum of potential energy of repulsive forces for a set of points of unit mass, if the magnitude of these repulsive forces is inversely proportional to the distance squared between the points:

$$\sum_{p=1}^{P} \sum_{q=p+1}^{P} \frac{1}{L_{pq}^{\,2}} \quad \rightarrow \quad \min \tag{3.7}$$

where L_{pq} is the distance between the points having numbers p and q ($p \neq q$).

The elaboration of the plans is time consuming, so each plan of experiment is elaborated only once and stored in a matrix characterized by the levels of factors for each of P experiments. For example, for a number of factors (design variables) $N = 2$ and $P = 10$, the matrix is

$$\begin{vmatrix} 8 & 10 & 4 & 6 & 2 & 3 & 9 & 5 & 7 & 1 \\ 1 & 7 & 10 & 6 & 8 & 5 & 4 & 2 & 9 & 3 \end{vmatrix} \tag{3.8}$$

The plan (3.8) is represented in Figure 3.5 and compared with a CCD for two design variables with 9 runs.

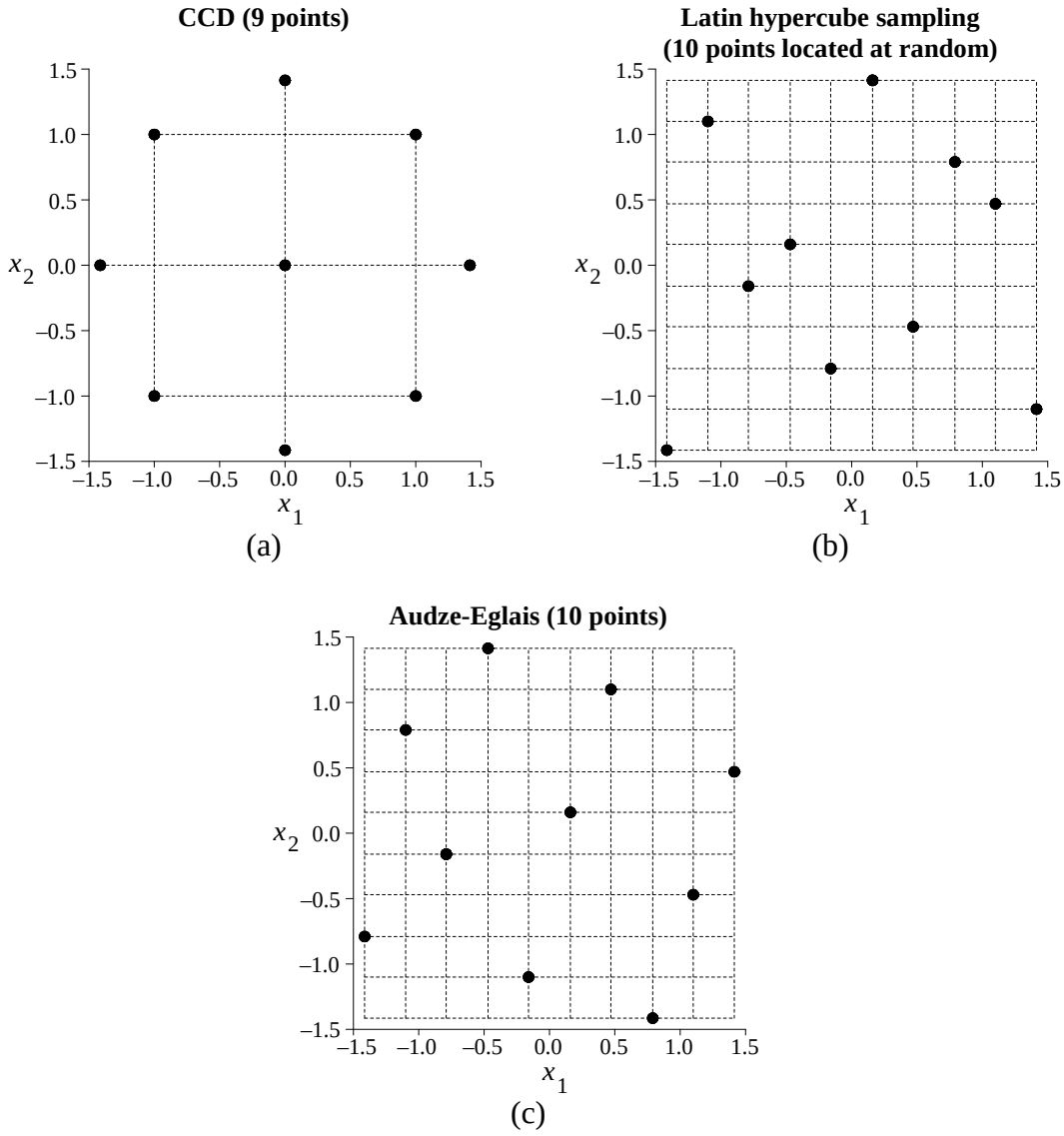

Figure 3.5 Comparison between CCD (a), Latin hypercube design (b) and Audze-Eglais design (c)

The advantages of this method are the space-filling property as shown in Figure 3.5 and the presentation of the data as tabulated designs. A disadvantage is that once a design has been defined, no extra points can be added to the initial set. This approach has been used by Rikards (1993) to design composite materials with predicted properties (weight, price, etc.).

3.3.7 Van Keulen's approach

In the course of an iterative optimization process modelled by approximations, new points must be generated in specified domains of the design variable space. A new scheme for the design of experiments (Van Keulen and Toropov, 1999) has been formulated with the following characteristics:

1. The scheme works efficiently even if only a single additional design point is generated to the existing plan. For a number of new design points, the algorithm is used several times.
2. The scheme remains effective if different types of functions are used within the same optimization task to approximate the objective function and the constraints.

The approach distributes points as homogeneously as possible in the sub-domains of interest. This is done by the introduction of the following cost function:

$$Q = \sum_{p=1}^{P} \frac{n^2}{\left\| \bar{\boldsymbol{x}}_p - \bar{\boldsymbol{d}} \right\|} + \sum_{p=1}^{P} \sum_{i=1}^{n} \frac{1}{\left([\bar{x}_i]_p - \bar{d}_i \right)^2} + \sum_{i=1}^{n} \frac{1}{\left(2\,\bar{d}_i \right)^2} + \\ + \sum_{i=1}^{n} \frac{1}{\left(2 - 2\bar{d}_i \right)^2} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{1}{\left(\bar{d}_i - \bar{d}_j \right)^2} \tag{3.9}$$

which is minimized with respect to the location of the new point *d*. Symbols denoted $\bar{\ldots}$ refer to coordinates which are normalized in the sub-domain of interest. The first term in the expression attempts to maximize the distance between points, and the second term promotes a homogeneous distribution along the coordinate axes. The third and fourth terms ensure that points do not belong to the boundary of the sub-

domain. The last term prevents points from aligning along the diagonal of the search sub-region when only a few points are available.

3.4 Conclusion

The response surface methodology analysis has been reviewed. RSM can be used for the approximation of both experimental and numerical responses. Two steps are necessary, the definition of an approximation function and the design of the plan of experiments. As concluded in Chapter 2, genetic programming is the method of choice to find a suitable approximation function and will be described in Chapter 4.

A review of different designs for fitting response surfaces has been given. A desirable design of experiments should provide a distribution of points throughout the region of interest, which means to provide as much information as possible on the problem. This "space-filling" property is a characteristic of three plans: Latin hypercube sampling, Audze-Eglais and van Keulen. All three plans are independent of the mathematical model of the approximation. However, Latin hypercube sampling distributes the points randomly in the space, while Audze-Eglais uses a distribution based on maximum separation between points. The Audze-Eglais plan has been chosen in this thesis.

It should be noted that if the model building is to be repeated within an iterative scheme (e.g. with mid-range approximations), van Keulen’s plan would become an attractive alternative as it adds points to an existing plan. This thesis is primarily focused on building global approximations.