A Better Paper Helicopter: Process Optimization Using Response Surface Methodology

Erik Barry Erhardt[1]

My motivation for writing this article is to share an experience I had my first graduate semester that immediately met my needs for creativity, discovery, growth, learning, and stimulation as a young statistician. The methods used to optimize a paper helicopter cover several core ideas in statistics and mathematics. Also, by designing and performing our own experiments, we have intimate connection with the data, which too often is taken as given.

1 Problem

A group of your friends are having a friendly contest to design a paper helicopter that remains aloft the longest when dropped from a height. To be fair, everyone starts with an initial helicopter pattern, given in Figure 1, which is an easy pattern to modify and replicate. With an interest and knowledge of statistics, and a desire to bring excellence to the project, you have an (unfair) advantage. I'll show the tools to win, introducing the ideas of response surface methodology along with how my classmate, Hantao Mai, and I optimized a paper helicopter design to become Worcester Polytechnic Institute's two-time paper helicopter champions (Erhardt & Mai, 2002).

I begin by introducing the ideas of response surface methodology, curve estimation with first- and second-order polynomials, and the sequential process of improvement and optimization. Next, regression and factorial designs are discussed. Then the fun — we'll begin experiments that lead us to a winning helicopter.

[1]Erik was introduced to RSM at and received his M.S. in statistics from Worcester Polytechnic Institute. He is currently a Ph.D. student and PIBBS Fellow at the University of New Mexico, and American Statistical Association Albuquerque Chapter representative. He enjoys teaching, and as a teaching assistant for every TA-able statistics course, both undergraduate and graduate, and a graduate course outside statistics department, he received a 2005–2006 UNM Outstanding Teaching Assistant Award, as well as a departmental teaching award. As a research assistant, Erik is statistician for the largest individual case-control study of Hodgkin's disease in children centered at the UNM Cancer Research and Treatment Center. Erik enjoys consulting and interdisciplinary collaborative research. In his free time he plays pool and snooker, contra dances, unicycles, and likes to continually improve things.
Department of Mathematics and Statistics
The University of New Mexico
Albuquerque, New Mexico 87131-1141
Phone: (505) 277-0757
Fax: (505) 277-5505
e-mail: erike@stat.unm.edu

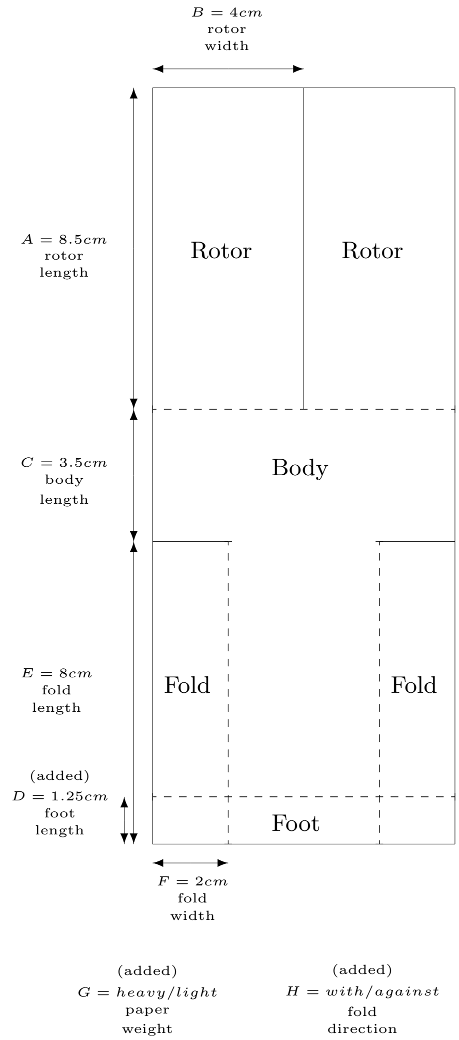

Figure 1: Initial Helicopter Pattern. Cut along solid lines, fold along dotted lines. Foot fold D, paper weight G, and fold direction H were not included as part of the initial pattern, but added after a brainstorm of potential factors influencing the flight time response.

2 Response Surface Methodology

Response surface methodology (RSM) is a collection of statistical and mathematical techniques used to parsimoniously explore and improve a process (Myers & Montgomery, 1995). The major tools used are (1) design of experiments (DOE), (2) multiple regression, and (3) optimization.

In the model, we assume the (possibly multivariate) response of flight time, y, depends upon k controllable factor inputs, such as rotor length and width, $\xi_1, \ldots, \xi_k$, in $y = f(\xi_1, \ldots, \xi_k) + \varepsilon$, where f is an unknown function (usually assumed smooth) and ε is random error with expected value $\mathrm{E}(\varepsilon) = 0$ and common variance $\mathrm{Var}(\varepsilon) = \sigma^2$. The **response surface** is the expected flight time for varying sets of factor inputs defined by $\mathrm{E}(y) = f(\xi_1, \ldots, \xi_k)$. For convenience, the natural variables, $\xi_1, \ldots, \xi_k$, in their original units are often transformed to unitless coded variables, $x_1, \ldots, x_k$ with zero mean $\mu_i = 0$ and common standard deviation. The response surface used is in terms of the coded variables $\mathrm{E}(y) = f(x_1, \ldots, x_k)$.

Because we often don't know the form of f, we need to approximate it, at least locally. For this purpose first- or second-order polynomials are useful, because of Taylor's theorem that surfaces look flat up close. First- and second-order polynomials have the form

$$\begin{aligned} \mathrm{E}(y) &\doteq \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j \\ &\quad + \sum_{i=1}^{k} \beta_{ii} x_{ii}^2. \end{aligned} \tag{1}$$

The first line of equation (1) includes the flat first-order terms including cross-products (also called interactions), and the second line includes the curved second-order terms. Four examples of possible response surfaces appear in Figure 2.

Our goal is to find the settings of factors that maximize flight time. We will design a set of sequential experiments that allow us to discover enough about the shape of the response surface to design a better and likely winning paper helicopter.

2.1 Sequential Process

The sequential process of RSM has three primary phases.

In PHASE 0 a **brainstorming session** reveals many potential factors influencing the response, everything you can think of that you have control over. Using a **screening experiment** eliminates many of the trivial variables, retaining the vital few important factors influencing the response. The dimension of the experimental region is often greatly reduced, exponentially decreasing the number of further experiments required.

In PHASE 1 an experiment is performed to see whether the settings of the important factors produce a near-optimal response. If the process is in a region (possibly) remote from the optimum, use a first-order model to move in the direction of **steepest ascent** toward a peak on the surface.

In PHASE 2 a near-optimum location has been found and a second-order model is used to exhibit the curvature near the optimum in order to **determine optimal conditions**. A

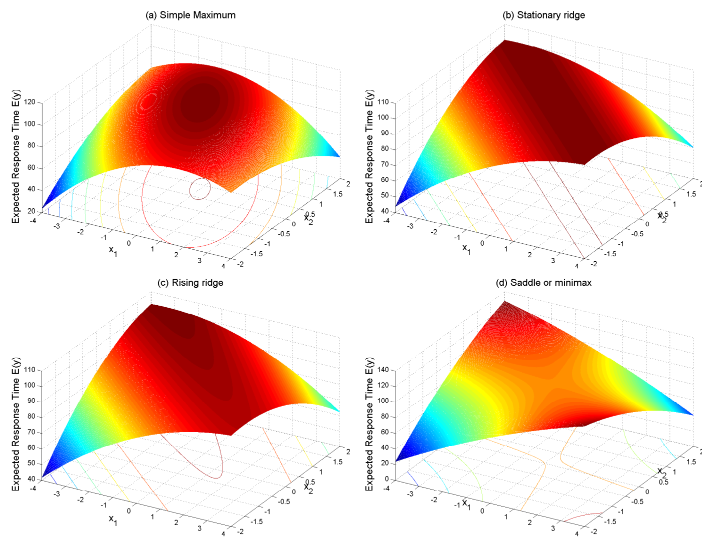

Figure 2: Examples of four common types of response surfaces defined by the second-order model in $k = 2$ variables. (a) Simple maximum has a unique peak at a point, (b) stationary ridge has a common maximum along a line (hyperplane), (c) rising ridge does not have a maximum, but a single direction which most steeply increases, and a (d) saddle has a point which is the maximum in one direction, but a minimum in another direction, hence its other name, minimax.

(a) Simple maximum,
$\mathrm{E}(y) = 100 + x_1 + 2x_2 - 2x_1^2 - 3x_2^2 - 3x_1x_2$.

(b) Stationary ridge,
$\mathrm{E}(y) = 100 + x_1 + 2x_2 - \frac{3}{4}x_1^2 - 3x_2^2 - 3x_1x_2$.

(c) Rising ridge,
$\mathrm{E}(y) = 100 + x_1 + 3x_2 - \frac{3}{4}x_1^2 - 3x_2^2 - 3x_1x_2$.

(d) Saddle or minimax,
$\mathrm{E}(y) = 100 + x_1 + 2x_2 - \frac{1}{2}x_1^2 - 3x_2^2 - 6x_1x_2$.

confirmatory experiment at the optimum location verifies the second-order prediction.

In each phase, multiple linear regression models are used to analyze experiments based on factorial designs. Next, I briefly introduce regression models, then describe factorial designs in some detail. Familiar with these, we will begin our experimentation and helicopter improvement.

2.2 Multiple Linear Regression Models

Linear models, multiple linear regression models in particular, are used to approximate the response surface. These models are linear with respect to the model coefficients, $\beta_1, \ldots, \beta_k$, but not necessarily in the variables, $x_i, x_i^2, \log(x_i)$. We will consider first- and second-order models as in equation (1). Note that nonlinear models can be used if there are theoretical justifications for the response surface to have a particular functional form. Use of linear regression models is advantageous because they are well studied and are implemented in nearly every statistics software package. We use SAS in this article.

3 Factorial Designs

Factorial designs are efficient designs for assessing the marginal and joint effects of several factors on a response. A 2^k design is a factorial design for k factors, each at two levels $(-1, +1)$. A graphical display of a 2^4 design is shown in Figure 3.

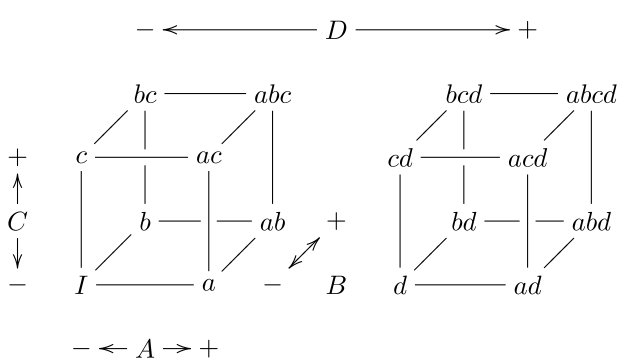

Figure 3: The 2^4 factorial design consists of all combinations of four factors each taking two levels $(-1, +1)$, represented here as the letter for $+1$ and no letter for -1. For example, point abd in the figure corresponds to factors A, B, and D set to high level $+1$ and C to low level -1.

The main applications of these designs to RSM are in the screening experiment to reduce the number of factors to a manageable number, for fitting first-order models to determine the path of steepest ascent, and as the basic building block for other RSM designs. The addition of center points, runs at level 0 of all factors, is an inexpensive way of checking curvature and obtaining an independent estimate of the error variance, and doesn't effect the orthogonality of the design.

3.1 Fractional Factorial Designs

Fractional factorial designs (FFDs) are used when the full factorial requires a prohibitive number of runs. When the number of factors k is large, a 2^k design will have a large number of runs. For example, $k = 8$ requires $2^8 = 256$ runs, having 255 degrees-of-freedom (df) to estimate factor effects. Yet, only 8 df are used to estimate main effects, 28 df are used to estimate two-factor interactions, and the remaining 219 df are used to estimate three-factor and higher interactions. If it is reasonable to assume that the effect of these higher-order interactions is negligible, then the main- and low-order effects can be estimated by running a fraction of the number of experiments, for example 2^{k-p}.

Successful use of FFDs relies on three key ideas. First, the **sparsity of effects principle** says there are only a few main effects and/or low-order interactions that account for almost all the variation in the response, making the higher-order effects negligible. Second, the **projection property** says FFDs can be projected into stronger designs, allowing analysis as a full factorial in some factors by ignoring others. Third, **sequential experimentation** allows combinations of FFDs to sequentially assemble a larger design, allowing additional experimental trials to be added to the existing set of trials.

The price paid for the efficiency of the 2^{k-p} design is the ambiguity introduced through an aliasing structure which confounds low-order with high-order effects. For example, with eight factors, there are $2^8 = 256$ effects: I is the identity; $a, b, \ldots, h$ are main effects; $ab, ac, \ldots, gh$ are two-factor interactions; up to the eight-factor interaction, $abcdefgh$. In our 2^{8-4} design used in our screening experiment, the aliasing structure based on $E = BCD, F = ACD, G = ABD, H = ABC$ confounds the three-factor and higher interactions with the main effects. (See Appendix A.1 in original project report for the full aliasing structure.) Thus if factor $h = abc$ is found to have a significant influence on the response, we wouldn't know whether it was main factor h or three-factor interaction abc, or both, that contributes to the effect observed.

To resolve this ambiguity we rely on philosophical principles and practical methods. The **Pareto principle** says there are vital few important effects and a trivial many. The **hierarchy of importance** expects lower-order effects to be more important than higher-order effects. The **knowledge of confounding patterns**, also called the aliasing structure, and reasoning would say that if ab is aliased with ch, and three effects a, b, and $(ab = ch)$ are all judged important, then it is likely that ab is important and ch is not, since main effects a and b were each important. Conducting further experiments can resolve ambiguities.

3.2 Blocking

Blocking is a device for reducing variation when all the runs of an experiment can't be performed under homogeneous conditions. Use the motto, "block what you can, randomize the rest." For example, if I drop some helicopters and Hantao drops the rest, then we would include a block effect for who drops since it is possible that dropping and flight timing differs between us. The analysis is handled by including a block effect and uses $b - 1$ degrees-of-freedom for b blocks.

We brainstorm to define conditions under which blocking is a necessary consideration. For helicopter creation, considerations include the paper stock, who marks the pattern on the paper, who cuts the paper, and who folds the paper. For helicopter dropping, considerations include the conditions of the day (atmospheric pressure, temperature, dew point), conditions of the time (whether processes such as climate control are running, movement by people, or other events effecting air circulation), drop location, who drops the helicopters, dropping method, and who records the time.

We resolve blocking issues by planning our experimental protocol to eliminate as many conditional factors as possible by being consistent in the conditions we can control. For example, Hantao plots the helicopter patterns on the paper and I cut and fold them. Hantao drops the helicopters while I record the time. We always drop the helicopters from the second floor balcony in Fuller Laboratories at WPI. Hantao releases the helicopters while I time the drop with a stopwatch from the ground floor. If a helicopter experiences disturbances during a drop, such as hitting a wall, or any other observable conditions which effects its performance, we redrop that helicopter. We limit ourselves to dropping only when the conditions are observed to be homogeneous, though we have no way of measuring these conditions. Under this situation, blocking in a single experiment is no longer a necessary consideration.

For quality control, when manufacturing helicopters, a long ruler, a liquid-ink pen, and a paper cutter provide consistent results. We limit error in measuring, cutting, and folding to no more than 0.1 centimeters, producing a replacement if these standards aren't met. Similarly, when dropping, if Hantao and I are not well synchronized for the drop and timing, we redrop that helicopter.

4 Let the contest begin!

4.1 Phase 0: Brainstorm and Screening Experiment

First a brainstorm about helicopter design: Bring your knowledge and experience forward about helicopter flight — make a few and drop them, try modifications. This is fun!

After careful consideration, we identify eight factors possibly influencing the flight time of the helicopter, the original five factors plus three we added (D foot length, G paper weight[2], H fold direction[3]), included in both Figure 1 and Table 1. Notice there is no issue regarding the mixture of continuous (lengths and widths) and categorical (weight and direction) factor levels. We decide to retain all the factors for experimentation. At the outset, it is better to have too many rather than too few factors.

In order to determine the vital few factors by the Pareto principle that contribute to flight time, we perform a screening experiment using a 2^{8-4} fractional factorial design with the initial helicopter pattern as our center point.

Factors	Coded Values −1	0	+1
A = Rotor Length	5.5	8.5	11.5
B = Rotor Width	3	4	5
C = Body Length	1.5	3.5	5.5
D = Foot Length	0	1.25	2.5
E = Fold Length	5	8	11
F = Fold Width	1.5	2	2.5
G = Paper Weight	light	(none)	heavy
H = Fold Direction	against	(none)	with

Table 1: Potential factors influencing the flight time of the paper helicopter and their levels for the screening experiment in coded and natural units.

It is likely that our initial helicopter pattern begins our experimentation outside the region of optimum conditions. Away from the peak of our hill, we can learn where the peak may be by heading in the steepest direction. At this stage a flat first-order approximation might be reasonable.

4.1.1 Initial Factor Levels

We assign the initial helicopter pattern in Figure 1 as the center point, coding the levels to 0. Then, we choose factor ranges seeming reasonable for this helicopter pattern, assigning values of the natural units at the low and high levels of the range to the coded values of −1 and +1, as in Table 1.

4.1.2 Screening Experiment

The goal during the screening experiment is to identify those factors with the most influence on the response and then eliminate the remaining factors from further analysis. We produce the 16 helicopters in random order at the design points listed in Table 2. Then we drop them in random order and record the drop time of each.

		Factors and Coded Levels								
Obs	Ord	A	B	C	D	E	F	G	H	Time
1	12	−1	−1	−1	−1	−1	−1	−1	−1	11.80
2	7	−1	−1	−1	1	1	1	1	−1	8.29
3	11	−1	−1	1	−1	1	1	−1	1	9.00
4	15	−1	−1	1	1	−1	−1	1	1	7.21
5	1	−1	1	−1	−1	1	−1	1	1	6.65
6	4	−1	1	−1	1	−1	1	−1	1	10.26
7	16	−1	1	1	−1	−1	1	1	−1	7.98
8	8	−1	1	1	1	1	−1	−1	−1	8.06
9	3	1	−1	−1	−1	−1	1	1	1	9.20
10	10	1	−1	−1	1	1	−1	−1	1	19.35
11	9	1	−1	1	−1	1	−1	1	−1	12.08
12	5	1	−1	1	1	−1	1	−1	−1	20.50
13	6	1	1	−1	−1	1	1	−1	−1	13.58
14	13	1	1	−1	1	−1	−1	1	−1	7.47
15	14	1	1	1	−1	−1	−1	−1	1	9.79
16	2	1	1	1	1	1	1	1	1	9.20

Table 2: Screening Drop Coded Factor Levels with Results

With only a single replicate there is no estimate of error, since all of the degrees-of-freedom are being used for estimating the factor effects. Therefore, we can't use the standard t-tests for testing the significance of factor effects. Let's look at two methods for identifying the important factors having a profound effect on the response when we can't estimate error. A large effect is an outlier, so how would you identify outliers? A graphical method uses a normal quantile

[2]Since Paper Weight is not a continuous factor, we confine ourselves to two levels. Light is from phone book whitepages and heavy is standard copy paper.

[3]Since Fold Direction is not a continuous factor, we confine ourselves to two levels. Against indicates the fold direction is opposite the direction of rotation. For example, if the rotor on right side of the helicopter folds away from you, the fold on the right side will fold toward you.

plot of the effects with outliers identified as significant effects (Daniel, 1959). The plot in Figure 4 does not strongly suggest any outliers, though A (rotor length), B (rotor width), and $G = ABD$ (paper weight) are the most likely candidates. Non-graphical methods include Lenth's (S)MOE methods for unreplicated factorials (Lenth, 1989). Lenth's method calculates MOE (margin-of-error) and SMOE (simultaneous margin-of-error) bars based on the median (not the mean) of m effects. We use SAS macro EFFECTS to compare the factor effect estimates with the MOE and SMOE at the 0.95 confidence level (Petruccelli *et al.*, 1999). The effects in Table 3 fail to exceed the inner MOE bars (± 4.94516) or outer SMOE bars (± 10.0394), so there is insufficient evidence that any factors have an effect on the response.

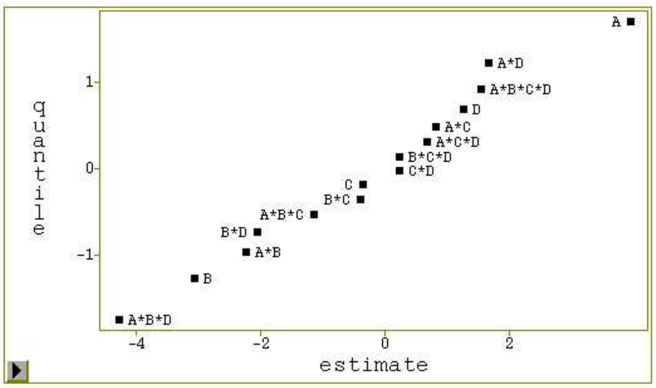

Figure 4: Normal Quantile Plot indicates A (rotor length), B (rotor width), and $G = ABD$ (paper weight) are the most likely significant effects.

Label	Factor	Estimate	
A	A	3.9900	*
B	B	−3.0550	*
C	C	−0.3475	
D	D	1.2825	
I234	$E = BCD$	0.2500	
I134	$F = ACD$	0.7000	
I124	$G = ABD$	−4.2825	*
I123	$H = ABC$	−1.1375	
I12	AB	−2.2175	
I13	AC	0.8400	
I14	AD	1.6850	
I23	BC	−0.3850	
I24	BD	−2.0350	
I34	CD	0.2475	
I1234	$ABCD$	1.5625	

Table 3: Lenth's (S)MOE effects estimates are compared to unreplicated 95% MOE bars (± 4.94516) and outer SMOE bars (± 10.0394) and no factors are found significant. After recalculation based on replicated center points, 95% MOE bars ± 2.4332 indicates A, B, and $G = ABD$ are significant.

No significant factors, even when the response times varied from 7 to 20 seconds?! Do not be discouraged. We are in control of our own experiment and additional trials are easy. We know having a larger sample reduces estimation error, so... one way to reduce the size of those (S)MOE error bars is to run replicated experiments. Design augmentation with center points provides an estimate of experimental (pure) error (PE), reducing the size of the error bars and allowing t-tests, and an indication of curvature, while retaining the orthogonality of the original design. Since factor G (paper weight) appears to have an effect, we run two sets of three center points, one set at the high and the other set at the low level of G. Our center points are not at the actual continuous center because there are only two categorical levels of G (paper weight) and H (fold direction), therefore it is necessary to compute the MOE and SMOE values manually. (S)MOE $= s_{\text{PE}} \times t_{c-1,\delta}$, where $\delta = \frac{1+L}{2}$ for MOE and $\delta = \frac{1+L^{1/m}}{2}$ for SMOE, and confidence level $L = 0.95$, and

$$s_{\text{PE}} = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)}} \quad (2)$$

is the weighted sample standard deviation of the two sets of $n_1 = n_2 = 3$ responses at the $c = 6$ center points.

In Table 4 we set each factor to center coded value zero, except G and H. We parsimoniously run H only its low level, and run G and both its levels in order to average on either side of its undefined center level.

		Factors and Coded Levels								
Obs	Ord	A	B	C	D	E	F	G	H	Time
1	5	0	0	0	0	0	0	−1	−1	10.52
2	1	0	0	0	0	0	0	−1	−1	10.81
3	3	0	0	0	0	0	0	−1	−1	10.89
4	2	0	0	0	0	0	0	1	−1	15.91
5	4	0	0	0	0	0	0	1	−1	16.08
6	6	0	0	0	0	0	0	1	−1	13.88

Table 4: Screening Drop Results of Center Points

With these center point trials we estimate $s_{PE} = 0.8764$, resulting in $\text{MOE}_{0.05} = 2.4332$ at the 0.05 significance level. At last, comparing estimates in Table 3 to the MOE based on center points, we find main effects A, B, and $G = ABD$ significant.

We have three main effects, but clearly $G = -1$ (the light paper) gives a higher response. Also, we fix $H = -1$ (fold against) since this setting both gave observably more stable helicopters and tends to have better performance.

With the reduction of factors through lack of significance (C, D, E, F and H) and confinement (G and H), we have only two factors to consider (A and B). Hereafter, we can use the full 2^2 factorial replicated design to start our optimization process.

4.2 Phase 1: **Process Improvement: Steepest Ascent**

What's the shortest path to the top of a hill? Always take the steepest possible step. The approximate method we'll use makes the first step the steepest possible and continues in that direction until two steps head downward. Then we look around (experiment) at our local maximum and reevaluate our course.

The goal is to move the process variables, x, to a new region of improved response by following the path of **steepest ascent** using experimental design, model building, and sequential experimentation. We'll assume initially that a flat first-order model is adequate.

The procedure is to first fit a first-order model based on an orthogonal design, such as 2^k or 2^{k-p}, perhaps with center points. Based on the model fit, compute the path of steepest ascent. Conduct experiments along this path until a local maximum is reached. Centered at the location of the local maximum, conduct an experiment using a first-order design with center points. If curvature or interaction lack-of-fit is found, explore the region more thoroughly for a local maximum with a second-order design, otherwise again compute the path of steepest ascent and repeat.

Using a regression model, calculating the path of steepest ascent is as easy as reading the coefficients of the model. The direction and magnitude of steepest ascent is the **gradient**. In one dimension this is the first derivative, the slope, and in two or more dimensions this is the direction of steepest slope for each point on the surface. For the response surface $\text{E}(y) = f(x_1, \ldots, x_k)$, the direction of steepest ascent as the gradient is $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_k}\right)'$, a vector of slopes. In particular, for the first-order model $\text{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, the path of steepest ascent is in the direction $(\beta_1, \beta_2, \ldots, \beta_k)'$. Fitting the model $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$, the resulting estimated direction of steepest ascent is $(b_1, b_2, \ldots, b_k)'$.

With only two factors it is practical to use a 2^2 replicated design. By the previously mentioned projection property, we reuse the previously folded helicopters made from light paper ($G = -1$), modifying all fold directions to against ($H = -1$). Half our experiment uses previously manufactured helicopters, and we fold eight new helicopters to complete the 2^2 design with four replicates, and reuse three for center runs.

These old and new blocks of helicopters, have different characteristics in the insignificant factors (C, D, E, F, and H) but should have comparable performance since these factors have little effect on the response. New measurements are taken for all helicopters, appearing in Table 5.

How can we use the three center runs in our experiment to check surface curvature? If the surface is flat, where will the mean response at the center point ($\bar{y}_C$) be relative to the mean of the responses at the corner factorial points ($\bar{y}_F$)? At the same place! How can we measure the curvature using $\bar{y}_C$ and $\bar{y}_F$? An estimate of the curvature is the difference $\bar{y}_F - \bar{y}_C$, and if this is large relative to experimental error, there is evidence that the response surface is curved. A simple two-sample t-test does the trick.

The estimated standard error of $\bar{y}_F - \bar{y}_C$ is

$$\hat{\sigma}_{\bar{y}_F - \bar{y}_C} = \sqrt{s^2_{\text{PE}}\left(\frac{1}{n_F} + \frac{1}{n_C}\right)} = \sqrt{0.8464\left(\frac{1}{16} + \frac{1}{3}\right)} = 0.5788 \quad (3)$$

and the critical cutoff at the 95% level ($L = 0.95$) is

$$\hat{\sigma}_{\bar{y}_F - \bar{y}_C} t_{c-1,(1+L)/2} = 0.5788(2.92) = 1.69 > |\bar{y}_F - \bar{y}_C| = 0.4389. \quad (4)$$

Obs	A	B	Time
1	−1	−1	10.24
2	−1	−1	9.11
3	1	−1	16.52
4	1	−1	16.99
5	−1	1	10.20
6	−1	1	9.26
7	1	1	10.02
8	1	1	9.94
9	−1	−1	11.31
10	−1	−1	10.94
11	1	−1	12.58
12	1	−1	13.86
13	−1	1	8.20
14	−1	1	9.92
15	1	1	9.95
16	1	1	9.93
17	0	0	11.67
18	0	0	10.74
19	0	0	9.83

Table 5: 2^2 replicated design results are used to determine the path of steepest ascent.

The difference is not larger than expected so there is insufficient evidence to conclude that there is surface curvature. We prepare to climb the flat face of the hill.

The estimated response surface is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 = 11.1163 + 1.2881 x_1 - 1.5081 x_2. \quad (5)$$

For our climb we need to decide how large our steps will be. We decide each step will be $1cm$ in factor A, corresponding to $\frac{1}{3}$ design units. The corresponding step size in design variable B is then $(b_2/b_1)(1/3) = (-1.5081/1.2881)(1/3) = -0.39$ in design units, giving natural units of $-0.39 \times 1 = -0.39cm$.

Before conducting steepest ascent, a pilot study in the insignificant factors helps to stabilize the helicopters. Experience suggests the factors determined insignificant do, in fact, effect helicopter flight. We want to find good settings for the insignificant factors D, E, and F using steepest ascent design points Base + 1Δ and Base + 2Δ. With these factors set, we will continue with our steepest ascent experiments. We decide to set factor C to 2 and then vary levels of factors D, E, and F in regions our experience suggests are good.

These quick experiments are not formally analyzed, and the helicopters are dropped from standing on a table rather than the high balcony otherwise used. The first experiment in the top of Table 6 suggest factor F (fold width) is preferable at $2cm$. The two experiments combined suggest a preferred level of $D = \frac{1}{4}E$. Also, factor E at $6cm$ gives a stable helicopter, so we will use that level.

Therefore, we use favorable levels of body length $C = 2$, foot length $D = 2$, fold length $E = 6$, and fold width $F = 2$, in their natural units. As we follow the path of steepest ascent, F may be relatively large enough to wrap around the "tail" of the helicopter, so if this becomes the case, we will confine ourselves to $F = \frac{2}{3}B$, resulting in fold widths dividing the tail into three equal parts.

Factors			Response Time	
D	E	F	Base + 1Δ	Base + 2Δ
0	9	1	2.98	3.30
2	9	1	3.77	3.64
4	9	1	3.35	3.95
5	9	1	3.43	3.91
4	9	2	3.82	4.07
2	9	2	3.72	3.67
0	9	2	3.36	3.64
2	6	2	3.90	4.70
4	6	2	4.03	3.22
0	6	2	3.44	3.89
0	6	2	3.26	3.50
2	6	2	4.08	3.82

Table 6: Steepest ascent preparation results for two experiments used to determine robust levels for the insignificant factors.

We predict five steps along the path of steepest ascent will bring us near the optimum region followed by a deterioration in response. Six runs are made along the path, starting at the base, given in Table 7. Eventually, severe deterioration in response occurs when the first order approximation is no longer valid. An approximate maximum is found at Base + 3Δ, so our follow-up experiment will involve an experimental design centered at Base + 3Δ.

	Factors			
Step	A	B	(F)	Time
Base	8.5	4.00	(2.0)	12.99
Δ = step size	1.0	−0.39		
Base + 1Δ	9.5	3.61	(2.0)	15.22
Base + 2Δ	10.5	3.22	(2.0)	16.34
Base + 3Δ	11.5	2.83	(1.5)	**18.78**
Base + 4Δ	12.5	2.44	(1.5)	17.39
Base + 5Δ	13.5	2.05	(1.2)	7.24

Table 7: Steepest ascent coordinates and results in natural units; maximum in three steps.

4.3 Phase 2: Determining Optimal Conditions: Second-order Response Surface

Near the optimum, the response surface is curved. The shape of the surface near the optimum value can be approximated by the second-order model in equation (1). A maximum or minimum will occur at a stationary point, the critical point where the derivatives are all 0, $\frac{\partial \hat{y}}{\partial x_i} = 0$, for $i = 1, \ldots, k$. Determining the shape of the surface is easiest done by plotting the surface or looking at contour plots if k is 1 or 2. If there are lots of x variables, then visualizing the surface is difficult, but information about curvature can be understood analytically. Canonical analysis takes a square matrix of the b_{ij} coefficients and calculates eigenvalues $\lambda_1, \ldots, \lambda_k$ and eigenvectors. The eigenvalues tells us the direction and magnitude of curvature, with sign indicating downward ($-$ =concave) or upward ($+$ =convex) curvature, and magnitude indicating fairly flat (near 0) to steeply peaked (far from 0). For example, with two x variables, $(\lambda_1, \lambda_2) = (-1, -2)$ indicates a maximum in both directions with the hill being steeper in the x_2 direction, and $(-0.2, 3)$ indicates a saddle point, curved down slightly in the x_1 direction, but curved up steeply in the x_2 direction. Wanting a maximum, we hope all our λs are negative.

4.3.1 Central Composite Design

The workhorses of second-order response surface models are **central composite designs** (CCDs). A graphical display of a two-dimensional CCD is shown in Figure 5. They consist of a (fractional) factorial design for estimating the first-order and two-factor interactions, a set of $2k$ axial points $(\pm a, 0, \ldots, 0)$, $(0, \pm a, 0, \ldots, 0)$, …, $(0, \ldots, 0, \pm a)$ for estimating the second-order terms, and repeated center points for estimating the second-order terms and to provide an internal estimate of (pure) error. The value of a can be 1, $\sqrt{k}$, or other values based on an optimality criteria.

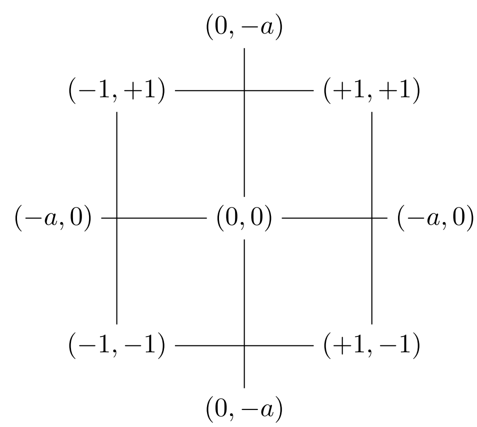

Figure 5: A two-dimensional central composite design.

At the top of our steepest ascent path we want to explore the space for curvature, and a spherical (circular, where $a = \sqrt{k} = \sqrt{2}$) 2^2 Central Composite Design with center points will do the trick. If curvature is not detected, we repeat steepest ascent.

The CCD is given in Table 8 in coded and natural units centered at the maximum steepest ascent point, Base + 3Δ. Since helicopters are easy, and fun, to make, we assume the risk of performing unnecessary runs at the axial points in the case where curvature is not detected in consideration of dropping all helicopters in one block. If curvature is found the axial points will be necessary for estimation of quadratic terms. The drop results are in Table 9.

Using equations (3) and (4) we again test for response curvature. As before, the estimated standard error of $\bar{y}_F - \bar{y}_C$ is

$$\hat{\sigma}_{\bar{y}_F - \bar{y}_C} = \sqrt{0.3342\left(\frac{1}{4} + \frac{1}{3}\right)} = 0.4415 \quad (6)$$

Factor	$-\sqrt{2}$	-1	0	1	$\sqrt{2}$
A	10.08	10.50	11.50	12.50	12.91
B	2.28	2.44	2.83	3.22	3.38
C			2.0		
D			1.5		
E			6.0		
F			1.5		
G			light		
H			against		

Table 8: CCD levels in coded and natural units for second-order experiment at the peak of steepest ascent path.

Obs	Ord	A	B	Time
1	7	-1	-1	13.65
2	3	$+1$	-1	13.74
3	11	-1	$+1$	15.48
4	5	$+1$	$+1$	13.53
5	9	0	0	17.38
6	2	0	0	16.35
7	1	0	0	16.41
8	10	$+1.414$	0	12.51
9	4	-1.414	0	15.17
10	6	0	$+1.414$	14.86
11	8	0	-1.414	11.85

Table 9: CCD experiment results.

and the critical cutoff at the 95% level ($L = 0.95$) is

$$\begin{aligned}\hat{\sigma}_{\bar{y}_F-\bar{y}_C} t_{c-1,(1+L)/2} &= 0.4415(2.92) \\ &= 1.289 < |\bar{y}_F - \bar{y}_C| = 2.6133 \qquad (7)\end{aligned}$$

providing sufficient evidence that there is surface curvature. Finding curvature completes our path of steepest ascent and we proceed with second-order modeling.

The estimated second-order model is

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &\quad +b_{12}x_1x_2 + b_{11}x_{11}^2 + b_{22}x_{22}^2 \\ &= 16.713 - 0.702x_1 + 0.735x_2 \\ &\quad -0.510x_1x_2 - 1.311x_{11}^2 - 1.554x_{22}^2. \qquad (8)\end{aligned}$$

To validate our model we check the significance of the model terms, perform a lack-of-fit test, and check our statistical assumptions.

Using t-tests for regression coefficients, Table 10 indicates all factors except the interaction term are significant at the 0.05 level. The lack-of-fit test suggests the model adequately describes the data because the p-value = 0.3737 in Table 11 is large.

The statistical assumptions of independent and identically distributed normal residuals are both satisfied. In Figure 6, the residuals vs. predicted value plot indicates the residuals are random with constant variance, and the residuals follow the line in the normal quantile plot, indicating normality. Additionally, the Shapiro-Wilk normality test has a large p-value = 0.7613 indicating the normality assumption is reasonable. Therefore, our second-order model appears reasonable.

Parameter	df	Estimate	Std Err	t Value	Pr > $\lvert t \rvert$
Intercept	1	16.71	0.408	40.98	0.0001
$x_1 = a$	1	-0.70	0.250	-2.81	0.0374
$x_2 = b$	1	0.73	0.250	2.94	0.0322
$x_{11} = a * a$	1	-1.31	0.297	-4.41	0.0070
$x_{12} = a * b$	1	-0.51	0.353	-1.44	0.2083
$x_{22} = b * b$	1	-1.55	0.297	-5.23	0.0034

Table 10: Second-order response surface parameter estimates show all but the cross product significant.

Residual	df	SS	MS	F Value	Pr > F
Lack of Fit	3	1.826	0.609	1.82	0.3737
Pure Error	2	0.668	0.334		
Total Error	5	2.495	0.499		

Table 11: Lack-of-Fit Test does not indicate the model does not fit.

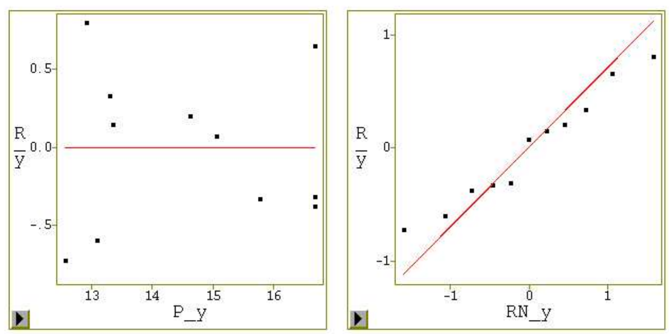

Figure 6: Residual plots does not suggest departures from normality.

Now we use our model to predict the **stationary point**, which we expect, and hope, is a maximum. SAS's RSREG (response surface regression) procedure gives us the location of the stationary point ($a = -0.32, b = 0.29$), the predicted response $\hat{y} = 16.9$, and the eigenvalues $(\lambda_a, \lambda_b) = (-1.15, -1.71)$ indicate a maximum — yes! The final pattern is defined by the values in Table 12 and shown in Figure 8. The estimated response surface in (8) is plotted in Figure 7 with the 90% confidence region (Castillo & Cahya, 2001) emphasized at the top. Our model $R^2 = 0.92$, indicating the model should give precise predictions. A model fitting as well as ours, having a confidence region moderate in area, implies some flexibility in choosing factor levels for a near-optimum. In many situations, estimation of optimum conditions is very difficult. However, RSM should still produce important information about the process.

To confirm the model prediction, a confirmatory experiment consisting of six new helicopters with the optimal setting are recorded in Table 13. This confirmatory experiment gives a mean response of 17.81 with standard deviation 1.67, with a 95% confidence interval on the mean response of $(15.83, 18.04)$.

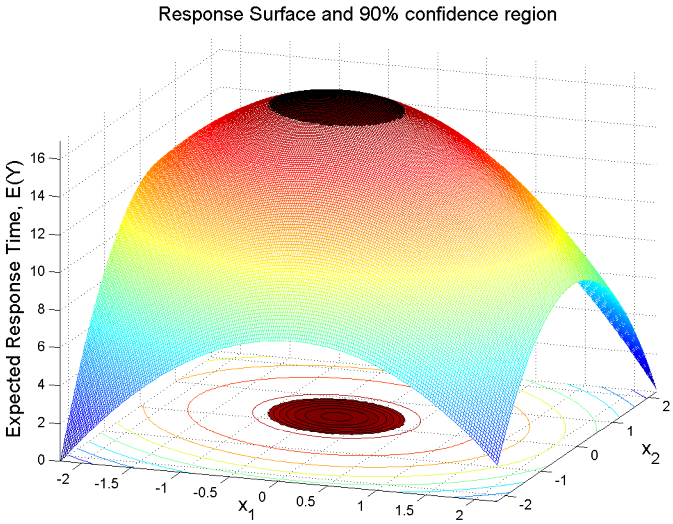

Figure 7: Estimated response surface with 90% confidence region for the coordinates of maximum response.

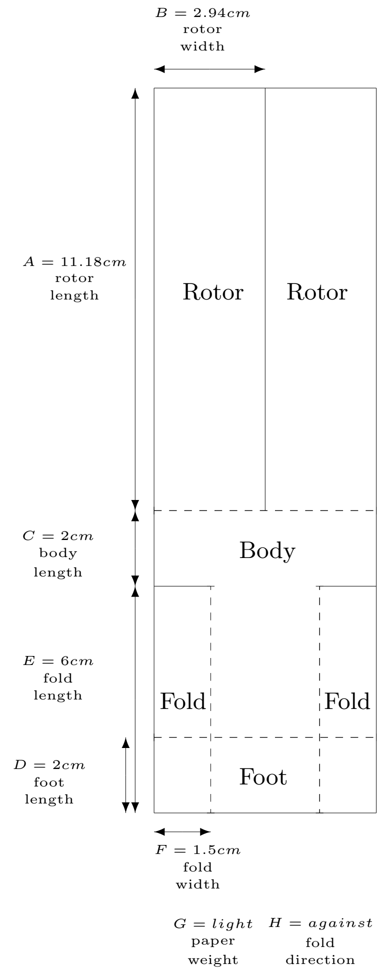

Figure 8: The optimal helicopter pattern is longer, narrower, and has a shorter base than the initial helicopter pattern. Cut along solid lines, fold along dotted lines.

Factor	Coded	Natural
A	-0.32	11.18
B	0.29	2.94
C		2.0
D		2.0
E		6.0
F		1.5
G		light
H		against

Table 12: Optimum helicopter factor levels.

Our model's predicted response $\hat{y} = 16.9$ falls within this interval, confirming the validity of our response model.

Obs	Time
1	15.54
2	16.40
3	19.67
4	19.41
5	18.55
6	17.29

Table 13: Confirmatory experiment results shows that the predicted mean response from the model is accurate.

An interesting factor that we hadn't considered was scale — how large should the helicopter be? In the second contest I also entered my "nanocopter", a new tiny paper helicopter developed in five minutes. It is $5cm$ long, $1cm$ wide, has a $2.5cm$ rotor length, and a $2cm$ tail that is $0.5cm$ wide. This helicopter flew as well as our optimum helicopter! Figure 9 shows a photograph of almost all of the helicopters we made for our experiments. Figure 10 shows a photograph of two helicopters in action. Figure 11 has a single plot displaying all the experiments and model fits; it is a map of the entire process.

The methods of RSM are much broader than what appear in this paper, and with them almost any process can be better understood and improved. Besides, uncontrollable noise factors change with time, so a set of optimum conditions may be a fleeting concept. What is more important is for the analysis to reveal important information about the process and about the roles of the variables. The computation of a stationary point, a canonical analysis, or a ridge analysis may lead to important information about the process, and this in the long run will often be more valuable than a single set of coordinates representing an estimate of optimum conditions.

References

Castillo, E.D. & Cahya, S. (2001). Statistical computing and graphics — A tool for computing confidence regions on the stationary point of a response surface. *The American Statistician*, **55**, 358–365, http://lib.stat.cmu.edu/TAS/BH/.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311–342.

Erhardt, E.B. & Mai, H. (2002). The search for the optimal paper helicopter, The original project report for MA 544 Statistical Response Surface Analysis at WPI.

Lenth, R.V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, **31**, 469–473.

Myers, R.H. & Montgomery, D.C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley.

Petruccelli, J.D., Nandram, B. & Chen, M. (1999). *Applied Statistics for Engineers and Scientists*. Prentice Hall, SAS and Minitab macros to accompany the text at http://users.wpi.edu/~jdp/downloads/book_macros/main.html.

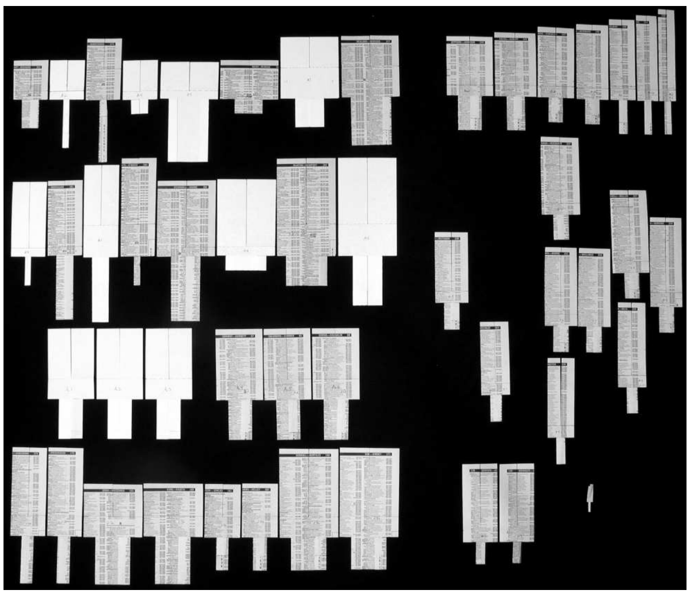

Figure 9: A photograph of most of the helicopters we made for our experiments showing the diverse range and combination of factor levels. The first two left rows are the 16 helicopters in the 2^{8-4} fractional factorial design screening experiment. The third left row are the 6 center runs. The bottom left row are the additional 8 helicopters used together with the whitepages helicopters in the first three rows for the 2^2 replicated design to determine the path steepest ascent. The top right row is the path of steepest ascent. The circular right group is the central composite design, the third and fourth steepest ascent helicopters (with shorter tails) were modified to compose the top-left portion of the central composite design. The bottom right gives 2 of the six optimum helicopters, with the "nanocopter" to the right.

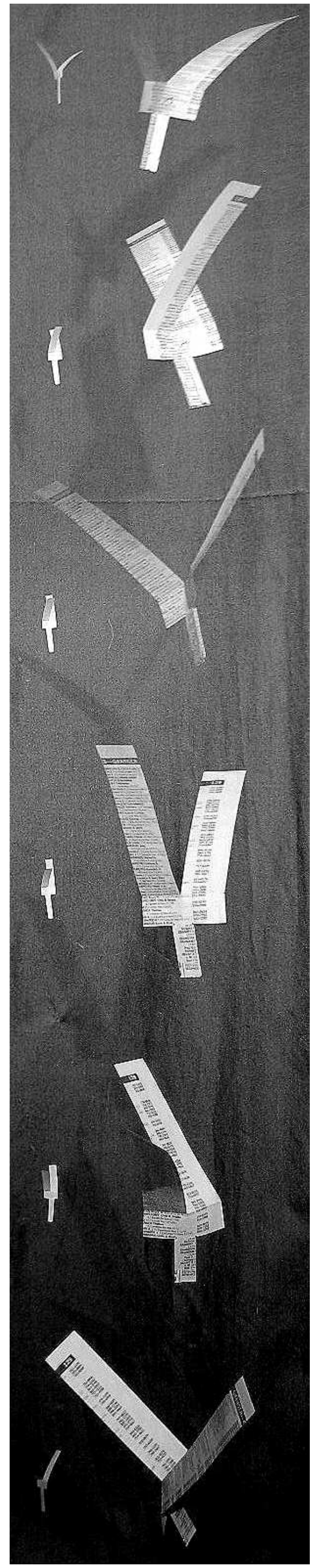

Figure 10: A strobe-series photograph of the "nanocopter" and an optimal helicopter in flight.

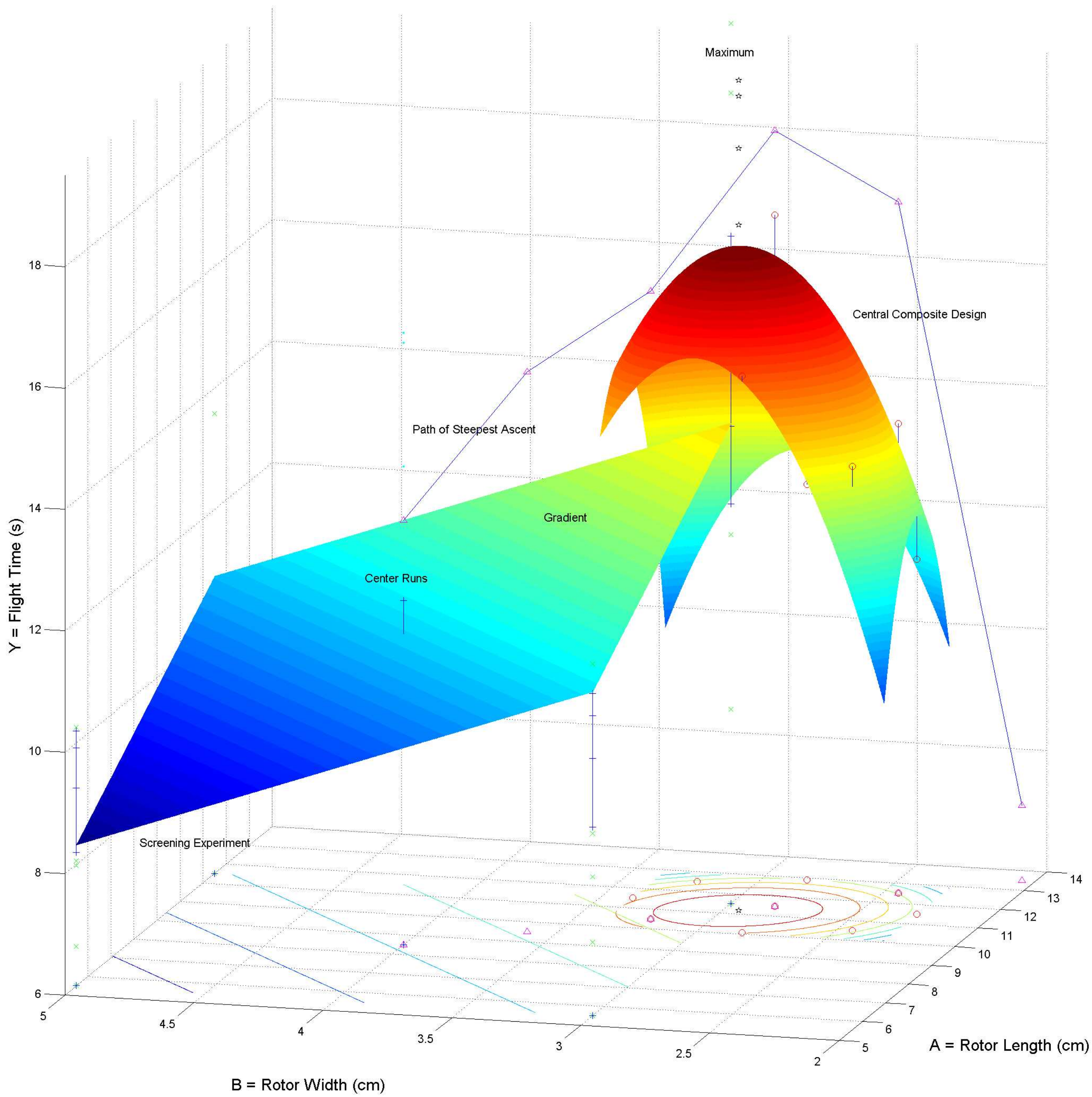

Figure 11: A perspective of all the experiments, their responses, and the response surfaces fit. Starting at the bottom-left, a screening experiment identifies the important factors influencing the response. Center runs are included to test for curvature. Since no curvature, we follow the gradient up the path of steepest ascent until the response sharply deteriorates. We run a Central Composite Design at the maximum of the path of steepest ascent. The CCD predicts a maximum, and we verify the predicted maximum response with replicated runs.