Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.2051

Published online 19 July 2016 in Wiley Online Library

Multiple Response Optimization for Higher Dimensions in Factors and Responses

Lu Lu,[a] Jessica L. Chapman[b] and Christine M. Anderson-Cook[c]*†

When optimizing a product or process with multiple responses, a two-stage Pareto front approach is a useful strategy to evaluate and balance trade-offs between different estimated responses to seek optimum input locations for achieving the best outcomes. After objectively eliminating non-contenders in the first stage by looking for a Pareto front of superior solutions, graphical tools can be used to identify a final solution in the second subjective stage to compare options and match with user priorities. Until now, there have been limitations on the number of response variables and input factors that could effectively be visualized with existing graphical summaries. We present novel graphical tools that can be more easily scaled to higher dimensions, in both the input and response spaces, to facilitate informed decision making when simultaneously optimizing multiple responses. A key aspect of these graphics is that the potential solutions can be flexibly sorted to investigate specific queries, and that multiple aspects of the solutions can be simultaneously considered. Recommendations are made about how to evaluate the impact of the uncertainty associated with the estimated response surfaces on decision making with higher dimensions. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: response surfaces; decision making; Pareto front; scalability; estimation uncertainty

1. Introduction

Balancing multiple responses to optimize a product or process is a common problem in a wide array of engineering and manufacturing fields.[1–7] Some strategies used to strike this balance, such as the desirability function (DF) approach of Derringer and Suich,[1] produce a single optimal solution by combining the multiple responses into a single metric that is based on subjective user inputs. On the other hand, a Pareto front (PF) approach for optimizing multiple responses (for example, refer to Refs. 8–11) identifies an objective set of solutions that are not inferior to any other solutions. Lu *et al.*[12] describe how to use Pareto frontiers as part of a two-stage decision-making process. This process first identifies an objective set of non-inferior solutions, which constitutes the PF, and then allows users to make a final selection that matches their priorities by strategically quantifying the impacts of subjective user inputs. The method was first proposed for generating computer-optimized design based on multiple criteria,[12] and then was successfully adapted for a variety of applications in design optimization[13–15] and reliability analysis.[16] A more general paradigm expands the approach to define–measure–reduce–combine–select (DMRCS),[17] which provides a structured roadmap for individual or team decision making. This approach can be thought of as being analogous to the define–measure–analyze–improve–control process for structured problem solving. The define and measure steps ensure that the optimization is over the right criteria with data on which the decision-maker can have confidence. The reduce step connects with the objective stage of the two-stage process, which eliminates non-contenders, while the combine and select steps involve using some of the tools and approaches from the second subjective stage to make measures on difference scales more comparable and to identify final choices that most closely match the users' priorities. Chapman *et al.*[18] illustrate how to use the two-stage PF approach in multiple response optimization to find the optimal operating conditions for a chemical process with two input factors and three response variables. This approach involves approximating the estimated response surfaces based on a grid of input values, identifying the Pareto frontier of non-dominated solutions, and then strategically incorporating subjective user choices to narrow down the set of possible solutions to a final decision on the best input value.

A consideration that often gets overlooked when simultaneously optimizing multiple responses is the uncertainty in the estimated parameters when fitting the response models, which can be affected by the shape of the response surfaces, the natural variability associated with the different responses as well as where the data were collected for fitting the models. Ignoring this uncertainty could potentially lead to sub-optimal input factor choices that do not perform well in practice and over-confidence about how reproducible

[a]*University of South Florida, Tampa, FL, USA*
[b]*St. Lawrence University, Canton, NY, USA*
[c]*Los Alamos National Laboratory, Los Alamos, NM, USA*
**Correspondence to: Christine M. Anderson-Cook, Los Alamos National Laboratory, Los Alamos, NM, USA.*
†*E-mail: candcook@lanl.gov*

the results and optimization are. Chapman *et al.*[19] describe how to incorporate this uncertainty into the PF decision-making process, and Chapman *et al.*[20] examine how the varied amount of variability in the response variables can impact the Pareto optimal solutions.

The graphical tools in the PF approach play a very important role in understanding the impact of user preferences and inputs, and hence are crucial in facilitating informed decision making.[12,18,19] However, the existing graphical tools are suitable for only relatively small numbers of input and response variables, which have limited the application and adaptation of the general methodology for larger scale problems. The main goal of this paper is to develop new graphical summaries that scale well to higher dimensions to equip the PF decision-making approach with more powerful tools that can help broaden its application in broader contexts. Because visualization in higher dimensions is often problematic, a key aspect of the new graphics is that the potential solutions can be sorted in multiple ways to investigate specific queries, and that several relevant aspects of the solutions can be simultaneously considered. The tools are currently implemented in R and are available from the authors by request. The sorting provides a way to link together different views of the same solution and to see patterns that would be otherwise difficult to extract. An alternative would be to use dynamic highlighting such as is implemented in the statistical software, JMP, which allows direct connection between graphs. While the methods presented work for larger numbers of input and response variables, we do recommend that decision-makers carefully think about the number of responses over which to optimize, as there can be increasingly severe trade-offs between choices and overall mediocre performance of many responses when a large number of responses are simultaneously optimized.

To illustrate the new graphical tools and demonstrate the improved scalability for higher dimensions, we revisit the tire tread optimization example described by Derringer and Suich.[1] The goal of the study, as specified in the define stage of DMRCS, is to find the ideal combination of three ingredients (hydrated silica, ζ_1; silane couple agent level, ζ_2; and sulfur level, ζ_3; all measured in parts per hundred) to optimize four characteristics of the tire (PICO abrasion index, y_1; 200% modulus [Mod200], y_2; elongation at break, y_3; and hardness, y_4). After the measure stage of the DMRCS process verified that the responses can be adequately measured, the objectives of the optimization are to maximize both PICO abrasion index and Mod200 and to achieve a target elongation of 500 and hardness of 67.5. The experiment data with 20 observations (reproduced in Table A.1 in the Appendix) were collected within the desired operating region with $\zeta_1 \in [0.38, 2.02]$, $\zeta_2 \in [33.6, 66.4]$, and $\zeta_3 \in [1.4, 3.2]$ to understand the relationship between inputs and responses; all input factors were coded (formula are given in Table A.1) to lie in the interval $x_i \in [-1.633, 1.633]$, $i = 1, 2, 3$. The data were generated from a three-variable, rotatable, central composite design with six center points.[1] Second-degree polynomial response surfaces were fit for each response variable, and the response surfaces fit via least square estimation appear in the succeeding texts.

$$\hat{y}_1 = 139.12 + 16.49x_1 + 17.88x_2 + 10.91x_3 - 4.01x_1^2 - 3.45x_2^2 - 1.57x_3^2 + 5.13x_1x_2 + 7.13x_1x_3 + 7.88x_2x_3$$

$$\hat{y}_2 = 1261.11 + 268.15x_1 + 246.5x_2 + 139.48x_3 - 83.55x_1^2 - 124.79x_2^2 + 199.17x_3^2 + 69.38x_1x_2 + 94.13x_1x_3 + 104.38x_2x_3$$

$$\hat{y}_3 = 400.38 - 99.67x_1 - 31.40x_2 - 73.92x_3 + 7.93x_1^2 + 17.31x_2^2 + 0.43x_3^2 + 8.75x_1x_2 + 6.25x_1x_3 + 1.25x_2x_3$$

$$\hat{y}_4 = 68.91 - 1.41x_1 + 4.32x_2 + 1.63x_3 + 1.56x_1^2 + 0.06x_2^2 - 0.32x_3^2 - 1.63x_1x_2 + 0.13x_1x_3 - 0.25x_2x_3$$

The desired operating region is chosen to be a spherical region centered at the origin of the coded values, with a radius of 1.633[1]. To approximate the response surfaces within the operating region, we can define a grid of locations that evenly span the space and estimate their associated response values using the fitted models mentioned in the preceding texts. Using a grid of 2301 points with coded factor values as multiples of 0.2, Table I summarizes the range of the estimated response values across the operating region as well as its size relative to the estimated variability of each response. For example, the estimated PICO values range between 50.5 and 242.2, with the range value being 34.2 times the estimated standard deviation of the process. This ratio of the range of response values to the estimated standard deviation can be considered as a measure of the signal-to-noise ratio, which captures the variability in the response values across the operating space calibrated by its natural variability (noise). A large signal-to-noise ratio indicates that the response changes considerably across the operating region relative to its noise. Hence, responses with larger signal-to-noise ratios are anticipated to have potentially larger influence on the shape of the PF when uncertainty is incorporated. Table I shows that, among the four responses, PICO and elongation have the largest signal-to-noise ratios at around 34, while Mod200 has the smallest signal-to-noise ratio at around 9. Therefore, it should not be surprising if the PICO and elongation have the largest impact on the formation of the PF, while Mod200 has the least influence due to estimation uncertainty.

Table I. Summaries of the range of estimated response values relative to the estimated variability as measured by $\frac{\max(\hat{y}_i)-\min(\hat{y}_i)}{\sqrt{MSE_i}}$ for the four responses in the tire tread optimization example from Derringer and Suich[1]

	Estimated response surfaces				
Response	$\sqrt{MSE_i}$	$\min(\hat{y}_i)$	$\max(\hat{y}_i)$	$\max(\hat{y}_i) - \min(\hat{y}_i)$	$\frac{\max(\hat{y}_i)-\min(\hat{y}_i)}{\sqrt{MSE_i}}$
PICO	5.611	50.5	242.2	191.7	34.2
Mod200	328.7	7.2	3010.2	3003	9.14
Elongation	20.55	153.3	846.9	693.6	33.8
Hardness	1.267	57.7	87.7	30	23.7

Based on the observed data in Table A.1, if we consider optimization of each response surface individually, we find that the first two responses, PICO and Mod200, tend to be best when all three input variables take on larger values. The elongation response tends to be near its target when the amount of silica is low, while the hardness response tends to achieve its target for low-to-moderate values of silane. When examining the estimated response surfaces separately, we clearly see that there is no single combination of input factors that is universally good for all four responses, and thus trade-offs and compromises are necessary to simultaneously optimize these four responses.

One of the key elements of the final three stages of DMRCS and PF two-stage decision-making processes is to use a structured approach paired with versatile graphical tools to facilitate informed decision making. This requires using efficient graphical summaries[12,13,16,18–20] to quantify and visualize the PF solutions as well as the contending choices with different possible user priorities. For multiple response optimization, visualization of the responses, the PF, and the contending candidate solutions becomes difficult even for moderate dimensions of both the response variables and input factors. For instance, in our case study, visualizing the different surfaces across the three-dimensional input space through contour plots requires many slices based on one of the inputs to gain an impression of the overall surface. This then needs to be repeated for each of the four responses. Hence, even the starting point of the analysis is tricky to visualize, and our goal is to offer suggestions about how to make understanding the optimization options available in the input space more manageable.

The remainder of this paper is organized as follows. Section 2 outlines the two-stage PF approach to decision making when the goal is to simultaneously optimize multiple response variables. The first objective stage of the decision-making process for the tire tread example, with new development in graphical summaries for higher dimensions, is illustrated in Section 3. Then, Section 4 illustrates more visualization tools developed for facilitating the structured subjective decision-making process in the second stage, with improved scalability. Section 5 contains some discussion and concluding remarks.

2. Multiple response decision making using a Pareto front approach

After carefully selecting the responses over which to optimize (define) and ensuring that adequate data are available (measure), the two-stage PF approach to decision making consists of an objective stage (which matches the reduce step of DMRCS) and a subjective stage (divided into the combine and select steps in DMRCS). The goal of the objective stage is to identify the most promising solutions – those on the PF – and eliminate all inferior solutions. A solution is said to be inferior (or dominated) if there is at least one other solution that has all criterion values at least as good as the inferior solution, with at least one that is strictly better. During this objective stage, it is helpful to understand the interrelationships between the response variables under consideration, as well as the dependent relationships between the input factors and the response variables, for the solutions on the PF within the desired operation/design region. Additionally, it is useful to investigate how estimation uncertainty affects the consistency with which locations appear on the PF. In the subjective stage, the decision-makers narrow down the set of contending solutions, based on their own priorities about the relative importance of the criteria and the performance of the contending solutions, until a final decision is made. We next describe key steps in both stages, and the new graphical summaries designed to be applicable to a larger number of response variables and a higher dimensional input factor space, in more detail before considering the illustrative example in Section 3.

After fitting the response models based on the initial data, the objective stage begins by defining a grid of locations in the operating region over which to approximate the response surfaces. A reasonable strategy for spacing the grid points in the operating region is to consider what level of practical precision there is for implementing solutions in the actual process for the input factor settings. It is important to choose this carefully, as grid points that are too close together result in too many locations, especially when there are a large number of input factors, and can increase computational complexity, while grid points that are too far apart may result in a poor approximation of the true response surface, and thus the selection of an sub-optimal solution. From this grid of locations in the operating region, the response surfaces are approximated based on the fitted models using the point estimates of model parameters, which are referred to as the mean models throughout the paper. Then, from the estimated response values from the mean models for all locations on the grid, the PF is identified. To keep the computational demand in higher dimensions manageable, it can be advantageous to use a coarser grid initially when examining the entire operating region then use a finer grid to fine tune the identified PF once inferior portions of the operating region have been eliminated.

After obtaining the PF based on the mean models, the objective stage explores general patterns between the inputs and responses for solutions on the PF as well as the trade-offs between different responses. To examine trade-offs among the variables, it is helpful to put all of the responses on a common desirability scale, scored from 1 for the best value of a response and 0 for the worst. There are many sensible ways to scale the responses to create these desirability scores; this could include the best and worst values for each response from the PF or from the entire grid of locations spanning the input factor space. To evaluate estimation uncertainty and keep the results consistent in their desirability scales, we choose to use the 99% confidence bounds across the operating space to determine the best and worst response values. More details on the scaling scheme are elaborated in Section 3.

Chapman *et al.*[19] describe two different figures to achieve the goals of stage 1. The first (Figure 1 of Ref. 19) allows the users to see what input factor values yield solutions on the PF. However, the limitation of this figure is that it works only for a two-dimensional space and is hard to adapt for cases with more than two input factors. The second figure (Figure 2 of Ref. 19) allows the users to examine trade-offs between the response values for the solutions on the PF, but it is displayed in a set of pairwise scatterplots, and the number of figures increases quickly as the number of responses increases. In this paper, we introduce a new and more

compact presentation that combines the information in those plots into a single, flexible graphical summary. This new graph allows users to visualize the input space, the desirability values for each response, and the trade-offs between the different responses for any number of input or response variables. This summary is flexible and informative because users can sort based on a number of different values, including the desirability values for any response or any input factor values, to get different views of the interrelationship between the responses and their dependent relationship with the input factors.

In addition to understanding the shape of the PF, the trade-offs between responses, and the relationship between the responses and input factors based on the mean model, we also want to understand the uncertainty in the estimated response surfaces, which as a result can lead to different PFs and possibly different solutions chosen as optimal. Chapman *et al.*[20] found that the set of PF optimal solutions can be influenced by the signal-to-noise ratio for the different response variables. It is important for realistic decision making to understand the consistency of the PFs in the presence of the estimation uncertainty as part of the optimization process. This can be evaluated via simulation, where numerous response surfaces consistent with the observed data are generated,[19] and for each set of simulated response surfaces, the PF is identified. Some useful graphical summaries, appropriate for a flexible number of response variables, are developed in this paper to quantify the uncertainty by examining the frequency with which locations appear on the simulated PFs.

Once we have developed our overall understanding of the contending solutions and eliminated inferior solutions, the decision-making process continues to the second, subjective stage of the process to hone in on a final decision. A DF will be used to rank the solutions for varied weighting preferences. Hence, the practitioner needs to choose an appropriate DF form (e.g., additive or multiplicative) to combine the multiple responses into a single summary score. The goal is to gain an understanding of which solutions are best for different weight combinations, as well as which locations are robust across many possible weights. A mixture plot[12,21] can be a very effective way to convey this information for scenarios with no more than three response variables. Lu and Anderson-Cook[14] adapt the mixture plot for four criteria by using slices of mixture plots at selected weights for one of the criteria. But it does not scale well for higher dimensions because it requires doing many slices at all selected weight combinations for a subset of criteria, and the number of figures can grow quickly as the number of responses increase. In this paper, we propose a new desirability–weight–input–volume (DWIV) plot that combines all the information more efficiently into a single figure that scales well to higher dimensions. From the DWIV plot, we can examine trade-offs between the different response variables and the relationship between the desirability of responses and the input factors and identify the range of weight combinations for which each solution is optimal, as well as the fraction of all weight combinations for which each solution is optimal. Because of the consistent sorting of the solutions across all of the different subfigures, the user is able to extract all meaningful summaries for a given solution of interest easily. The plot is designed to be straightforward to sort based on any of the inputs, responses, or locations in the design space. The DWIV plot can be very informative to help the decision-maker better understand which solutions are best for certain weight combinations and their associated input factor values and which solutions are robust across a variety of interested weight combinations.

With the rich information captured in the DWIV plot, the practitioners are now in a much better position to make a tailored final decision. There are two paths forward from this point in general: If the decision-makers have a priori preferences about the relative weightings of the responses, they should focus on solutions that match those preferences and narrow down from there to make a final decision; alternatively, if there is large amount of ambiguity about weighting preferences, a decision-maker could focus on solutions that are most robust across a variety of weight combinations. For the example in this paper, we take the former approach and use the DWIV plot to more closely examine solutions that perform best in a more focused weight range. Based on this result, we can further narrow our choices down to a couple of small regions in the input factor space. Once the location space of interest is reduced, we can examine these small regions using a finer grid for improved precision for approximating the response surfaces. In addition, estimation uncertainty can be evaluated using simulations, and its impact can also be incorporated into the DWIV plot by including uncertainty bounds on trade-off plots and examining the average fraction of weight combinations for which the solutions are optimal. This allows the decision-makers to develop a realistic sense about what range of desirability values are actually possible for the solutions that match their priorities, as well as how frequently those solutions are selected as best for some weight combinations. The decision-makers can then make a decision that takes into account the estimation uncertainty rather than simply using only the mean model, which can lead to sub-optimal solutions and over-confidence in the results. Once the decision-maker has a small number of contending solutions that match their priorities, the fraction of weight space (FWS) plot,[16,19,22] a dimension-free summary that quantifies the relative merits of individual solutions, can be used to aid in the final decision.

3. Improving scalability in the objective stage

This section describes the objective stage in the modified process, which aims to identify promising solutions using the PF and eliminate the non-contenders from further consideration. In addition, we want to explore the general pattern of the PF, including the trade-offs between responses as well as how the input factor settings in the design space impact the performance of different responses. Lastly, it is helpful to understand the consistency of the PFs in the presence of estimation uncertainty to facilitate making an informed decision.

In this objective stage, we take a two-step approach. In the first step, we use the mean models as an approximation of the true response surfaces and identify the PF for the four responses using a relatively coarse grid of locations that span the entire operating region. By using a grid of points with input factor values as multiples of 0.2 for the tire tread optimization example, we choose 2301

operating locations to get approximate estimates of the response surfaces using the mean models. A PF consisting of 872 locations was identified from the 2301 operating locations. For any location on the PF, there is no operating location in the grid of design space locations strictly dominating the location on the PF by having at least as good performance for all responses and strictly better performance for at least one response.

The top panel in Figure 1 illustrates the trade-offs between the four responses by showing the desirability scores (scored from 1 for the best to 0 for the worst) for the 872 points on the PF from the mean models. Four colors with different symbols are used to distinguish the responses. The locations on the PF are sorted based on their desirability scores for PICO (black), Mod200 (blue), elongation (red), and hardness (green), sequentially. As we discuss later, one of the flexible options for this plot is the ability to sort the values in the plot based on different summaries (such as responses, inputs factors, or labels for the locations). This allows the user to extract features and patterns based on their current priorities. As the dimension of either the number of responses or the number of input factors increases, the ability to generate flexible plots to dynamically explore the data based on different queries allows for improved understanding of the results.

The desirability scores are scaled using 99% confidence bounds to determine the best and worst response values. For each response, we calculate the 99% two-sided confidence interval for the mean response at all operating locations, which is given by $\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{0.005,n-p}\sqrt{MSE\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$ at a particular location $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})'$. In this formula, $\hat{\boldsymbol{\beta}}$ denotes the vector of the estimated model parameters, *MSE* is the mean squared error of the fitted response model, $\mathbf{X}$ denotes the model matrix of $n \times p$ dimension spanning over all model terms, and $t_{0.005,n-p}$ is the 99.5% percentile of the Student *t*-distribution with $n-p$ degrees of freedom. Then, the best and worst response values are determined based on the best and worst confidence bound values across the entire operating region. In particular, for PICO and Mod200, which we aim to maximize, we use the largest upper bound over all locations as the best value and the smallest lower bound over all locations as the worst value. For elongation and hardness, where we want to hit targets, the best value is selected as zero if at least one of the confidence intervals contains the target value. The worst value is selected as the largest absolute deviation from the target for all upper and lower bounds across all locations.

Several patterns can be observed from the top panel in Figure 1. First, comparing across all four responses, there is the least amount of trade-off between PICO and Mod200. As the desirability increases for PICO, the blue line for Mod200's desirability closely tracks the black line for PICO, with relatively little fluctuation. On the other hand, there is considerably more fluctuation in the desirability values of elongation and hardness as PICO's desirability increases. This indicates more trade-offs and less association between these two responses and PICO. PICO and elongation show the largest trade-off among all the responses. Second, the desirability scores show different ranges for the different responses. The elongation desirability values are spread between 0 and 1 for all points on the PF. The majority of desirability scores for PICO are between 0.3 and 1, with only a few locations having very poor

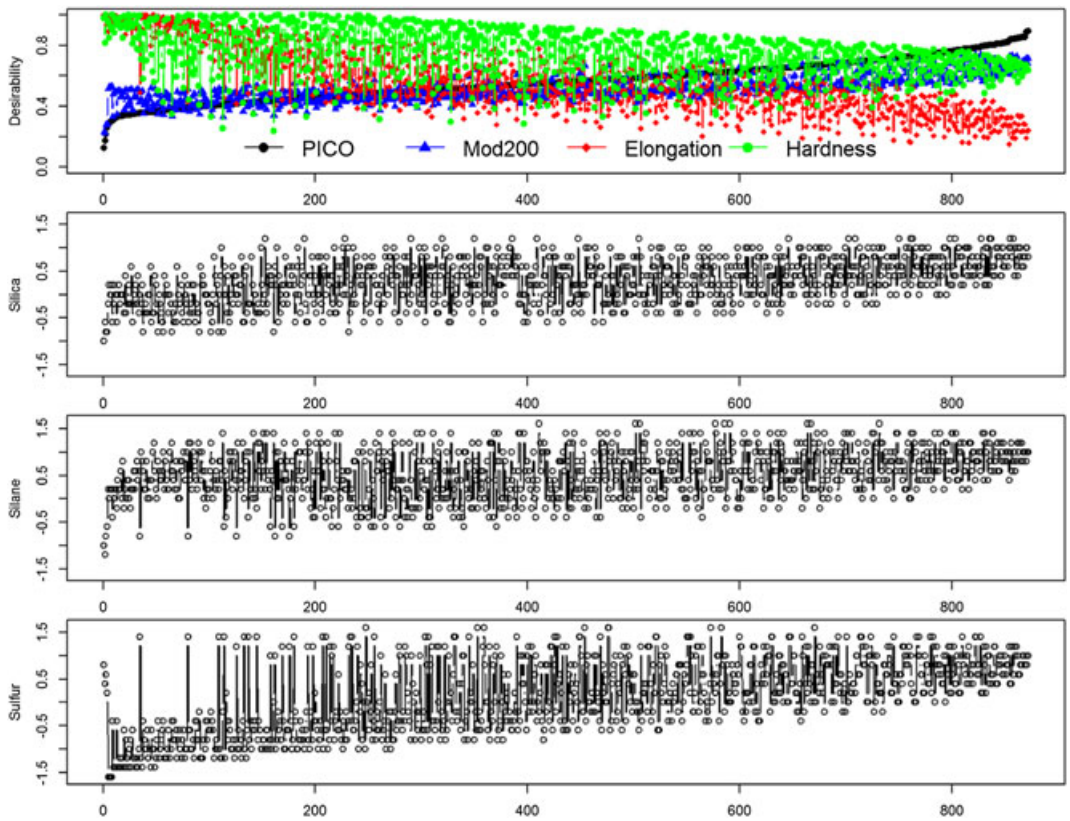

Figure 1. The trade-off plot with desirability values of the four responses stacked with input factor values for the Pareto front (PF) based on the mean models. The 872 designs on the PF are sorted by the desirability values for PICO, Mod200, elongation, and lastly, hardness. [Colour figure can be viewed at wileyonlinelibrary.com]

desirability scores below 0.3. The Mod200 range rarely achieves the extremes, with most desirability values between 0.2 and 0.7. In contrast, more points on the PF have relatively high desirability score for hardness, mostly above 0.4 and all above 0.2. Altogether, this indicates that the solutions on the PF either have good PICO but poor elongation scores or have poor PICO but good elongation scores, while the locations on the PF generally have mediocre Mod200 values, and the hardness scores vary considerably between fair to good. Therefore, by examining the trade-off plot, decision-makers can better understand the amount of trade-off required between the different responses. In this example, the most severe trade-off, which is the most important aspect for driving a decision, occurs between PICO and elongation. The Mod200 and hardness responses tend to have less influence on the final decision, because there are less pronounced trade-offs between these responses and the others and because of their good to moderate desirability scores across a number of the PF locations. This matches our intuition about the relative impacts of the responses based on their signal-to-noise ratios from Table I.

The bottom three panels in Figure 1 show the corresponding values for the three input factors for the 872 locations on the PF based on the mean models. This approach, with the input factor levels shown for each solution, builds a connection between inputs and responses, while allowing details on the operating location values to be extracted, even as the number of input factors increases.

Again, there are a couple of observations that can be made. First, most of the points on the PF have silica values above -1 on the coded scale, while the range of values for silane and sulfur covers the entire operating range between -1.5 and 1.5. This shows that the operating region with low silica values (<-1) is generally associated with non-contending solutions, and these locations should not be considered further. Second, higher silica and sulfur values are generally associated with higher PICO, but lower elongation desirability values. In addition, silane has a strong impact on hardness's desirability, with higher silane values being associated with lower hardness desirability scores. This can be better observed if we sort the locations primarily based on hardness (Figure B.1 in the Appendix).

Recall that one feature that the plot from Figure 1 offers is the ability to sort the desirability values based on different measures, such as inputs or responses, to allow for pattern extraction. Figure 2 shows an alternate view of Figure 1, where the locations on the PF are sorted according to the values of different input factors. The top panel is still the desirability scores for the four responses, shown with colors, symbols, and scaling consistent with Figure 1. In contrast to Figure 1, the locations in Figure 2 are sorted according to the input factor values based on silica, silane, and then sulfur. The obvious systematic pattern shown in the bottom three plots indicates how the locations are spread throughout the design space. While Figure 1 and other variations sorted according to the responses are helpful for showing the interrelationship between the responses, variations of the plot similar to Figure 2 allow us to see patterns related to the dependent relationships between the responses and the input factors.

More observations can be made to help us understand the response surfaces and also confirm findings from Figure 1. Because Figure 2 is sorted primarily based on silica, the features relating this input to the responses are most prominent. First, there is a

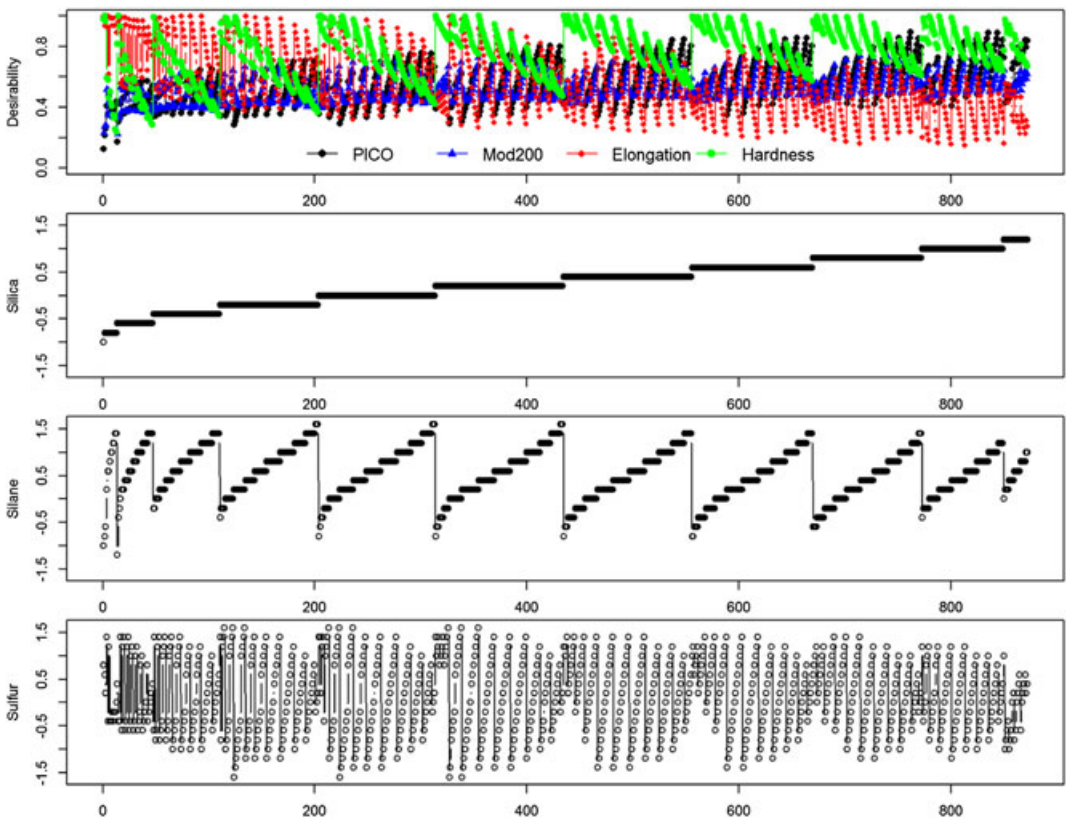

Figure 2. The trade-off plot sorted by the input factor values for silica, silane, and then sulfur. [Colour figure can be viewed at wileyonlinelibrary.com]

moderate positive correlation between silica and PICO's desirability and a negative correlation between silica and elongation's desirability. Second, as silane increases, there is a strong decreasing pattern in the desirability of hardness. By adjusting the order of the input factors based on which we sort the locations, we can also observe that the desirability of elongation tends to drop as sulfur increases (Figure B.2 in the Appendix). Also, if we fix silane, the desirability of hardness tends to drop as sulfur increases (Figure B.3 in the Appendix). This improved understanding of the relationships between the individual responses and the input factors allows us to better connect the desired response performance with the corresponding input factor regions and hence help decision-makers to quickly identify the desired regions once they have a better understanding of their goals and the relative contributions from each of the responses.

After understanding the general patterns of the PF based on the mean models for the responses and the connection to design space locations, next we want to understand the uncertainty associated with the identified PF. The uncertainty stems from the estimated response surfaces based on the observed data. This allows us to see how robust our decision based on the mean models is to the approximations of the response surfaces. To evaluate the uncertainty of the PFs, we did a simulation to get 1000 approximations of the response surfaces consistent with the observed data. The simulation used model parameters from a multivariate normal distribution, with the mean equal to the point estimates obtained by using least square estimation, and variance as the inverse of the estimated information matrix and calculated by $MSE(\mathbf{X}'\mathbf{X})^{-1}$. For each iteration of the simulation, the responses were calculated at all locations in the grid of 2301 locations using the generated model parameters for all four response models. More details for generating the simulated response surfaces are available in Chapman *et al*.[19]

For every set of simulated response surfaces, the PF was identified. Due to the uncertainty in the estimated model parameters, as well as the responses, the relative performance of the locations varies from simulation to simulation, and hence the locations on the PF can change. If the true response surfaces are known, then the true PF is only determined by the interrelationship of the response surfaces, and the user's choice on the fineness of the grid affects the level of the discrete approximation to the continuous true PF. However, when there is uncertainty, the relationship between locations can be distorted by the uncertainty of the estimated responses and how big it is compared with the differences in the response values. The former is affected by both the natural variability of the individual responses and the location of the input factors in the design space. The latter is affected by the shape and interrelationship of the response surfaces, as well as the coarseness of the grid the users chose to explore the operating region. For example, if the variation is large compared with the difference in the response values for nearby locations (a small signal-to-noise ratio), then all the locations tend to share the opportunity of appearing on the PF, and hence the identified PFs may vary a lot across simulations, and each location has a smaller frequency of appearing on the PF. But if the variation is quite small, then the simulated PFs are more consistent, and the dominating locations have a higher appearance frequency. Generally, the response surfaces with more curvature and smaller variation tend to have more consistent PFs, while the response surfaces with less curvature and larger variation can lead to a split in the adjacent locations in their role for contributing to the PF and hence more variation in the identified PFs. Also, the user's choice of the coarseness of the grid can also affect the relative magnitude of the change in the response values between adjacent values and the associated uncertainty at different locations. The use of a coarser grid tends to result in more consistent PFs, while the finer grid is usually associated with more competition between solutions in nearby locations and hence more variation in the identified PFs. However, due to the complexity of the interrelation between the different contributing factors for different applications, it is difficult to precisely anticipate the impact of uncertainty on the results without actually examining it through simulation.

Figure 3 summarizes the 'typical' performance of the 872 allocations on the PF based on the mean models across the 1000 simulated response surfaces as well as their frequency of appearance on the PF. The top panel shows the median desirability scores for all four responses over all simulated response surfaces, where the locations are again sorted based on PICO, Mod200, elongation, and then hardness. In general, we can observe similar trade-off patterns between the responses as in Figure 1, with slightly more local fluctuations. The most severe trade-off is still observed between PICO and elongation. Mod200 follows the trend of PICO, with slightly lower desirability for more desired values of PICO. The desirability of hardness generally has the largest fluctuation among the three responses other than PICO, and its fluctuation reduces for more desirable PICO values. The bottom panel shows the frequency of appearance on the PFs over the 1000 simulated response surfaces. There is a large amount of fluctuation in how frequently adjacent locations in the plot appear on the PF. The range of the frequency is between 0 to 80%, and there are many locations appearing on the PF in 40% or fewer of the simulations. There is a slight ascending trend in frequency toward the right side of the plot where PICO has higher desirability scores. As with Figure 1, it is possible to explore specific relationships and patterns by sorting the observations based on several factors, including the responses or the frequency of appearance on the simulated PFs.

Figure 4 compares the performance of points on the PF based on the mean models, with all locations evaluated in the entire design space in terms of the frequency of appearance on the PF across the simulated response surfaces. The horizontal axis in the plot shows different frequency thresholds ranging between 0 (never on the simulated PFs) and 1 (always on the simulated PFs). The vertical axis displays the proportion of points, with frequency of appearance above or equal to the certain thresholds. Closed circles with the solid curve denote locations on the PF based on the mean models, while open circles with the dashed curve are the entire collection of 2301 locations across the operating region. We can see that locations on the PF from the mean model have generally larger proportions with higher appearance frequency. However, this advantage diminishes as the frequency threshold increases to above 60%. This shows that the PF based on the mean models does have slightly more robust appearance performance across the simulated response surfaces. However, there is still a considerable amount of variation associated with the PF as estimation uncertainty is incorporated. While using the mean model is helpful to identify promising locations, it is clear that more exploration and graphical summaries are needed to help select a best overall location that performs well for the user's identified priorities in the presence of estimation uncertainty.

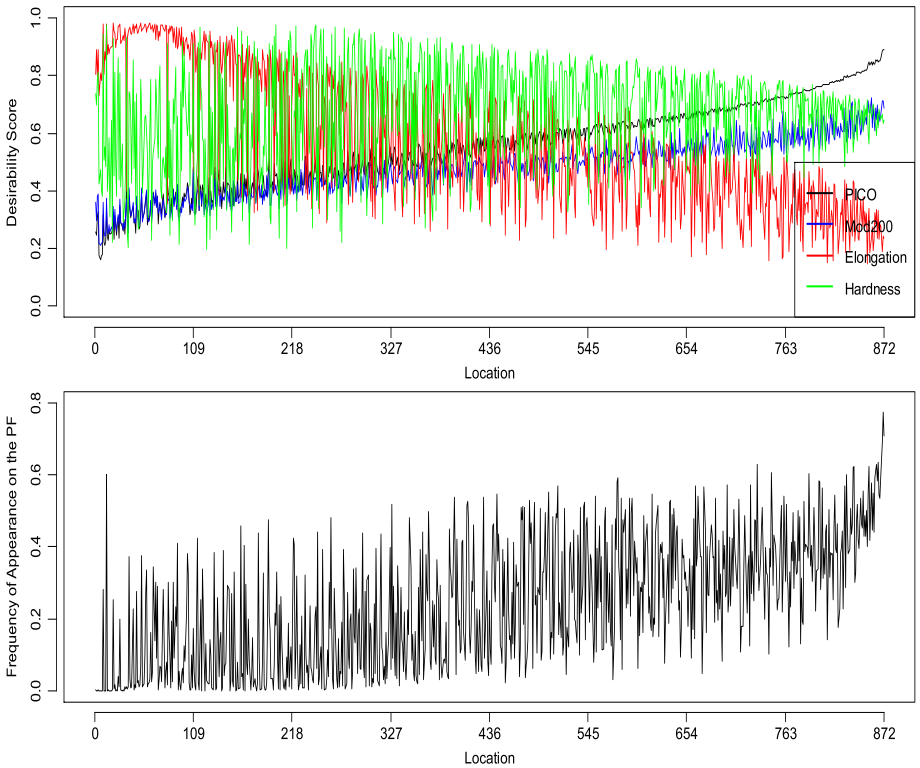

Figure 3. The top panel (a) the trade-off plot of the median desirability scores across 1000 simulations for all 872 locations that are on the mean model PF. The locations are sorted based on the desirability values, for PICO, Mod200, elongation, and lastly, hardness. The bottom panel (b) the frequency of appearance on the PF across the 1000 simulations. [Colour figure can be viewed at wileyonlinelibrary.com]

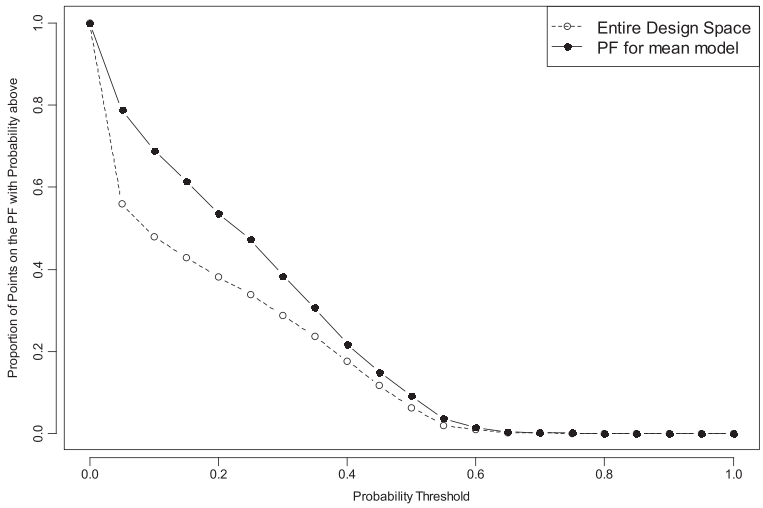

Figure 4. The proportion of points that appear on the PF with at least a certain probability threshold for (a) all 2301 solutions on the grid over the entire design space (shown with the open circles and the dash curve) and (b) the 872 solutions identified on the PF using the mean models (shown with the closed circles and the solid curve)

Considering how quickly the complexity of the decision increases as the dimensions of the problems increase, there are advantages to eliminating non-contending solutions based on the mean models in the objective stage to keep the decision-making process more manageable, but it is also important to consider solutions with good robustness to uncertainty. The practitioners need to be aware that the mean model approximation is not a panacea, and there still could be large amount of uncertainty in the identified set of promising solutions. In this example, the mean model approximation is a sensible compromise that was intentionally made to control the complexity and to help with the decision-making process. However, the uncertainty is still an important facet to understand when making a realistic decision, and we quantify its impact on final decision in the later stages of our structured procedure.

4. Improving scalability in the subjective stage

In the objective stage, we have reduced our choices from 2301 locations in the operating region down to the 872 locations on the mean model PF. In addition, we have gained an understanding about the interrelationship and trade-offs between the responses and how these are associated with different input factor values, as well as the impact of estimation uncertainty on the identified PF and its solutions. Next, we further narrow down the choices by incorporating user preferences and priorities. This is referred to as the subjective stage in our decision-making process. Due to the complexity of the problem and the large number of choices on the PF, a couple of structured steps are used to gradually narrow down our choices.

First, we want to further evaluate the performance of solutions on the PF for different weighting preferences to help identify some potential operating regions associated with more desirable performance. The performance of a solution for a particular weighting choice is evaluated based on an overall desirability score obtained by combining the multiple response desirability scores based on a chosen DF form. To make the analysis comparable to Derringer and Suich,[1] we chose to use the multiplicative DF of the form $DF_{\text{mult}}(j, \mathbf{w}) = \prod_{i=1}^{4} C_i(j)^{w_i}$, where $C_i(j)$ represents the scaled desirability score (between 0 and 1) for response i for location j, and $\mathbf{w} = (w_1, w_2, w_3, w_4)'$ is the weight vector with $\sum_{i=1}^{4} w_i = 1$ and $0 \le w_i \le 1$. We propose a new graphical tool that is easier to adapt for high-dimensional cases than the sliced mixture plot by Lu and Anderson-Cook.[14]

The DWIV plot, developed for this purpose, is shown in Figure 5. To produce this plot, we evaluate all the solutions found on the mean model PF at a fine grid of weighting choices, $\mathbf{w} = (w_1, w_2, w_3, w_4)'$, across the entire weighting space, and identify the optimal solution for each set of weights. Then, a subset of the top choices that are optimal for a larger proportion of weight combinations is selected, because this subset contains solutions more robust to user ambiguity about weighting preferences. The weights were chosen to have all entries as multiples of 0.02, that is, with 0.02 spacing for adjacent points on the grid. As a result, 23 426 sets of weight combinations were evaluated to find the optimal solution for each set of possible weights. The top 85 solutions that are optimal for at least 0.5% of the total weighting region (or ≥117 sets of weighting choices) were identified. Figure 5 summarizes key information for the selected 85

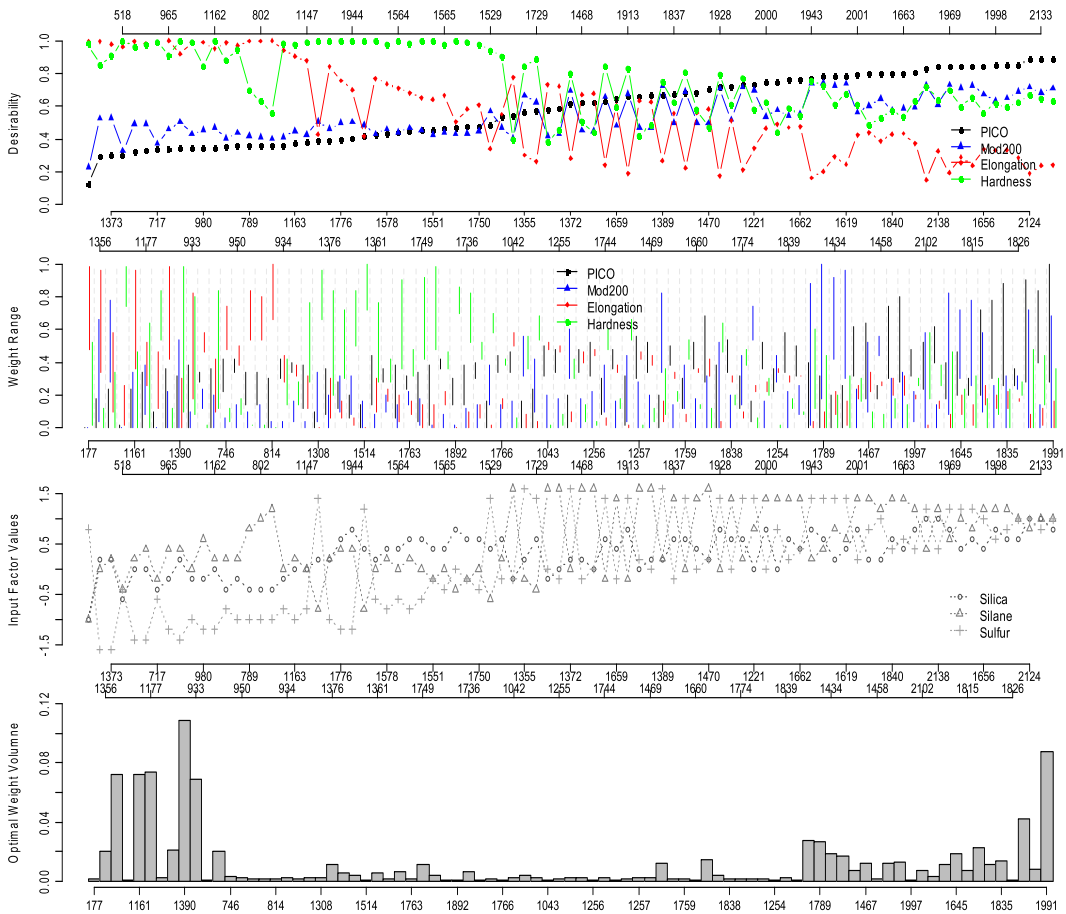

Figure 5. The desirability–weight–input–volume (DWIV) plot for the 85 solutions that are optimal for at least 0.5% of all possible weights based on using the multiplicative desirability function and the mean models. The subplots from the top to the bottom show (a) the desirability values of all four responses, (b) the range of weights for which a design has optimal desirability value for individual responses, (c) the input factor values, and (d) the fraction of volume of the weights for which a solution is optimal. [Colour figure can be viewed at wileyonlinelibrary.com]

solutions using four subplots. The top panel shows the scaled desirability scores for the responses. Similar to Figures 1 and 2, colors and symbols are used to distinguish different responses. The 85 solutions in all four panels are sorted by their desirability scores for PICO, Mod200, elongation, and lastly, hardness. Even though this subset contains only less than 10% of the locations on the entire PF, this plot resembles the general patterns observed in Figure 1. In addition, it is more obvious that the solutions seem to be divided into two groups. The first group on the left-hand side of the plot generally has relatively low PICO and Mod200 desirability scores but high elongation desirability scores and very good hardness desirability values. The second group on the right-hand side tends to have good PICO and fairly good Mod200 and hardness but poor elongation desirability scores. This is consistent with our understanding from the previous figures that the most trade-off occurs between PICO and elongation, which is a main factor in the final decision.

The second panel shows the weight ranges for all responses for which the solutions are optimal. We can see for each location that there are four vertical line segments shown in different colors; the colors of the line segments match the colors for the four responses in the top panel. Each line segment connects the maximum and minimum weights for which the solution is optimal for a particular response. For improved visualization, a gray vertical dashed line separates adjacent locations. The relative position of the line segment in the [0, 1] vertical space indicates where the optimal weights are located for each particular solution. For example, Location 117, on the left side of the plot, is optimal when elongation is weighted more than PICO, hardness is weighted less than half, and Mod200 is weighted less than 20%. Long line segments for all the responses indicate more robust performance for an optimal solution. Most importantly, it allows the users to visualize where optimal weights are located and how closely they match with user priorities.

The third panel in Figure 5 displays the input factor values for each of the 85 selected locations. Different symbols and shades of gray are used to distinguish the three input factors. This plot allows the users to directly connect their choices with the physical locations in the operating region. We can see a clear distinction between the two groups of choices with different response performance. The first group with lower PICO and higher elongation desirability scores is generally associated with low levels for sulfur, while the second group of locations with higher PICO and lower elongation desirability scores tends to be associated with high levels for sulfur. The silica values are generally around the middle levels, with the second group having slightly higher values than the first group. This matches with our observation from Figure 2 that higher silica and sulfur values tend to be associated with higher PICO but lower elongation desirability scores. The silane value has a strong negative association with the hardness desirability score. Hence, the second group has more locations at higher silane levels than the first group.

The last panel of Figure 5 is a bar plot showing the volume of the optimal weight region for each selected location. This gives direct quantification of the robustness characteristic for each solution. One interesting observation is that the solutions with more trade-offs between PICO and elongation tend to be associated with larger volumes of weights and hence better robustness, while the locations in the middle region of the plot which represent more balanced performance of PICO and elongation are less robust to the ambiguity in the specified weights. This information is helpful for choosing between individual solutions when they have similar performance within the user-preferred weighting regions.

Figure 5 is informative for showing general patterns as well as the individual performance of competing solutions. Again, because of the consistent sorting of the solutions across all of the different subfigures, the user is able to extract all meaningful summaries for any particular solution easily. This can help the users to make direct connections between the user priorities, the associated response performance, and the corresponding operating region. From Figure 5, the user can make choices that shape the remainder of the decision-making process. To arrive at a final optimal location, the user should (i) use a finer grid to explore more interesting regions to get a more accurate approximation of the response surfaces and (ii) examine the impact of the uncertainty associated with the estimated responses when the subjective user priorities are incorporated.

For scenarios when users have little idea of how to value and prioritize the responses, Figure 5, which examines the entire weighting space, can certainly help the users identify more interesting regions for further explorations. For scenarios when users have more focused weighting preferences, this figure can be adapted to match the more focused user's priorities. For the case study of Derringer and Suich,[1] the original analysis used equal weights for all the responses, that is, a weight vector $\mathbf{w} = (0.25, 0.25, 0.25, 0.25)'$ was used to optimize the DF. In reality, it is often hard for the practitioners to justify a very precise weighting preference, and it is common to have some ambiguity when determining the sensible weighting choices. Hence, we consider using a focused weight region similar to that proposed by Lu, Anderson-Cook, and Lin,[22] with approximately equal weighting to allow some extra flexibility. More specifically, our weighting region includes all weight vectors $\mathbf{w} = (w_1, w_2, w_3, w_4)'$ that satisfy $0.2 \le w_i \le 0.3$ and $\sum_{i=1}^{4} w_i = 1$. This corresponds to a tetrahedron at the center of the weighting space. We still use the grid of weights as multiples of 0.02, and there are 146 weight combinations to explore in this focused region. Note that the focused priorities should be determined based on an in-depth understanding of the particular application and all related factors. Without such knowledge or information, it is safer to consider a larger weight region to avoid making overly simplified subjective choices that could lead to sub-optimal solution.

Figure 6 shows another DWIV plot for the focused central weighting region chosen based on roughly equal weighting of all responses. As we zoom in, the number of optimal solutions reduces considerably. There are only seven locations optimal for at least one set of weights in this focused region. The first five locations are quite similar in response values as well as their location in the design space, while the last two are quite similar to each other but differ from the first five. From the top panel, the first five locations 1177, 1390, 1162, 1591, and 1376 all have high hardness and quite good elongation desirability scores, fair Mod200, and relatively poor PICO desirability scores. The last two locations, 1834 and 1835, are adjacent locations on the grid. Both have fairly good PICO, fair Mod200, and hardness desirability scores, and relatively poor elongation desirability scores. In the second panel, in addition to

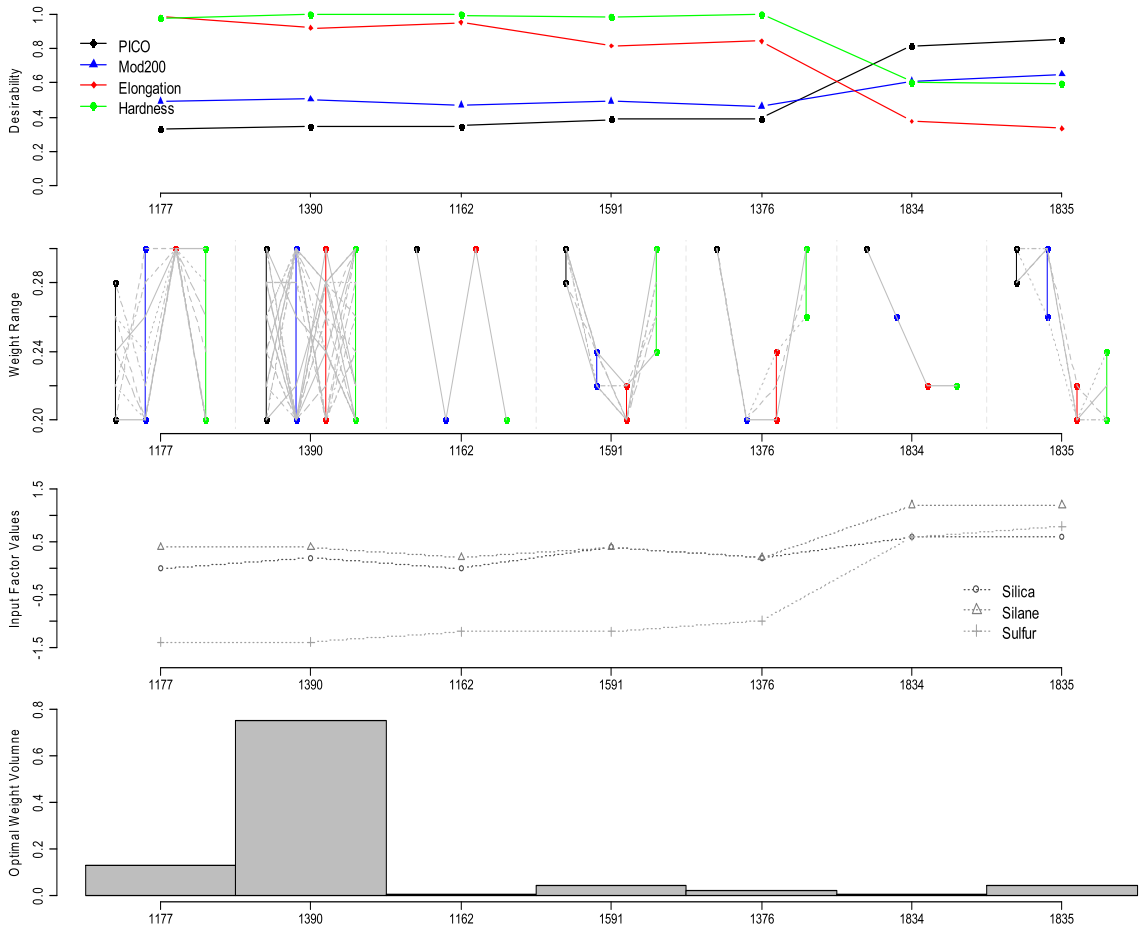

Figure 6. The DWIV plot for the seven designs that are optimal for at least one set of weights within a focused weighting region with $0.2 \leq w_i \leq 0.3$ for $i = 1, \cdots, 4$. In the weight range plot (the second from the top), the gray dash lines display a subset of representative weight combinations for which a solution is optimal. [Colour figure can be viewed at wileyonlinelibrary.com]

using the line segments for showing the general range of optimal weights for each selected location, there is also a set of gray dashed lines to showcase some representative weighting choices for which the particular location is optimal. From this, the user can extract some particular weight combinations. Location 1390 is optimal for a large number of possible weights across the entire focused weight region, with individual weight ranging between 0.2 and 0.3. Location 1177 is optimal for widely spread weights between 0.2 and 0.3 for Mod200 and hardness, and weight below 0.28 for PICO, but it is only optimal when elongation is weight close to 0.3. The longer line segments for all four responses for Location 1390 indicate more robustness, which is also evidenced by the larger volume (above 0.7) of the optimal weights shown in the bottom bar plot. The last five locations all have relatively short line segments in the second panel, which matches the small volume in the bottom plot. The third panel shows input factor values for the two design regions of interest in this focused region. The first region, labeled OR1, includes the first five locations on the left-hand side of the plot, which can be defined by $(x_1, x_2, x_3) \in [-0.1, 0.5] \times [0.1, 0.5] \times [-1.6, 0.8]$. The second operating region, labeled OR2, is based on the last two locations in the plot, which can be defined by $(x_1, x_2, x_3) \in [0.5, 0.7] \times [1.1, 1.3] \times [0.5, 0.9]$. From Figure 6, OR1 (particularly around Location 1390 with $(x_1, x_2, x_3) = (0.2, 0.4, -1.4)$) seems most promising, as it is optimal for the majority of the weight choices in the focused region. However, before making a final decision, it is important to consider how model parameter uncertainty may impact this choice.

Therefore, after identifying the two interesting small operating regions, we explore these regions with a finer grid, with the input factor values specified as multiples of 0.05 (four times finer in each dimension) to get a more precise approximation of the response surfaces. With the finer grid, we evaluate 1879 locations in OR1 and 201 locations in OR2. In addition, we include uncertainty to understand which locations have a higher probability of being optimal for more weighting choices in the focused region. Using the same simulation method, we simulated 1000 sets of response surfaces for the two interested operational regions using the finer grid of locations. For each simulated PF, we find the optimal location among the 2080 (=1879 + 201) locations for each set of 891 weight combinations in the focused weighting region (using a finer weight grid with 0.01 unit spacing). Then, we find the average optimal weight volume as the proportion of the weighting region for which each location is optimal averaged over the 1000 simulated PFs. This summary takes into account not only the robustness of a location's performance to the user specified ambiguity about the weights but also how frequent a location is considered optimal while the estimation uncertainty is incorporated. Based on this summary, we choose the top locations with the highest average optimal weight volume across the simulations, which represent greater robustness to both the weight ambiguity and estimation uncertainty.

Table II summarizes the average optimal weight volume and the corresponding input factor values for the top 20 solutions with the highest average optimal volumes. The closest match on the finer grid to the Derringer and Suich[1] solution, labeled Location 1660, is also added for comparison. First, all top 20 solutions come from the OR1. The highest ranking for locations from the OR2 is 225. This matches with our intuition from Figure 6 that OR1 is associated with more robust optimal solutions for the user's priorities. The top choice is Location 212, with $(x_1, x_2, x_3) = (0.2, 0.4, -1.5)$, which is optimal for 0.6% of the total interested weight region averaged over 1000 PFs. Note that the average optimal volume value is quite small compared with the optimal volume for the top choice based on the mean model (Location 1390 in Figure 6). This is because for many simulated PFs when this location was not best for any weight combinations, the optimal volume has a value of zero, which brings down the average volume across the simulated PFs. This happens quite often because as we use finer grid, more nearby locations with similar performance split the optimal volume as the response surfaces fluctuate across the simulations. This can be also observed from the close input factor values among the top choices in Table II. The bottom row of Table II shows the closest match to the Derringer and Suich[1] solution on the grid. This location (1660) is optimal for virtually 0% of the focused weighting region across the simulations, which indicates that it is unlikely to be an optimal solution for any approximation of the response surfaces consistent with the observed data. The two-stage PF approach has identified better solutions for optimizing the responses with near equal weighting.

Figure 7 shows an adapted DWIV plot to incorporate the model parameter uncertainty for the top 20 solutions selected based on the average optimal weight volume within the focused weight region. The closest match to the Derringer and Suich[1] solution, Location 1660 on the finer grid, is also added for comparison. To incorporate the associated uncertainty in estimation, the top panel shows the 2.5, 50, and 97.5% empirical desirability scores across the 1000 simulations for all responses. Due to the high similarity of the top solutions, the uncertainty bands for PICO, Mod200, and hardness are almost flat. There is slightly more fluctuation for elongation, but the overall difference in the desirability scores among the solutions is small. Location 1660 has similar PICO, hardness, and elongation desirability scores but generally smaller Mod200 desirability score. The second panel shows the range of optimal weights across the simulations for all responses at each location. This plot appears to be of very little interest, because all the top 20 solutions have optimal weights that cover the complete range of weighting choices between 0.2 and 0.3 from some of the simulations. This 'dull' pattern shows that estimation uncertainty can lead to all 20 top solutions being best across the user's interested weighting region. In contrast, Location 1660 is consistently not optimal for any weighting combinations within the region across the simulations. The third panel again confirms how similar the locations of the top solutions are in the operating region, particularly in the OR1. They all have very low values for sulfur, and slightly above middle level values for silica and silane. The best Derringer and Suich[1] solution has lower silica and silane values but a higher sulfur value. The bottom panel shows the average optimal weight volume across the simulations. The locations are ordered from the high-to-low rank based on this last panel.

Figure 7 shows that the top solutions identified using our structured procedure have consistent and robust performance with respect to weighting ambiguity and estimation uncertainty. For our particular user's priorities, they all represent similar degrees of trade-off not only between PICO and elongation but also somewhat with hardness and Mod200. All the selected solutions outperform the Derringer and Suich[1] solution based on the evaluation of the region surrounding equal weighting for all responses. The similarity of the top locations is reassuring that we have identified an operating region with overall optimal performance for the user priorities.

Table II. Summary of the top 20 solutions based on the Pareto front approach and the closest match to the Derringer and Suich[1] solution

Location index in finer grid	Average optimal volume	Silica	Silane	Sulfur
212	0.006659	0.2	0.4	−1.5
240	0.006301	0.35	0.5	−1.5
211	0.005976	0.15	0.4	−1.5
122	0.005811	0.2	0.45	−1.55
213	0.005596	0.25	0.4	−1.5
351	0.00547	0.15	0.5	−1.45
459	0.005376	0.35	0.45	−1.4
241	0.005296	0.4	0.5	−1.5
337	0.005118	0.1	0.45	−1.45
217	0.005088	0.45	0.4	−1.5
226	0.005068	0.25	0.45	−1.5
467	0.005066	0.1	0.5	−1.4
692	0.004905	0.3	0.45	−1.3
79	0.004644	0.1	0.25	−1.55
104	0.004642	0.25	0.35	−1.55
101	0.004552	0.1	0.35	−1.55
586	0.004457	0.2	0.5	−1.35
327	0.004389	0.25	0.4	−1.45
460	0.004384	0.4	0.45	−1.4
357	0.00436	0.45	0.5	−1.45
1660 (Derringer and Suich)	0.000000	−0.05	0.15	−0.85

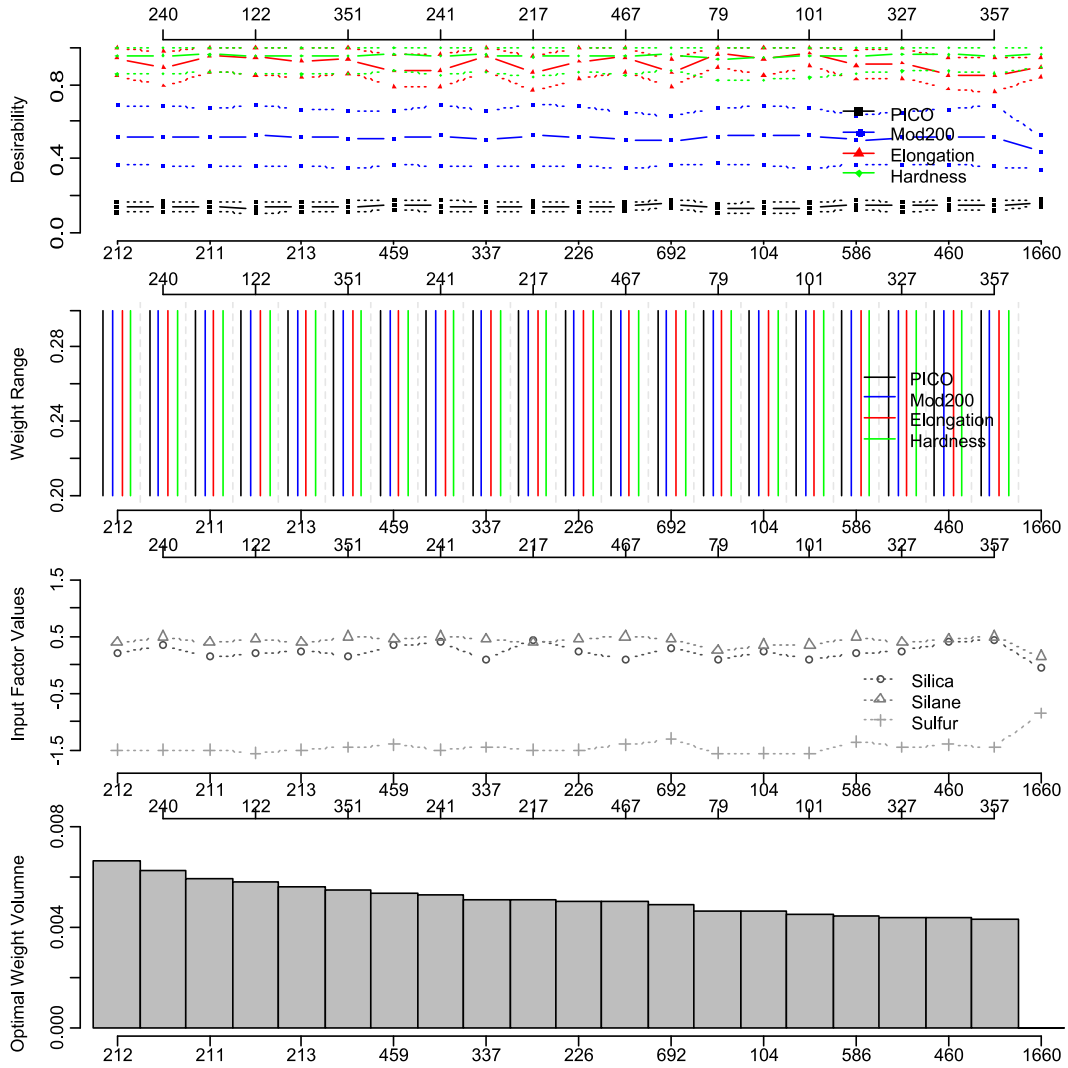

Figure 7. The DWIV plot for the top 20 solutions selected based on the average optimal weight volume within the focused weigh region with $0.2 \leq w_i \leq 0.3$ for $i = 1, \cdots, 4$ across 1000 simulations. Note that in the top desirability plot, the uncertainty bands with the 2.5 and 97.5% percentiles of the desirability values across the 1000 simulations are shown. The second weight range plot shows the range of optimal weights across 1000 simulations within the focused weighting region for which a solution is optimal. The last solution Location 1660 is the closest match to the Derringer and Suich[1] solution on the finer grid. [Colour figure can be viewed at wileyonlinelibrary.com]

Note from Table II that the top choice from the simulation while incorporating the uncertainty, Location 212 on the finer grid, has the input factor values $(x_1, x_2, x_3) = (0.2, 0.4, -1.5)$, and the top choice from the use of the mean model, Location 1390 on the coarse grid with $(x_1, x_2, x_3) = (0.2, 0.4, -1.4)$, is the closest match from the choices on the coarse grid to the top solution based on incorporating uncertainty. This similarity of results across the different stages is reassuring that the use of mean model approximation was a sensible simplification in the early stages when there were many aspects to consider and we needed some control to keep our task manageable. However, it is unlikely that the results based on the mean model and with uncertainty included will universally match. Hence, we recommend to always consider the impact of uncertainty whenever possible to avoid making a premature and possibly inferior decision. It is helpful to compare results between the different approaches to see if there is consistency or if different optimal locations are suggested, because an awareness of the similarity of results is helpful in the final decision making and in having confidence in the choice.

After finding a small set of contending solutions that match our priorities, an FWS plot, shown in Figure 8, can be used to compare the solutions based on their actual performance. The FWS plot[16] is a dimension-free summary for quantifying the relative performance of an individual solution to the best possible option at any weights summarized across all weight combinations of interest. The plot is convenient as a summary because it scales well across any number of criteria of interest and can be flexibly adapted to show the entire set of weights or only for a portion of weights for a focused region.[22] It was later adapted[18] to incorporate estimation uncertainty using uncertainty bounds. The relative performance of jth solution at the weight combination $\mathbf{w}$ was measured by the synthesized efficiency,[13] defined as $SE(j, \mathbf{w}) = \frac{DF(j,\mathbf{w})}{\max_j DF(j,\mathbf{w})}$. The relative performance is then summarized over the weighting region of interest to report what fraction of the interested weighting space the synthesized efficiency of a solution is above a certain threshold. The 95% uncertainty bounds obtained using the 1000 simulated PFs are added for quantifying the uncertainty.

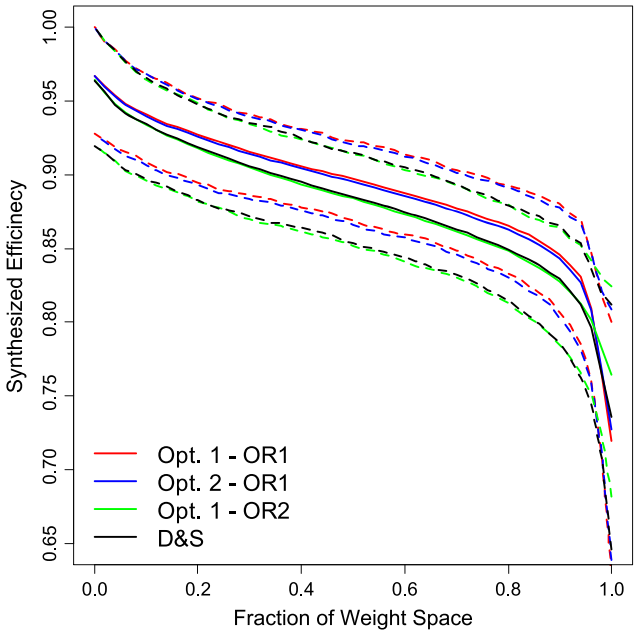

Figure 8. The fraction of weight space plot with uncertainty bounds summarized over 1000 simulations for four solutions: the top two solutions from OR1 (Locations 212 and 240 in Table II, shown in red and blue), the top solution from OR2 (Location 2062 ranked as 225 over the 2080 locations on the finer grid and shown in green), and the closest match to the Derringer and Suich[1] solution on the finer grid (shown in black). [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 8 compares four representative solutions for their relative performance in the focused weight region where $0.2 \leq w_i \leq 0.3$ for $i = 1, \cdots, 4$ for the four responses. The four selected solutions include the top two choices using our two-stage PF approach, that is, Location 212 with $(x_1, x_2, x_3) = (0.2, 0.4, -1.5)$ and Location 240 with $(x_1, x_2, x_3) = (0.35, 0.5, -1.5)$, which are both from OR1; the top choice from OR2 (Location 2062 with $(x_1, x_2, x_3) = (0.7, 1.2, 0.85)$, which is ranked 225 among all locations on the finer grid in both OR1 and OR2); and the closest match to the Derringer and Suich[1] solution on the finer grid (Location 1660 with $(x_1, x_2, x_3) = (-0.05, 0.15, -0.85)$). The top two solutions from our selection approach (shown in red and blue) have consistently higher synthesized efficiency across almost the entire focused weighting region (except for the very tail end near 100%). For more than 90% of the selected weights, their worst-case performance across all of the simulations is above 80% of the best possible performance at that weight combination. The top choices from OR2 (labeled in green) and the Derringer and Suich[1] solution (labeled in black) have similar performance to each other across most of the weight region. However, their curves are generally below those of our top two solutions, indicating consistent underperformance for most of the weighting choices within the interested region.

It is interesting to note that while the top two solutions outperform the other solutions, there is considerable overlap in the uncertainty bands from the four solutions. Hence, we are again reminded of the importance of considering the uncertainty associated with estimating the model parameters as we look to optimize the process. On one hand, the uncertainty shows that there is no clear universal winner when selecting a final set of input factor values. However, the positive part of this result is that we are reminded that there are multiple competitive solutions that each could be lead to good results when implemented. Including the uncertainty bands adds realism to the interpretation and comparison of results and what can be expected when a final set of input levels are determined.

5. Conclusions

As the number of input factors and responses over which to optimize increase, identifying a best solution for the priorities of a study becomes increasingly difficult. The DMRCS process with the embedded two-stage PF optimization approach provides a promising roadmap to structured decision making. The reduce step eliminates the non-contenders in the objective stage. In the combine and select steps, the user's priorities are incorporated through the use of a DF with user specified weights. Initially focusing on the mean model can be a practical strategy that allows the decision-maker to identify general patterns and trade-offs between the responses. As with the tire tread optimization example, it is important to understand which responses can be simultaneously near their optima as well as those which require difficult trade-offs to be considered when beginning to narrow the choices for a final solution.

Just as it is difficult to visualize larger numbers of raw response surfaces across a higher dimensional input space, understanding the characteristics of the PF in higher dimensions is also challenging. The trade-off plot with stacked input factor values allows details about the fronts to be easily examined. The consistent sorting of the solutions across all of the different subfigures allows the user to be able to easily see the different summaries for a given solution. To help with the interpretation of these plots, it is helpful to look at multiple versions where the user can choose how to sort based on the responses, inputs, or location numbers. The plots are flexibly

designed to allow for straightforward generation based on any desired sorting order. The graphics shown throughout the paper were generated using code written in R and are available by request from the authors. Considering several tailored versions of the plots with flexible sorting can partially help mitigate the curse of higher dimensions for visualization.

To gain an understanding of how uncertainty can impact the robustness of results, it is helpful to use the simulated response surfaces to show how the PFs change due to variability in the estimated response surfaces for the different criteria. Examining the frequency of appearance on the PF across the simulations shows patterns about the stability of the front. Different regions of the input space may be more consistently on the front than others.

During the second subjective stage of the optimization process, it is again helpful to initially focus on the results from the mean model to understand ranges of weights where different input factor combinations perform well. The DWIV plot is a compact summary that shows trade-offs between responses, weight combinations for which promising solutions are best, the location in the operating region, and the robustness of results across different weight combinations. Variations of which solutions to consider allow for great flexibility for understanding of results across all weight combinations or for specific ranges of weights of primary interest to the user. The DWIV plot can also be adapted in the later phase of the subjective stage to examine robustness of the results to estimation uncertainty.

The use of a coarse initial grid of input factor combinations and then subsequent refinement of the grid for regions of interest allows for both high-level understanding of patterns as well as details when narrowing down to a small set of choices. The sequential approach of initially using a coarser grid and then later switching to a finer grid also keeps the computational burden of higher dimensional input spaces more manageable.

Throughout the optimization process, balancing use of the mean model and an assessment of the role of estimation uncertainty can provide a beneficial view of promising solutions as well as realism about how these solutions might perform in any future implementation. Creatively combining multiple plots and examining several versions of the same plot sorted by different sources can all help with understanding the solution space and where it can be optimized for a particular user's needs. There is merit to thinking carefully about study priorities in order to limit the number of responses over which to simultaneously optimize. This can avoid having to compromise too much on any individual response. However, if either the number of input factors or responses becomes large, we have proposed some graphical and numerical summaries that can allow understanding and decision making based on data-driven methods that incorporate the uncertainty associated with the responses.

References

1. Derringer G, Suich R. Simultaneous optimization of several response variables. *Journal of Quality Technology* 1980; **12**(4):214–219.
2. Hari Krishna S, Chowdary GV. Optimization of simultaneous saccharification and fermentation for the production of ethanol from lignocellulosic biomass. *Journal of Agricultural and Food Chemistry* 2000; **48**(50):1971–1976.
3. Kim K-J, Lin DKJ. "Simultaneous optimization of mechanical properties of steel by maximizing exponential desirability functions. *Journal of the Royal Statistical Society: Series C: Applied Statistics* 2000; **49**(3):311–325.
4. Correia SL, Hotza D, Segadaes AM. Simultaneous optimization of linear firing shrinkage and water absorption of triaxial ceramic bodies using experiments design. *Ceramics International* 2004; **30**:917–922.
5. John RP, Sukumaran RK, Nampoothiri M, Pandey A. Statistical optimization of simultaneous saccharification and L(+)-lactic acid fermentation from cassava bagasse using mixed culture of lactobacilli by response surface methodology. *Biochemical Engineering Journal* 2007; **36**(3):262–267.
6. Ribeiro JS, Teofilo RF, Augusto F, Ferreira MMC. Simultaneous optimization of the microextraction of coffee volatiles using response surface methodology and principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2010; **102**(1):45–52.
7. Hasan HA, Abdullah SRS, Kamarudin SK, Kofli NT. Response surface methodology for optimization of simultaneous COD, NH_4^+–N and Mn^{2+} removal from drinking water by biological aerated filter. *Desalination* 2011; **275**:50–61.
8. Kasprzak EM, Lewis KE. Pareto analysis in multiple optimization using the collinearity theorem and scaling method. *Structural Multidisciplinary Optimization* 2001; **22**:208–218.
9. Gronwald W, Hohm T, Hofmann D. Evolutionary Pareto-optimization of stably folding peptides. *BMC-Bioinformatics* 2008; **9**:109.
10. Trautmann H, Mehnen J. Preference-based Pareto optimization in certain and noisy environments. *Engineering Optimization* 2009; **41**:23–38.
11. Costa N, Lourenco J, Pereira Z. Multiresponse optimization and Pareto frontiers. *Quality and Reliability Engineering International* 2012; **28**:701–712.
12. Lu L, Anderson-Cook CM, Robinson TJ. Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier. *Technometrics* 2011; **53**:353–365.
13. Lu L, Anderson-Cook CM. Rethinking the optimal response surface design for a first-order model with two-factor interactions, when protecting against curvature. *Quality Engineering* 2012; **24**:404–422.
14. Lu L, Anderson-Cook CM. Balancing multiple criteria incorporating cost using Pareto front optimization for split-plot designed experiments. *Quality and Reliability Engineering International* 2014; **30**:37–55.
15. Lu L, Johnson ME, Anderson-Cook CM. Selecting a best two-level 16-run screening design from the catalog of non-isomorphic regular and non-regular designs for six to eight factors. *Quality Engineering* 2014; **26**:269–284.
16. Lu L, Chapman JL, Anderson-Cook CM. A case study on selecting a best allocation of new data for improving the estimation precision of system and sub-system reliability using Pareto fronts. *Technometrics* 2013; **55**:473–487.
17. Anderson-Cook CM, Lu L. Much-needed structure: a new 5-step decision-making process helps you evaluate, balance competing objectives. *Quality Progress* 2015; **48**(10):42–50.
18. Chapman JL, Lu L, Anderson-Cook CM. Process optimization for multiple responses utilizing the Pareto front approach. *Quality Engineering* 2014a; **26**(3):253–268.
19. Chapman JL, Lu L, Anderson-Cook CM. Incorporating response variability and estimation uncertainty into Pareto front optimization. *Computers and Industrial Engineering* 2014b; **76**:253–267.
20. Chapman JL, Lu L, Anderson-Cook CM. Impact of response variability on Pareto front optimization. *Statistical Analysis and Data Mining* 2015; **8**:314–328.
21. Cornell J. Experiments with Mixtures: Design, Models, and the Analysis of Mixture Data (3rd edn). Wiley: New York, 2002.
22. Lu L, Anderson-Cook CM, Lin D. Optimal designed experiments using a Pareto front search for focused preference of multiple objectives. *Computational Statistics and Data Analysis* 2014; **71**:1178–1192.

Appendix

A. Original experimental data table

Table A.1. The experimental data from the Derringer and Suich[1] solution

Obs.	Silica $x_1 = \frac{\zeta_1 - 1.2}{0.5}$	Silane $x_2 = \frac{\zeta_2 - 50}{10}$	Sulfur $x_3 = \frac{\zeta_3 - 2.3}{0.5}$	PICO Y_1	Mod200 Y_2	Elongation Y_3	Hardness Y_4
1	−1	−1	1	102	900	470	67.5
2	1	−1	−1	120	860	410	65
3	−1	1	−1	117	800	570	77.5
4	1	1	1	198	2294	240	74.5
5	−1	−1	−1	103	490	640	62.5
6	1	−1	1	132	1289	270	67
7	−1	1	1	132	1270	410	78
8	1	1	−1	139	1090	380	70
9	−1.633	0	0	102	770	590	76
10	1.633	0	0	154	1690	260	70
11	0	−1.633	0	96	700	520	63
12	0	1.633	0	163	1540	380	75
13	0	0	−1.633	116	2184	520	65
14	0	0	1.633	153	1784	290	71
15	0	0	0	133	1300	380	70
16	0	0	0	133	1300	380	68.5
17	0	0	0	140	1145	430	68
18	0	0	0	142	1090	430	68
19	0	0	0	145	1260	390	69
20	0	0	0	142	1344	390	70

B. Supporting figures:

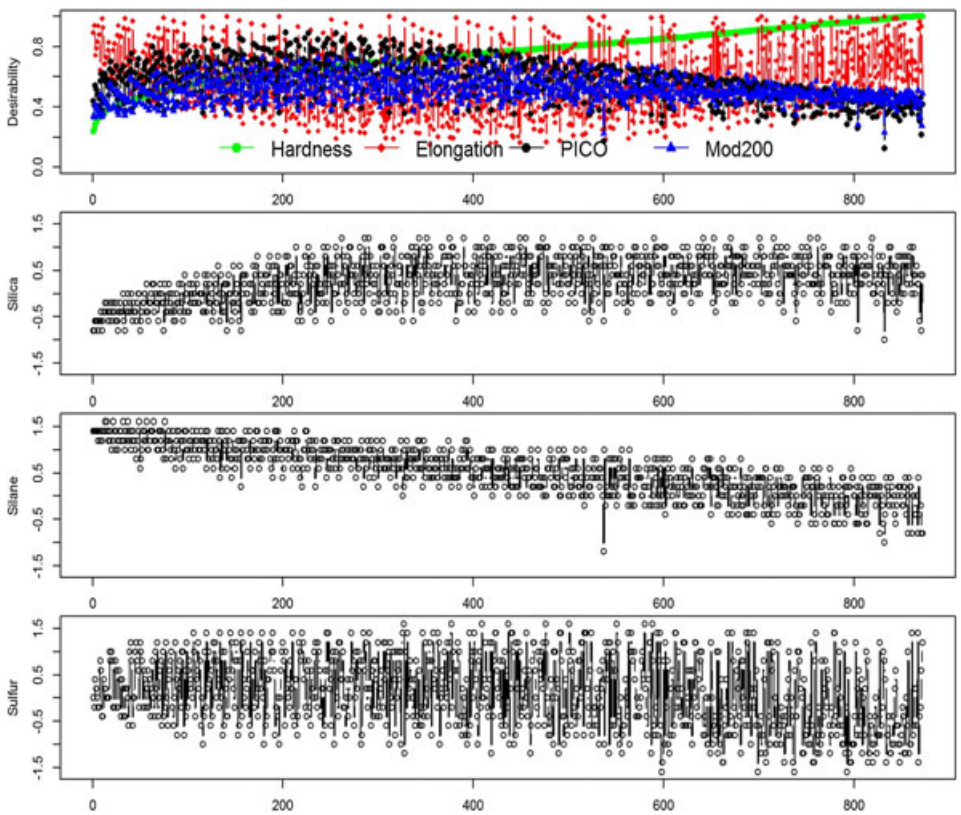

Figure B.1. The trade-off plot for the 872 locations on the PF of the mean models sorted primarily based on hardness. [Colour figure can be viewed at wileyonlinelibrary.com]

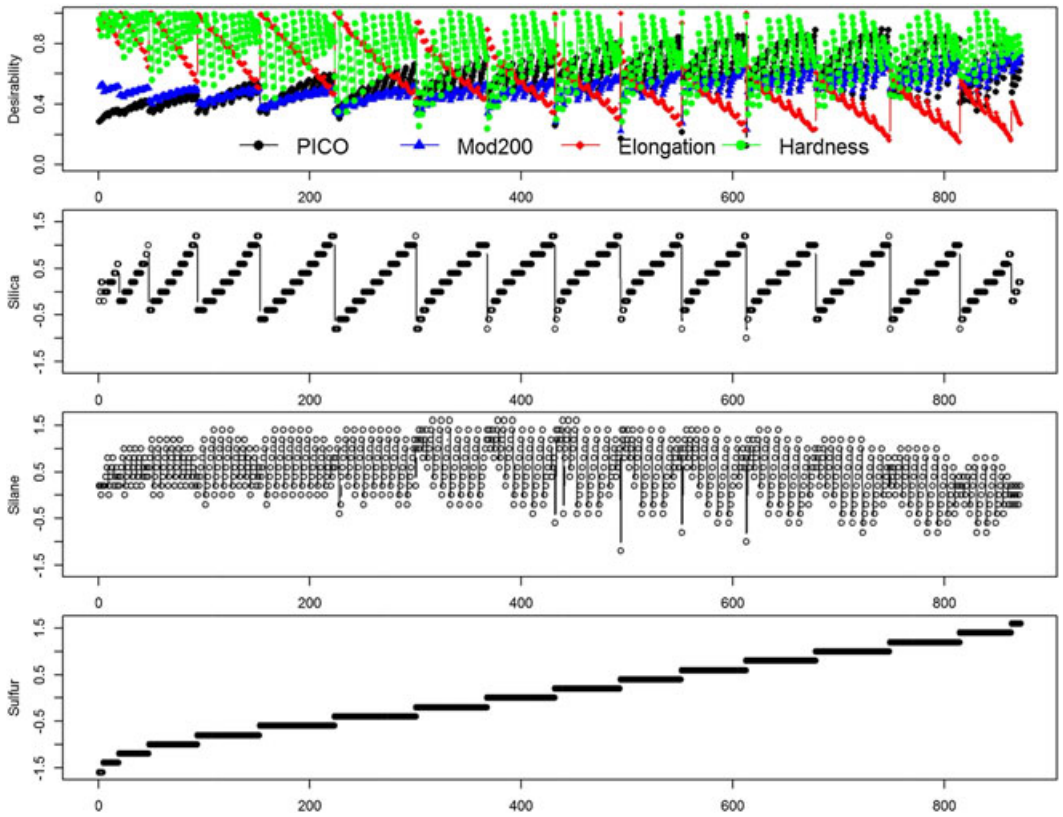

Figure B.2. The trade-off plot of the PF sorted primarily based on the input factor sulfur. [Colour figure can be viewed at wileyonlinelibrary.com]

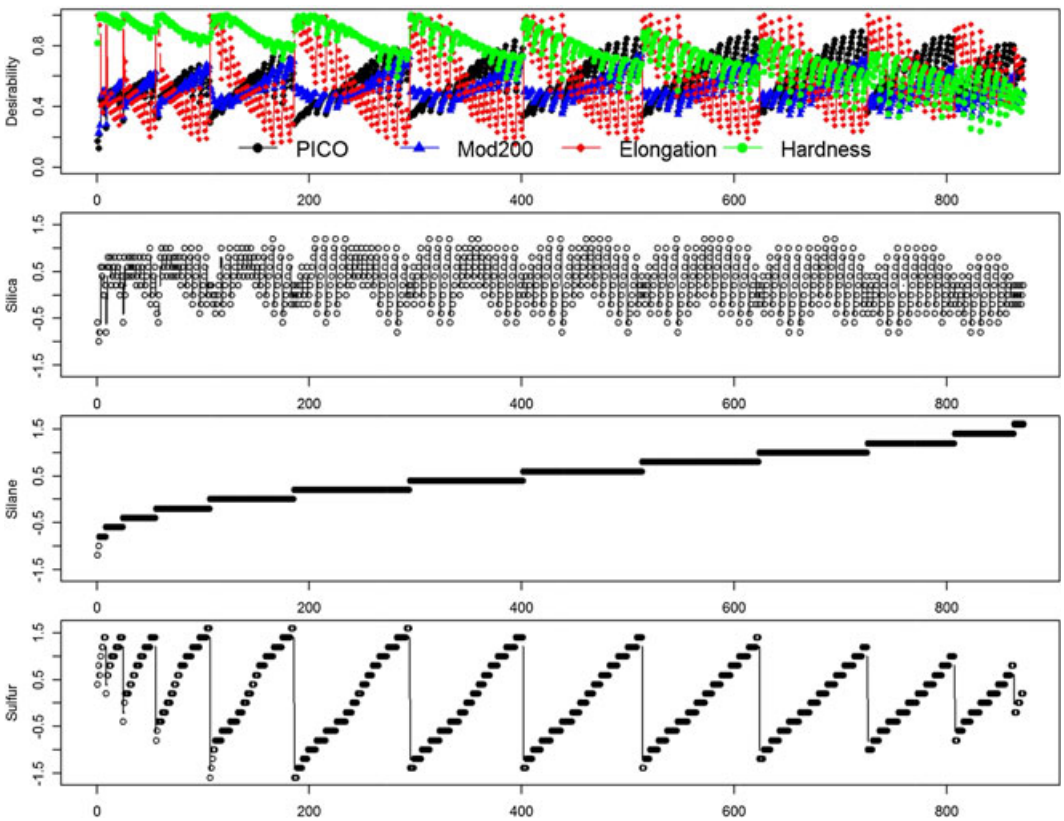

Figure B.3. The trade-off plot of the PF sorted primarily based on the input factor silane. [Colour figure can be viewed at wileyonlinelibrary.com]

Authors' biographies

Dr. Lu Lu is an assistant professor in the Department of Mathematics and Statistics at the University of South Florida. Her research areas include reliability analysis, design of experiment, response surface methodology, multiple objective optimization, and survey sampling.

Dr. Jessica L. Chapman is an associate professor in the Department of Mathematics, Computer Science, and Statistics at St. Lawrence University in Canton, NY. She earned a doctorate in statistics from Iowa State University. Her research areas include reliability, multiple criterion optimization, response surface methodology, and statistics education. At St. Lawrence University, she holds the Grace J. Fippinger '48 Professorship in the Sciences, an endowed position encouraging the success of women in the sciences.

Dr. Christine M. Anderson-Cook is a research scientist at Los Alamos National Laboratory. Her research areas include reliability, design of experiments, multiple criterion optimization, and response surface methodology. She is a fellow of the American Statistical Association and the American Society for Quality.