Employing Response Surface Methodologies for Design Optimization

Tech Brief 150901 DE

Integrated Systems Research, Inc.

September, 2015

steve.carmichael@isrtechnical.com

Abstract:

This Tech Brief discusses the use of Response Surface Methodologies (RSM) in the process of optimizing design performance. Identifying design points where performance is maximized and variation in system response minimized is central to implementing successful designs. This paper discusses a method for evaluating the significance of parameter estimations in generating response surface models and how to use them in the decision making process.

Background:

Typically, Response Surface Methodologies (RSM) are used in building models from empirical data. The power of simulation, however, is in the ability to create virtual prototypes of systems and evaluate their performance. This provides another source for RSM other than empirical data. The simulations are typically executed at much lower costs and in less time than actual prototypes. Additionally, the evaluations can be done at the beginning of a design process where decisions will have the largest impact on project, design and production costs.

Finite element models are often used to analytically simulate the performance of options in the design space. The response surfaces generated from the simulations can provide vital information in developing paths that will minimize costs and maximize performance. The limitations, however, of response surface models need to be appreciated in order to properly use them in the decision making process.

Estimating the goodness associated with the parameters defining the response surface is important in understanding the value of a model. Generating the surface or hyper surface from response data is referred to as a regression analysis. In the case of using results from deterministic models, the variability associated with the regression is due to the ability of the imposed model to explain the results not due to random variation within the data.

When reducing empirical data, the variation seen in the reduction is due to limitations inherent in both the imposed regression model and measurements associated with the data. With data obtained from analytical models the variation is associated only with the chosen regression model.

Using Response Surface Methodologies:

The approach typically taken with the Response Surface Methodology (RSM) is to first determine through a screening process the independent variables which have the greatest influence on the system's response. This allows the engineer to either develop the most efficient full or partial factorial design which will capture the main and the cross coupling effects of the variables that drive performance. Typically, if a variable doesn't have a large main effect it seldom has a significant cross coupling with another variable and hence is not included in the design space under evaluation. Depending on the specific partial factorial design some of the cross coupling effects will be aliased but not necessarily all of them.

The second step is to determine the extent of the design space where a 2^{nd} order polynomial will provide a good representation of the system's response. The idea behind this approach is that virtually any system response can be well characterized by a second order function as long as the domain of the analysis is sufficiently small. The typical method for determining the goodness of the fit is to use the regression coefficient as the primary metric. Using this metric alone, however, can lead to an incorrect conclusion regarding the utility of the characterization in evaluating robustness of a design to system variation. This topic is taken up in the section entitled *Checking Model Adequacy*.

An important feature to remember is that even when an adequate 2^{nd} order polynomial has been identified it does not mean that extrapolation can be employed with any confidence. The primary reason for this is that the polynomial is very likely not the model that best corresponds to the actual physical phenomenon. Even though the polynomial can provide an excellent explanation of the variation seen in the design space it does not necessarily capture the causality associated with it. Correlation is necessary but is not sufficient for causality. The regression can simply provide a good fit within the chosen design space without providing the best relational model for the data.

Once an adequate model has been created, the response surface can be used as a means of identifying local maxima or minima at which to operate. Taking partial derivatives around the operating point enables the engineer to assess the tolerance of the design to variation in operating conditions. Additionally, if the surface is a ridge the engineer can also identify the direction to take the design space to further optimize desired system characteristics.

This brief provides a couple of simple examples to illustrate these RSM concepts and techniques.

Checking Model Adequacy:

Building numerical models involves fitting an equation to an overdetermined population of data. (e.g., more degrees of freedom in the data than are in the model). The technique employed in fitting an overdetermined data set is typically some variation of a least squares approach. For a given model, the least squares approach provides a result that minimizes the error between the fit and the actual data points. It's a minimization method that mimics what is observed in physical systems.

The difference between the model and actual data points (e.g., residuals) will sum to zero. The variance between the data and model is captured by looking at the sum of the square of the difference. In general, the smaller the variance the greater explanatory power of the model. The explanatory value is accounting for variation between the model and the data not causal relationships between the variables. The regression coefficient is typically the parameter used to evaluate goodness of fit.

As degrees of freedom are added to the model, the tendency is for the regression coefficient to approach unity (e.g., maximum explanatory power). The variance, however, can actually increase with additional terms and is masked by an increase in the regression coefficient. An adjustment to the regression coefficient is often employed to determine whether or not the additional terms add value to the model.

Figure 1 provides a simple example of this process.

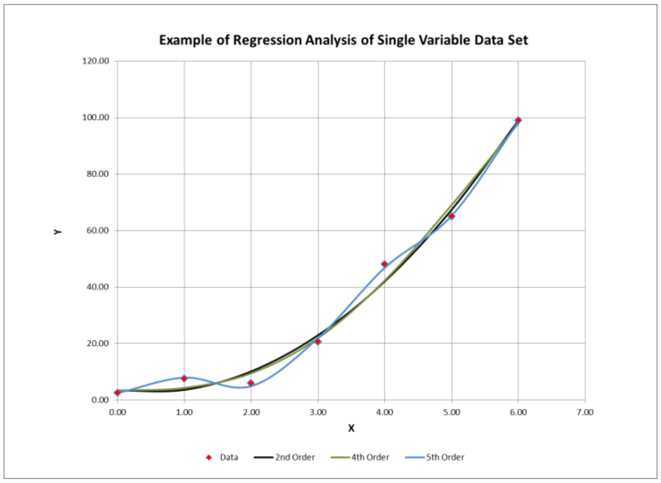

Figure 1

The least squares process creates a square matrix from the overdetermined independent variable matrix [X], by pre-multiplying each side of the equation by the transpose of that matrix $[X]^T$.

$$[X]^T\{Y\} = [X]^T n[\overset{m}{X}]\{\beta\}$$

Equation 1.0

The system is over determined (e.g., n > m). The resulting independent variable square matrix is inverted and both sides of the equation are multiplied by it to obtain the coefficient vector β for the regression model.

The regression coefficient is obtained from the ratio of the regression sum of the squares and the total sum of squares.

The regression sum of squares is

$$SS_R = \{\beta\}^T[X]^T\{Y\} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Equation 2.0

The total sum of the squares is

$$S_{yy} = \{Y\}^T\{Y\} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Equation 3.0

The regression coefficient is

$$R^2 = \frac{SS_R}{S_{yy}}$$

Equation 4.0

As stated before, the regression coefficient will trend towards unity as the number of terms in the independent variable matrix increases. The actual benefit, however, provided by additional terms can be non-significant and potentially create noise when evaluating the sensitivity of the system to input variations at a given operating point. Applying Occam's razor[1] is recommended when determining the best model to use in evaluating a system.

The adjusted regression coefficient is a means of determining whether or not additional terms are being added beyond what is necessary. The ratio of the number of degrees of freedom in the data to the remainder left from the model is used to scale the regression coefficient. The adjusted regression coefficient will decrease when additional terms in the model are unnecessary.

$$R_{adj}{}^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right)$$

Equation 5.0

In Figure 1, the R^2 for a second degree polynomial is 0.9899 while for a fourth degree fit it is 0.9907. At first blush it appears that adding the additional terms associated with the fourth degree polynomial improves the model. If the adjusted regression coefficient is used, however, the coefficient becomes lower with the added terms (0.985 compared to 0.972) indicating they are unnecessary.

The adjusted regression coefficient for the fifth degree polynomial does increase, but it only leaves a single unconstrained degree of freedom (n-p). Since the adjusted regression coefficient for the fourth degree polynomial is lower than the second, it would be advisable to either employ the second degree model for evaluating the data or obtain more data in the domain to verify whether or not the variation that appears to be explained by the higher order polynomial actually exists.

Using RSM at Decision Points:

Figure 2 provides a plot of a stepped bar used to evaluate the stress concentration at the shoulder of the step as a function of fillet radius and the step ratio. The bar is loaded axially with a stress ratio R (not to be confused with the regression coefficient R^2) of -1. The nominal stress is 10 ksi.

The stepped bar provides results that are intuitively obvious but illustrates the process of using a response surface to evaluate the performance and robustness of a component's performance to design and/or manufacturing variations.

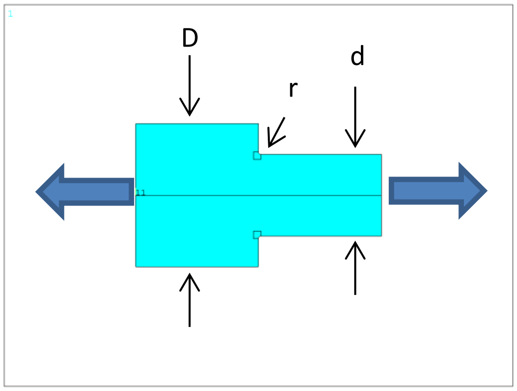

Figure 2

The stress concentration (K$_t$) response surface for the step bar is provided in Figure 3. The surface is fitted with a 2nd polynomial which includes the cross coupling term. The adjusted R^2 for the regression is 0.9995.

$$K_t = 34.56\beta_1^2 - 0.224\beta_2^2 - 9.005\beta_1 + 1.679\beta_2 - 3.52\beta_1\beta_2 + 1.341$$

Equation 6.0

β_1 is the fillet ratio (r/d) and β_2 is the step ratio (D/d).

The cross coupling term for the fillet and step ratios has significance in terms of goodness of fit. When the cross coupling term is eliminated the adjusted R^2=0.9966. This is lower than the adjusted R^2 for the polynomial that incorporates all the terms. The cross coupling term, therefore, should be kept in the

[1] Entities must not be multiplied beyond necessity

response surface. This term has a relatively large influence on the tolerance evaluation as a function of the step ratio.

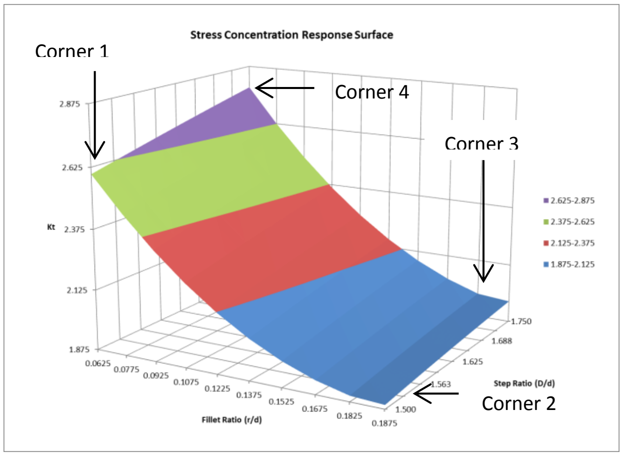

Figure 3

The cost and geometry constraints could be overlaid or projected onto the surface to identify the viable commercial region that exists for the design. Using partial derivatives of the numerical model, the robustness of the stepped bar in terms of fatigue life can be evaluated in that region.

Since the fatigue life is a very strong power function it would be difficult to obtain a good curve fit over any sizable design space with a 2nd order polynomial. Having this *a priori* knowledge enables the engineer to generate a response surface using another variable other than fatigue life (e.g., stress concentration) which can be curve fit relatively well with a polynomial. Then employing a secondary calculation (e.g., Basquin's equation) applied to the concentrated stress, evaluation of the component's robustness in terms of fatigue life can be evaluated at a given design point.

In this example, absent cost and geometry constraints, the corner points of the response are where the local maxima and minima of the design space exist. Employing the Basquin equation for the life evaluation, Tables 1 and 2 summarize the estimated percent change in fatigue life for a percent change in a given design variable.

The percent change in life as a function of the two design variables, fillet radius and step ratio, were obtained by taking partial derivatives at the corners of the response surface. The change in stress concentrations were then used to compute the change in life predictions.

Table 1
Percent Change in Life as Percent Change in (r/d)

Corner	Low	High
1	4%	-3%
2	2%	-2%
3	2%	-1%
4	4%	-3%

Table 2
Percent Change in Life as Percent Change in (D/d)

Corner	Low	High
1	-2%	2%
2	-1%	1%
3	-1%	2%
4	-2%	2%

Table 3 provides the estimated -3σ fatigue life predictions at the corner points.

Table 3
Fatigue Life Predictions

Corner	Cycles
1	3.01E+05
2	2.93E+06
3	2.18E+06
4	1.82E+05

Three key performance characteristics are easily identified from these tables. First, as expected, the maximum life in the design space occurs where the fillet radius is a maximum and the step ratio is a minimum.

Secondly, the design has its greatest tolerance to manufacturing variations at the same design point as where the life is maximized. Operationally this is an optimal point in the design from both a yield perspective as well as tolerance to manufacturing variations. Thirdly, the tolerance evaluation indicates that the variation in the fillet radius is of greater significance than the step ratio.

Attempting to extrapolate outside the design space, results in significant error. For example, when extrapolating with the response surface function, the K_t value is predicted to be 3.152 with a ratio (r/d) of 0.031 at a (D/d) of 1.75. The actual value is 3.582.

The extrapolated K_t value results in a life prediction that is 2.5 times greater than when the correct stress concentration value is used. Although the polynomial provides an excellent fit of the design space, the ability to extrapolate with it does not exist. This is because the function that actually relates the stress concentration to the r/d ratio is a power relationship. The 2nd degree polynomial does not capture that relationship outside the design space.

Summary:

Response surfaces are useful tools at design decision points in that they enable the engineer to both evaluate the expected performance and the tolerance of the design to input and manufacturing variations.

When costs and geometry/manufacturing constraints are mapped onto the surface the commercial design space is readily identified and can be evaluated. This is extremely helpful in determining if the costs associated with relaxing given restraints are both functionally and commercially warranted.

The adequacy of a model can be evaluated with the adjusted regression coefficient as a function of terms used in the polynomial. Terms should only be used in the equation if they actually reduce the variance seen between the data and the imposed model. Employing insignificant terms in the response surface model not only creates more variance between the model and the data but introduces possible noise when evaluating the robustness of the component or system.

The particular yield or response of a system may not lend itself to a polynomial curve fit over a reasonable design space. When this is the case, it may be possible to identify a parameter that is closely linked to the desired response and use it in the response surface evaluation. The response can then be used in subsequent calculations to obtain the desired performance information. Such is often the case with fatigue life predictions.

Extrapolation of response surfaces should not be used at decision points. Extrapolations can be useful in guiding the direction for moving the design space, but decision points should only be made with evaluations within the domain of the design space.