Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: http://www.tandfonline.com/loi/lqen20

Prioritization of stockpile maintenance with layered Pareto fronts

Sarah E. Burke, Christine M. Anderson-Cook, Lu Lu & Douglas C. Montgomery

To cite this article: Sarah E. Burke, Christine M. Anderson-Cook, Lu Lu & Douglas C. Montgomery (2017): Prioritization of stockpile maintenance with layered Pareto fronts, Quality Engineering, DOI: 10.1080/08982112.2017.1390585

To link to this article: https://doi.org/10.1080/08982112.2017.1390585

QUALITY ENGINEERING
https://doi.org/10.1080/08982112.2017.1390585

Taylor & Francis
Taylor & Francis Group

CASE STUDY

Prioritization of stockpile maintenance with layered Pareto fronts

Sarah E. Burke[a], Christine M. Anderson-Cook[b], Lu Lu[c], and Douglas C. Montgomery[d]

[a]Scientific Test & Analysis Techniques Center of Excellence, The Perduco Group, Wright-Patterson Air Force Base, Ohio; [b]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, New Mexico; [c]Department of Mathematics and Statistics, University of South Florida, Tampa, Florida; [d]School of Computing, Informatics and Decisions Systems, Engineering Department of Industrial Engineering, Arizona State University, Tempe, Arizona

ABSTRACT
Difficult choices are required for a decision-making process where resources and budgets are increasingly constrained. This article demonstrates a structured decision-making approach using layered Pareto fronts to identify priorities about how to allocate funds between munitions stockpiles based on their estimated reliability, the urgency of needing available units, and the consequences if adequate numbers of units are not available. This case study, while specific to the characteristics of a group of munitions stockpiles, illustrates the general process of structured decision-making based on first identifying appropriate metrics that summarize the important dimensions of the decision, and then objectively eliminating non-contenders from further consideration. The final subjective stage incorporates user priorities to select the four stockpiles to receive additional maintenance and surveillance funds based on understanding the trade-offs and robustness to various user priorities.

KEYWORDS
DMRCS; decision-making; multiple criteria; Pareto front; reliability; statistical engineering

Introduction

Many of us have participated in team decision-making meetings where sharply different priorities of each team member made reaching consensus difficult and personalities, not data, drove the decision. This paper presents a case study for a complex budget allocation situation among stockpile programs to enhance stockpile performance. The decision was particularly contentious since it involved allocating a budget to stockpile programs where the managers who stood to benefit from the choices were part of the decision-making process. Historically, decisions were based on which stockpile manager had better powers of persuasion. This case study illustrates an improved strategy using a structured decision-making process shaped by the Define-Measure-Reduce-Combine-Select (DMRCS) process, (Anderson-Cook and Lu 2015), to quantitatively balance difficult trade-offs and evaluate best choices based on multiple criteria. DMRCS uses elements of statistical engineering (Hoerl and Snee 2010 and Anderson-Cook et al. 2012a, 2012b) to combine statistical tools for solving complex problems as part of a data-centric approach. The details of the DMRCS process and related statistical tools are described specifically for the case study, but the general process for making decisions when team members have different priorities is one that occurs frequently across all industries.

We begin with some context for this particular decision-making process and how individual stockpiles are managed. Each year, choices are made about which stockpiles (populations of munitions) should receive additional funding to enhance their existing budgets and ensure that units are available and ready for their intended use. The current government fiscal environment limits funds and hence difficult choices must be made to effectively use limited resources. Potentially dire consequences can result from poor decisions as an unreliable or inadequate supply of units could endanger warfighters and compromise missions. Typically, each stockpile is comprised of units of different ages. The reliability of the stockpile is estimated and projected into the future using the population reliability as a function of the time from present after modeling the individual reliability as a function of age and/or

CONTACT Sarah E. Burke seburke89@gmail.com Scientific Test & Analysis Techniques Center of Excellence, The Perduco Group, 2950 Hobson Way, Wright-Patterson AFB, Wright-Patterson AFB, OH 45433.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.
Supplemental data for this article can be accessed on the publisher's website.

usage (Lu, and Anderson-Cook 2011a, 2011b). While statisticians are aware of the risks of extrapolation, this decision-making process requires managers to assess the future health of the stockpile (Collins, Anderson-Cook, and Huzurbazar 2011) to anticipate its readiness for projected usage. Generally, reliability is predicted with an associated uncertainty interval. When the lower bound of the predicted reliability is projected to cross a regulated threshold, units are deemed unreliable and pulled from service. Each stockpile has a specific threshold defined by the requirements for the munitions. Since the lower bound is used to determine when to pull units out of service, increased surveillance (more data collected) can help extend the service life by reducing the width of the uncertainty interval, and hence "raising" the lower uncertainty bound for the estimated reliability, even if the point estimate of reliability is unchanged.

Because of the proprietary nature of the data, it is not possible to examine the details of the original decision. However, a dataset with similar aspects has been constructed and the process used mirrors the actual procedure. In this case study, we consider a group of 42 small caliber stockpiles from four families (A with 15 stockpiles, B with 11, C with 11, and D with 5). These assets are relatively inexpensive compared to more complex systems, and the costs of maintenance/surveillance are similar across different stockpiles. The 5 member decision-making team includes four managers responsible for the families of stockpiles and the sponsor who provides the funds to be distributed. Since the available budget allows only 4 stockpiles to receive additional funding, the goal is to identify the top four stockpiles with the most critical needs for enhancing the health and quality of reliability estimation across several aspects of the stockpile.

Historically, the process for selecting the stockpiles was handled in a single meeting where the four stockpile family managers made their case for why some of their stockpiles deserved additional funds. After the meeting, the sponsor announced their choice. Unsurprisingly, managers presented arguments that changed from year to year and made the most compelling argument to maximize the chances of success for their assets. The meeting was highly contentious with clear winners and losers. The decision was often dictated by the effectiveness of the manager's presentation, rather than data or direct comparison of the stockpiles' needs. Based on frustration with the existing process, an alternate approach was implemented based on the DMRCS process (Anderson-Cook and Lu 2015).

The goal of this paper is to provide a detailed case study to illustrate the process for an important, high-impact problem with data-driven approaches advocated in the statistical engineering literature. The remainder of the paper is organized as follows: First, we provide background on the structured decision-making process and layered Pareto fronts. The next section explains how the necessary data were collected for the stockpiles and how the criteria were chosen to quantify the multi-faceted needs of the stockpiles. Next, we illustrate the objective stage of eliminating non-contenders with layered Pareto fronts, and then follow with a subjective stage of incorporating user priorities to select the four most critical stockpiles. Finally, we conclude with a discussion on assessing the robustness across some of the subjective choices and generalizing the process to other scenarios.

Background

DMRCS structured decision-making

Similar in spirit to the Six Sigma Define-Measure-Analyze-Improve-Control (DMAIC) (Hoerl and Snee 2012, p 128–137) approach, the Define-Measure-Reduce-Combine-Select (DMRCS) process provides structure for identifying and comparing alternatives based on multiple objectives. Table 1 shows a summary of the general elements included in each step of the process, as well as what was done for this case study. The *Define* step focuses on brainstorming and selecting the most important characteristics over which to optimize. As with DMAIC, this step is critical to ensure that the correct problem is solved. The *Measure* step also overlaps with DMAIC and focuses on using appropriate data and ensuring all key facets of the decision are characterized with representative metrics. The criteria for the comparison are ideally quantitative, of high quality with known pedigree (Snee and Hoerl 2012), and trusted by the decision-makers to allow fair and consistent comparisons among choices.

The *Reduce* step simplifies the choices in two ways. First, triage of the decision priorities focuses on a manageable number of dimensions. Similar to designing an experiment where many potential factors are initially identified before focusing on the most important ones, this allows preliminary consideration of many

Table 1. Summary of the Define-Measure-Reduce-Combine-Select (DMRCS) process.

Stage	General Steps	Steps in Stockpile Example
Define	- Identify choices under consideration	- 42 stockpiles were under consideration
	- Identify the aspects of the decision which are most important	- After initial brainstorming, two aspects of reliability, two aspects of urgency, and consequence were selected as key to the decision-making process
Measure	- Identify a quantitative metric suitable to characterize the aspects chosen	- Current reliability and time to threshold were defined for reliability, available supply and availability of alternate were defined for urgency, and a metric for consequence was defined
	- Gather the relevant data for each metric for all choices	- Subject matter experts defined what characteristics to define with each of the measures - Scores were assigned in [0,10] for all metrics for all choices
Reduce	- Eliminate some criteria from further consideration - Eliminate non-contending choices	- Some criteria of lesser importance or for which no good data were available were removed from consideration - A 4-layer Pareto front was constructed which eliminated 16 of the 42 stockpiles from further discussions
Combine	- Evaluate trade-offs between choices	- An additive desirability function based on scaling [0,10]→[0,1] for all criteria was used to rank choices (robustness also explored)
	- Incorporate subjective weighting of criteria for all team members	- Identified a universally agreeable sub-region of weights to select the final stockpiles for the Top 4
Select	- Identify top solutions - Explore performance of top choices relative to competitors - Finalize choices and how process can be defended to outside scrutiny	- Top 3 choices emerged as clear winners - Used synthesized efficiency plot and (N+1) plot to understand relative performance of final choices - Examination of final choices consolidation, choice of metrics, and process. - Discussion about whether 4 stockpiles could be expanded to include more in the future

potential choices before focusing on the right subset of criteria. The second type of reduction is to eliminate non-contending choices from further consideration. Constructing a Pareto front (PF) (discussed later in this section) is an objective and efficient way to achieve this goal.

The *Combine* step examines trade-offs between different criteria, which often represent diverse (and potentially conflicting) facets of the decision. Here, the priorities of the decision-makers and how much they value each criterion are keys to elevating the top choices. This step provides better understanding of the impacts of the different priorities for making an informed and justifiable decision.

Finally, the *Select* step identifies the top choices best suited to the decision-makers' priorities and provides tools for comparing close contenders. Both the Combine and Select steps involve a subjective element because they (1) involve making different criteria comparable and (2) include the decision-makers' relative emphasis of the criteria on the final decision. After the DMRCS process, the decision-makers have identified the top choices and can articulate why they are best. With experience in data collection, the impact of uncertainty on analysis results, and a system-level view of processes, statisticians have important and influential roles in the DMRCS process (Anderson-Cook, 2016).

While the details of each step will differ for each application, the process of focusing on what problem to solve, how to make it data-centric with suitable metrics, and how to select the best choices for the priorities of the team members will allow this approach to be flexibly used across different fields.

Desirability functions and Pareto fronts

To combine multiple criteria, a common choice for making diverse criteria (potentially measured on different scales) comparable is to use a desirability function (DF) (Derringer and Suich 1980). Each original criterion is assigned a desirability score, d_i in the interval [0, 1], where 1 is the most desirable and 0 is least desirable. This scaling allows maximizing, minimizing, or hitting a target to be handled with the same scoring system. The choice of how to map the original units to the desirability scores provides flexibility when comparing a given value for different DF scores. The desirability scores for all criteria can be combined into a single value for any weighting choice, using an additive or multiplicative structure:

$$Add\,DF_j = \sum_{i=1}^{k} w_i d_{ij} \quad [1]$$

$$Multi\,DF_j = \prod_{i=1}^{k} d_{ij}^{w_i} \quad [2]$$

The $w_i s$ are the weights ($0 \leq w_i \leq 1$, $\sum w_i = 1$) assigned to the k criteria and reflect the priorities of the decision-makers. For a particular set of weights, the best solution maximizes the value of the DF. However, the choice of the $w_i s$ is subjective, and each decision-maker may prefer different values of the $w_i s$. Different $w_i s$ generally lead to different rankings of the choices, and hence the DF approach alone may not be sufficient for reaching consensus for the team.

A Pareto front (PF) (Lu et al. 2011) identifies multiple solutions best for all the possible weights. A PF consists of all choices with values at least as good as any other choice across all criteria and a strictly better value for at least one criterion. The PF contains all solutions with maximum overall DF value for variations of the additive and multiplicative DFs in the form of L_p-norm for any p and across all $w_i s$ (Lu et al. 2011). Constructing the PF does not require specification of weight preferences, and hence is an objective summary of leading choices.

In this case study, the goal is not to identify just a single optimal solution, but the top 4 solutions from an enumerated list of candidates. To accommodate multiple solutions, the PF approach was adapted to include layered PFs for the top N solutions (Burke et al. 2016). The layered PF approach is well suited for one of two objectives: (1) given multiple quantitative criteria, identify the top N solutions to accomplish a task; or (2) make a decision by evaluating several primary quantitative criteria as well as secondary qualitative priorities. The first scenario matches our goals, where we are interested in finding a collection of critical solutions; namely, the top 4 stockpiles to receive additional funding. While the traditional PF looks for non-dominated solutions, layered PFs identify potential solutions that lie just behind the PF as candidates. While these solutions are never top choices, they could be highly competitive for regions of the $w_i s$ when multiple top solutions are of interest. For the top N solutions being sought, it is recommended to use N layers of the Pareto front (Burke et al. 2016).

The layered PF approach divides the choices into multiple layers of PFs with ranked solutions. The choices on the top layer PF are strictly better than those on the second layer PF and so on. Because the top N solutions for any choice of weights must necessarily be included in the top N layers of PFs (Burke et al. 2016), the N PF layers provide an objective set of superior choices before considering the subjective weighting choices. For the stockpile prioritization case study, any of the 42 stockpiles not on the top 4 layered PFs, can therefore be excluded from further consideration. This useful reduction in solutions to consider can make the decision-making process and discussions more manageable.

Expert elicitation of stockpile criteria scores

A key advantage of the DMRCS process is the early emphasis (in Define and Measure) on determining important characteristics and which criteria to use in the decision-making process. Long before individual stockpiles are discussed and compared, the decision-making team considers what would lead to a good decision. This upfront discussion of criteria plays an important role in grounding the decision-making in data and prevents early maneuvering by managers to sway the decision. The stockpile prioritization team began the exercise by brainstorming different facets of the stockpiles.

Historically, the estimated *reliability* curves were the primary focus, but even this metric had several aspects to consider. For example, for a population reliability curve plotted as a function of time from present (y-axis = reliability and x-axis = time from present), the vertical difference between the lower bound for current reliability and the threshold where action was mandated is of interest. Alternately, in the x-direction, the estimate time until the lower bound crossed the threshold is also relevant. Figure 1 illustrates both metrics for a generic population reliability curve with associated uncertainty. Both metrics had previously been used to justify a decision and the choice between the two generally depended on which made a more compelling case for the individual stockpile family manager.

A second category identified as important in the decision was *urgency*. The essence of this category was to evaluate the current supply of units available relative to projected needs. A critical stockpile here suggested that with projected usage, there would be a shortage in supply. In addition, for most stockpiles, established documentation exists for which alternative stockpiles can be substituted in the event of a shortage. Hence, another dimension to urgency was which other stockpiles might be used in the event of diminished supply. The more alternatives available, the less urgent it is to request maintenance or enhancement funds.

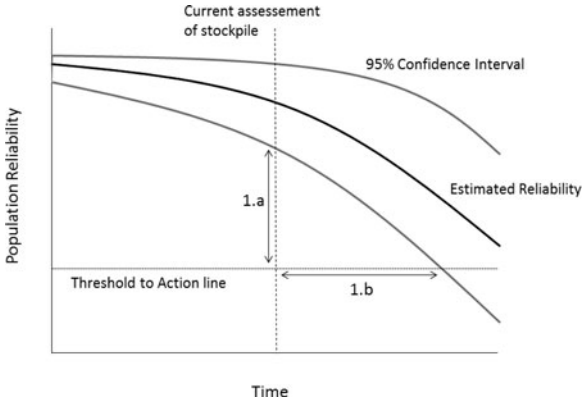

Figure 1. Sample population reliability curve with threshold to action line with the two reliability-based criteria illustrated. Current reliability (1.a) measures the distance between the lower bound of the estimated uncertainty interval and the threshold, while the time to threshold (1.b) measures the time until the lower bound of the uncertainty interval crosses the threshold.

The final category that emerged was *consequence*, which considered the impact on missions. This aspect was influenced by whether the current demand was for combat or training missions. Consequence can be a key factor in decision-making, especially when a combat situation is encountered.

Other aspects initially discussed, but not ultimately included were the number of historical failure problems, subject matter expertise on problems with usage, and the quality of the reliability testing procedures. In each case, after some discussion, these alternative criteria were removed from consideration. In some cases, these criteria were deemed less essential, while in others, getting objective, quantifiable data was nearly impossible.

In the Measure step, the definition of each characteristic specified in the Define step was made more precise, and the method for scoring was clarified. The team obtained data for each stockpile, and rigorous scoring criteria on a 0–10 scale (0 = least critical, 10 = most critical) was defined for the metrics:

1. Reliability (illustrated in Figure 1):
   - 1.a Current reliability – the difference between the threshold for action and the lower bound of estimated uncertainty interval for the reliability curve at the present time.
   - 1.b Time to threshold – the number of months until the lower bound of estimated uncertainty interval for the reliability curve is projected to cross the threshold to action value.
2. Urgency:
   - 2.a Available Supply – the degree of discrepancy between current supply and projected demand.
   - 2.b Availability of Alternate – the number and health of acceptable alternate stockpiles available as replacements in the event of a shortfall of units.
3. Consequence: Assessment of the impact if a shortage occurred based on training vs combat and how important the munition is in the combat missions. This criterion was the most difficult to quantitatively characterize, and involved considerable debate among the subject matter experts, in part because it was based on imprecise knowledge of the future. However, the stockpile managers ultimately were comfortable with the consistency of the assigned scores made by the subject matter experts.

The choice to use a score of 0–10 for each of the 5 metrics was based on historical precedent, and was based on comparing available data for each of the stockpiles to standardized definitions for each of the score values (Anderson-Cook 2013). The standardized definitions were established by subject matter experts (different than the decision-making team). Each stockpile was then assessed by several experts to obtain a final score for each metric. The fact that the scores were determined by the subject matter experts who do not directly benefit from the allocation of the funding and not directly involved with the decision-making process helps eliminate potential bias. By using experts from all aspects of the stockpile design, maintenance, and surveillance, the gathering and scoring of the data for each of the stockpiles was a labor-intensive process. However, there were additional benefits of transparency and consistency of assessment. After detailed evaluation and discussion, the scores in Table 2 were obtained for the 42 stockpiles.

Since the two metrics for reliability and the two metrics for urgency were complementary summaries of similar aspects, the decision was made to combine 1.a and 1.b into a single reliability summary, "Overall Reliability," and 2.a and 2.b into "Overall Urgency." In each case, the summary was a simple average of the two values. In the final section, we look at robustness of results to this choice. The three primary summaries (Overall Reliability, Overall Urgency, and Consequence) have different ranges:

Overall Reliability	Min = 5.75	Max = 9.5
Overall Urgency	Min = 6	Max = 9.5
Consequence	Min = 2.5	Max = 9.5

While the possible range for all the criteria was 0–10, the least critical (smallest) value for consequence was considerably lower than the other metrics. We later consider the impact of this.

Prioritization of stockpile maintenance with layered Pareto fronts

A portion of the Reduce step of DMRCS was already considered in the specification of metrics. Some stockpile characteristics initially identified were discarded from formal consideration. The second portion of the Reduce step seeks to objectively remove candidates with values not sufficiently critical to justify their inclusion. The use of the layered PFs allows for quick determination of leading choices. A JMP Add-In "TopN-PFS" (Top N Pareto Front Search) generated the top 4 PF layers. The JMP Add-In is available at https://community.jmp.com/t5/JMP-Add-Ins/Top-N-Pareto-Front-Search-for-Structured-Decision-Making/ta-p/36527.

Figure 2 shows a pairwise scatterplot of the top 4 layers, which by definition contain the top 4 solutions across all different prioritizations (Burke et al. 2016). As motivation for using more than one PF layer, consider stockpile A4 with criteria values (Overall Reliability, Overall Urgency, Consequence) = (OR,OU,C) = (9,8.75,9). It is not on the first PF, since stockpile A3 (with values (9.5,8.75,9)) dominates it; however, when compared to most other stockpiles, it has very high

Table 2. Data used to rank the 42 small caliber stockpiles based on reliability, urgency, and consequence. The columns in bold are the combined scores based on multiple measures within a category.

Stockpile	Current Reliability	Time to Threshold	**Overall Reliability**	Available Supply	Availability of Alternate	**Overall Urgency**	**Consequence**
A1	8.5	9.5	**9**	7	5	**6**	**2.5**
A2	7.5	5.5	**6.5**	7.5	10	**8.75**	**5.5**
A3	9.5	9.5	**9.5**	9	8.5	**8.75**	**9**
A4	9.5	8.5	**9**	8.5	9	**8.75**	**9**
A5	7	7	**7**	8	9	**8.5**	**4.5**
A6	7	7.5	**7.25**	8	8	**8**	**3**
A7	7.5	8.5	**8**	9	6.5	**7.75**	**3.5**
A8	7.5	9.5	**8.5**	8.5	6	**7.25**	**3.5**
A9	7.5	6.5	**7**	7	6	**6.5**	**3**
A10	8	7.5	**7.75**	8.5	9	**8.75**	**5**
A11	8	9	**8.5**	8	6	**7**	**3.5**
A12	7	6	**6.5**	8	5.5	**6.75**	**8.5**
A13	7.5	8	**7.75**	9	5	**7**	**5.5**
A14	8	9	**8.5**	7.5	8.5	**8**	**4**
A15	7.5	7	**7.25**	8.5	8	**8.25**	**3**
B1	6.5	8	**7.25**	8.5	6	**7.25**	**7**
B2	9	7.5	**8.25**	7.5	5	**6.25**	**7**
B3	7	6	**6.5**	6.5	6	**6.25**	**8**
B4	8.5	9	**8.75**	9	9	**9**	**9**
B5	8.5	8.5	**8.5**	7.5	8.5	**8**	**3**
B6	9	8	**8.5**	8	8	**8**	**3**
B7	6	7.5	**6.75**	8	8	**8**	**9.5**
B8	6.5	5.5	**6**	9	5.5	**7.25**	**4.5**
B9	9	7.5	**8.25**	7.5	8.5	**8**	**4.5**
B10	6.5	6	**6.25**	7	9.5	**8.25**	**6**
B11	9	7.5	**8.25**	9	6	**7.5**	**9**
C1	5.5	7	**6.25**	9	5.5	**7.25**	**4.5**
C2	8.5	9	**8.75**	9.5	6.5	**8**	**3.5**
C3	9.5	6	**7.75**	7	9	**8**	**9**
C4	6.5	5	**5.75**	8	5.5	**6.75**	**2.5**
C5	8	9	**8.5**	8.5	6	**7.25**	**6**
C6	8	8	**8**	7.5	9.5	**8.5**	**5**
C7	7.5	7.5	**7.5**	9	6	**7.5**	**7.5**
C8	8.5	6.5	**7.5**	9	5.5	**7.25**	**6.5**
C9	7	7.5	**7.25**	9	9	**9**	**4**
C10	8.5	6	**7.25**	7.5	5	**6.25**	**9**
C11	7	7.5	**7.25**	9	6.5	**7.75**	**8**
D1	7.5	7	**7.25**	8	5.5	**6.75**	**2.5**
D2	7	8	**7.5**	6.5	6	**6.25**	**9**
D3	10	8	**9**	8	6.5	**7.25**	**4**
D4	9	7	**8**	7	9	**8**	**9**
D5	6.5	6	**6.25**	9.5	9.5	**9.5**	**8**

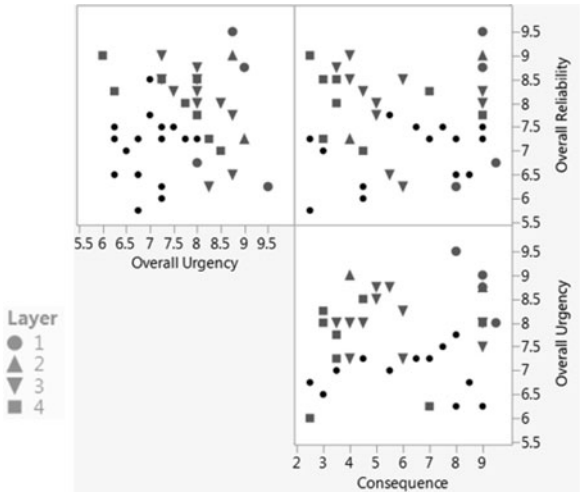

Figure 2. Top 4 Pareto front layers for the stockpile prioritization example. Each layer is identified with a different symbol and the most critical stockpiles have large values for all of the criteria and hence are located in the top right corner of each subplot. Stockpiles shown with black small dots are not on the top 4 layers.

scores for all three criteria and hence is likely to be among the top four choices for certain weights. Excluding these additional PF layers in the process when looking for multiple top choices could lead to flawed decision-making.

Table 3 lists the solutions in each PF layer (4 stockpiles on the top PF, 2 on second layer, 11 on third, and 9 on fourth). Stockpiles B5 and B6 have identical values for the three criteria and are both included in PF layer 4. In general, if two (or more) solutions are tied for all criteria, all are included. Without including these in the same PF layer, potential solutions could be missed in the decision-making process. Sixteen of the 42 stockpiles are not in the top 4 layers, and so can be objectively eliminated. The 26 remaining stockpiles (10 from A, 8 from B, 5 from C, 3 from D) continue under consideration. Hence, the Reduce step of the DMRCS process eliminated approximately 1/3 of the stockpiles as non-contenders. While the total number of choices is not huge for this problem, often the exhaustive list of alternatives can feel overwhelming. By constructing the layered Pareto fronts, some candidates can be eliminated as never being in the top N, which makes the problem feel more manageable and allows the team to make progress before embarking on the subjective phases of the decision making.

Table 3. Top 4 Pareto front layers for stockpile prioritization.

Stockpile	Overall Reliability	Overall Urgency	Consequence	PF Layer
A3	9.5	8.75	9	**1**
B4	8.75	9	9	**1**
B7	6.75	8	9.5	**1**
D5	6.25	9.5	8	**1**
A4	9	8.75	9	**2**
C9	7.25	9	4	**2**
A2	6.5	8.75	5.5	**3**
A10	7.75	8.75	5	**3**
A14	8.5	8	4	**3**
B9	8.25	8	4.5	**3**
B10	6.25	8.25	6	**3**
B11	8.25	7.5	9	**3**
C2	8.75	8	3.5	**3**
C5	8.5	7.25	6	**3**
C6	8	8.5	5	**3**
D3	9	7.25	4	**3**
D4	8	8	9	**3**
A1	9	6	2.5	**4**
A5	7	8.5	4.5	**4**
A7	8	7.75	3.5	**4**
A8	8.5	7.25	3.5	**4**
A15	7.25	8.25	3	**4**
B2	8.25	6.25	7	**4**
B5	8.5	8	3	**4**
B6	8.5	8	3	**4**
C3	7.75	8	9	**4**

The Combine step focuses on evaluating the trade-offs between contenders by leveraging the DF approach. The team considered several choices on how to weigh the different scores for reliability, urgency, and consequence. The first choice for the team was the scaling of the individual desirability scores. There are several possible choices to define the mapping to [0,1]: (1) use the natural range of each metric and map the 10 to a desirability score of 1, and the worst possible value (0) to 0; (2) use the range of the observed data (for example, for Overall Reliability, 5.75 → (maps to) 0, and 9.5 → 1)); or (3) use the range of data on the top N PFs layers (here for Overall Reliability, 6.25 → 0, and 9.5→ 1). The choice of scaling does not impact the PFs (i.e., the exclusion of non-contenders), but does impact the Combine step when scores from the different criteria are combined and compared.

Another important choice is choosing between the additive and multiplicative forms of the DF in Eqs. [1] and [2]. The multiplicative DF more severely penalizes low criterion values than the additive DF. For example, if one stockpile has a small score for any of reliability, urgency, or consequence, it is difficult for other criterion scores to overcome this poor rating.

The decision-making team deliberated on the merits of the different alternatives, and the majority agreed to use the natural range from 0 to 10 mapping to 0 and 1 on the desirability scale. The definition of consistent metrics across categories meant that a particular value (say, 5 on the natural scale or 0.5 on the desirability scale) would be comparable across metrics. The choice between using the additive or multiplicative

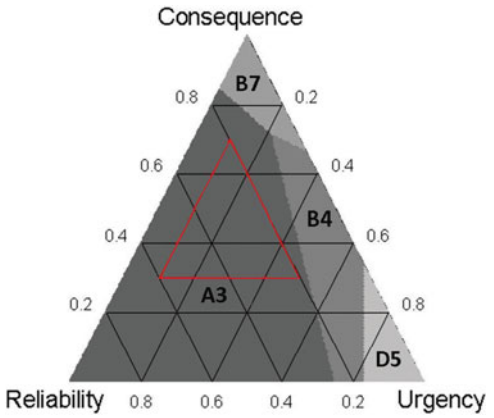

Figure 3. The mixture plot for showing the top choice for different possible weights of the three criteria, reliability, urgency, and consequence, considering only stockpiles on the top layer of PF. The red triangle shows the agreed upon sub-region of interest in the weight space requiring Overall Reliability ≥20%, Overall Urgency ≥10%, and Consequence ≥30%.

DF led to more vigorous debate, since whether a low score could eliminate a stockpile from consideration was not as clear. In the end, the consensus was to focus on the additive DF. The team did not want to completely exclude other scaling or DF form options. Hence in the discussion section, we examine other choices. In general, we recommend select a primary scaling and DF form, but also explore the impact of these choices on results before making a final decision. The use of layered PFs allows efficient evaluation of different scaling and DF choices by avoiding repetitive and unnecessary evaluations of non-contenders across different scenarios.

Figure 3 shows a mixture plot (Lu et al. 2011) for the top choice (those on the first PF layer) for all weight combinations using the additive DF scaled [0,10] → [0,1]. Regions close to a vertex emphasize a single criterion. Regions close to an edge emphasize the criteria for the adjacent vertices while down weighting the third criterion. Regions in the interior give non-zero weight to all three criteria. Only 4 stockpiles are most critical: Stockpile A3 with (OR,OU,C) = (9.5,8.75,9) is most critical for the majority of weight combinations (more than 70% of weights), particularly when reliability is weighted more heavily (≥25%) than the other criteria. Stockpile B7 with (OR,OU,C) = (6.75,8,9.5) is most critical when consequence is heavily weighted ($w_C \geq 70\%$), and D5 with (OR,OU,C) = (6.25,9.5,8) is top when urgency is highly prioritized ($w_{OU} \geq 70\%$). Stockpile B4 with (OR,OU,C) = (8.75,9,9) is top when reliability is down-weighted ($w_R \leq 25\%$), and urgency and consequence are weighted similarly. These results illustrate the benefits of the PF approach, with both stockpiles with top individual scores and those with balanced high scores are identified as critical.

Because the top 4 stockpiles are of interest for this decision, we now explore the additional information in the layered PFs. Figure 4 shows the top 4 choices for two slices of weighting choices from the mixture plot in Figure 3. Fixing the weight of consequence (w_C) to be 40% (0.4) matches the horizontal line labeled 0.4 on the left side of Figure 3. Consequence set at 50% corresponds to another horizontal line between 0.4 and 0.6 in Figure 3. The different shades of gray in Figure 4 indicate the top 4 choices (black = most critical, lightest gray = 4th most critical) for a given set of weights. First, note that the results are consistent with Figure 3. For example, when $w_C = 0.4$, A3 is the top choice for w_{OR} between 0.15 and 0.6, and B4 is the top choice when w_{OR} is between 0 and 0.15.

The summaries in Figure 4 allow exploration of the top 4 choices for any weight combination. Suppose that one team member chooses (w_{OR},w_{OU}, w_C) = (0.35,0.25,0.4). In this case, the top 4 choices (in order) are stockpiles A3, A4, B4, and D4. For weights (w_{OR},w_{OU},w_C) = (0.3,0.3,0.4), there is a tie for second rank between A4 and B4. Similarly, there is a 2-way tie for 4th most critical stockpiles between B11 and D4 when (w_{OR},w_{OU},w_C) = (0.4,0.2,0.4). The JMP TopN-PFS Add-In allows dynamic exploration of the 4 most critical stockpiles over varying weights by selecting different fixed weights for one criterion and then looking across the possible weights of the remaining two. Stockpiles A3, A4, and B4 are most often among the top 4 choices for both w_C values.

In addition to exploring individual weights, the team was also interested in the robustness. Figure 5(a) shows the proportion plot from the TopN-PFS Add-In that summarizes how often different stockpiles were identified in the top 4 across all weight combinations. The stockpiles are sorted from most frequently in the top 4 to least, and all stockpiles that appear anywhere in the top 4 for any weight combination are shown. Although the team started with 42 stockpiles (with 26 in the top four PF layers), only 14 were ranked in the top 4 for this scaling and DF choice. Different shades of gray again indicate the ranking achieved by the stockpile. As with Figure 3, note that only stockpiles A3, B4, B7, and D5 are ever shown as the

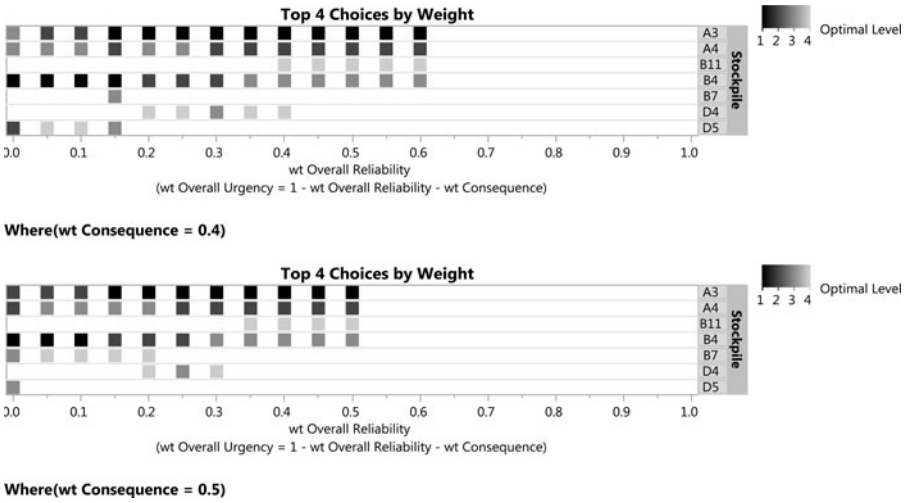

Figure 4. Sample mixture plot summary for top 4 choices when Consequence is weighted 40% and 50% using an additive DF and scaling [0,10]→ [0,1]. (Top) Where (wt Consequence = 0.4); (bottom) Where (wt Consequence = 0.5).

top choice (black bars in the stacked proportions) in Figure 5(a).

This plot shows that three stockpiles (A3, B4, A4) are in the top 4 for almost all the possible weights. Stockpile A3 is most critical for approximately 70% of all weights, and always in the top 4. Stockpiles B4 and A4 are in the top 4 for over 95% of weight combinations. This plot highlights the importance of including additional layers of the PF, as A4 (on the second PF layer) would not have been considered as a solution using only the top PF layer. When the team saw these results, it was clear that A3, B4, and A4 should receive additional funding. Seeing the right summary made some of the decision-making easy and non-controversial.

Deciding on the 4th stockpile was more difficult. From Figure 5(a), the 4th through 7th place stockpiles in the top 4 were not easy to distinguish. For example, both B11 and D5 were in the top 4 for 25% of the weights. The next step for the decision-making team was to identify a universally agreeable range of weights. Identifying a focused region that matches decision-maker goals can narrow the search (Lu et al. 2014). None of the team thought eliminating any criterion (i.e., setting a weight to zero) was appropriate. After some discussion, the team reached a consensus on the sub-region with Overall Reliability ≥20%, Overall Urgency ≥10%, and Consequence ≥30% (the overlaid triangle in Figure 3). All the team members agreed

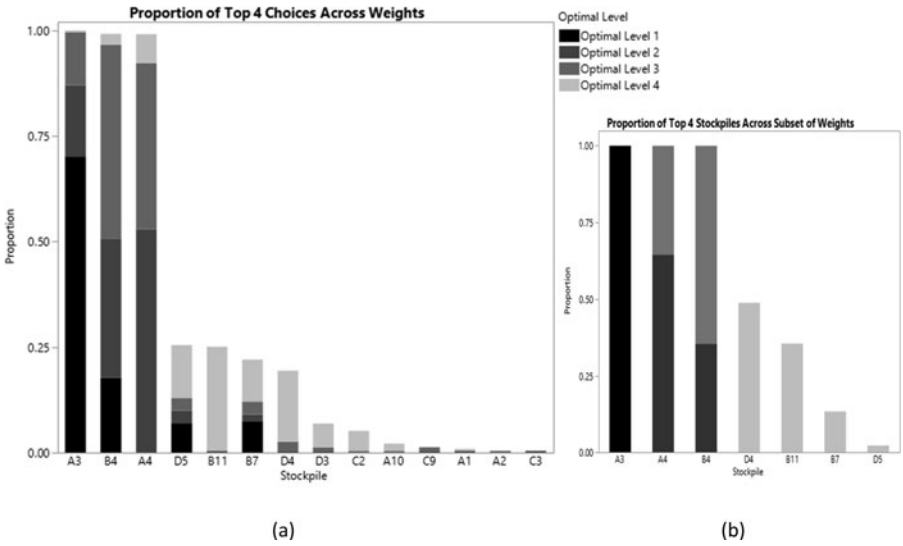

Figure 5. Proportion plot summary for top 4 choices using an additive DF and scaling [0,10]→ [0,1] (a) across all possible weight combinations, and (b) across the specified subset of weight combinations requiring Overall Reliability ≥20%, Overall Urgency ≥10%, and Consequence ≥30%.

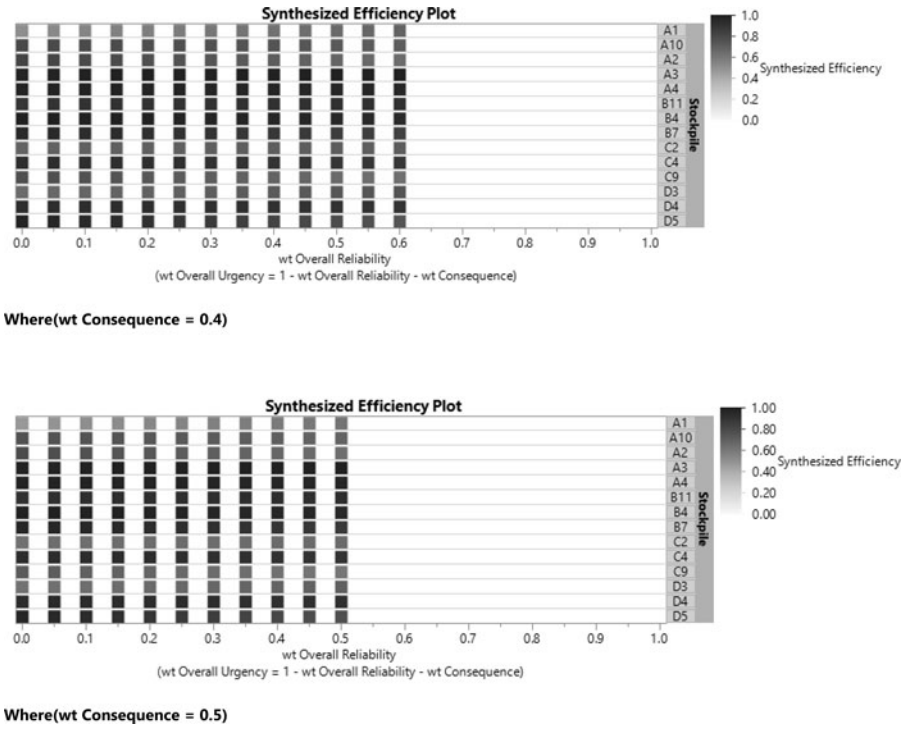

Figure 6. Synthesized efficiency plot when Consequence is weighted 40% and 50% using an additive DF and scaling [0,10]→ [0,1]. (Top) Where (wt Consequence = 0.4); (bottom) Where (wt Consequence = 0.5).

consequence should always be prominently considered, while urgency was least important. A new proportion plot for this subset of weights is shown in Figure 5(b). The choice of the fourth stockpile now becomes clearer for this region of weights. D4 is the 4th most critical solution for approximately 50% of this subset of weights.

From the different numerical and graphical summaries, the team selected A3, A4, B4, and D4 as the four stockpiles to receive the additional funds. To finalize the decision, the team considered the dynamic synthesized efficiency plot (also in the TopN-PFS Add-In). This adapted version of the synthesized efficiency plot for a single PF (Lu and Anderson-Cook 2012) allows exploration of how top solutions compare to the best available solution at a given weight combination. In

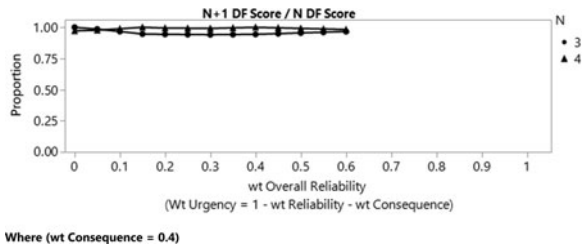

Figure 7. N + 1 comparison plot for N = 4 and N = 3 when Consequence is weighted 40% using an additive DF and scaling [0,10]→ [0,1]. Where (wt Consequence = 0.4).

Figure 6, the same two slices of weights for consequence fixed at 0.4 and 0.5 are shown. Note that from Figure 4, stockpile A3 is the top choice for $w_{OR} \geq 0.15$ and is therefore shown with the darkest shade in Figure 6. Note that even for $w_{OR} < 0.15$ when $w_C = 0.4$ or $w_C = 0.5$, stockpile A3 has near optimal performance with synthesized efficiency close to 1 (very dark shade). This suggests stockpile A3 is nearly universally most critical for all weights in Figure 6. In addition, all the top 4 stockpiles (A3, A4, B4, and D4) have dark shades throughout the range of weights shown, indicating they are leading choices. A few other stockpiles (such as B7, B11, and C4) also have similar shades for much of the ranges, which shows that more than 4 stockpiles could merit consideration. This plot allows the managers to assert to the funder the benefits of receiving even more funds to support additional stockpiles.

The final decision for the top 4 leads to a natural question about whether the choice made by the sponsor of N = 4 stockpiles was sensible. The N + 1 comparison plot (Figure 7) shows the impact of this choice on the final decision. This plot shows the ratio of the DF scores for the (N + 1)th best solution to the Nth best solution for given weight combinations. The ratio is naturally bounded between 0 and 1 with high values indicating the (N + 1)th solution is very close

to the Nth best choice. If the ratios across all weights are relatively low, it indicates the next best choice is not competitive with the Nth best choice. Figure 7 shows the N + 1 comparison plot for N = 3 and N = 4 using an additive DF and the scaling [0, 10] → [0, 1]. We can see that when N = 4 and $w_C = 0.4$, the fifth best solution has very close (almost indistinguishable) performance compared to the fourth. For N = 3, there is bigger separation. For some of the weights, the fourth best choice is below 90% efficient (in terms of the DF score) compared to the third best choice. If the criteria scaling was based on the range of values on the layered PFs, there is an increased change in the DF scores between the Nth and (N + 1)th solutions because the desirability scale is mapped to a narrower range of criteria values (Figure SM1.1 in the online supplementary materials [SM]).

In the end, three of the four managers received additional funding for at least one of their stockpiles. While the manager of the C family was disappointed about the outcome, there was much more clarity about how the decision was made and he understood justification for the choices.

Discussion and conclusions

One potential issue with the approach is the concern that the tools and graphical methods described in the paper may require extensive training before a manager is able to comfortably use them for decision-making. One of the authors (CAC) has introduced these methods to multiple groups with different scientific and quantitative backgrounds. A helpful strategy begins the decision-making process with discussion about the process that will be followed. The DMRCS process is quite intuitive once explained and it is helpful if the participants understand that there will be time later in the process when their opinions will be discussed. When the graphics for the Reduce, Combine, and Select steps are needed, it is helpful to introduce them with a simple example of similar dimension, unrelated to the current decision to be made, and walk through how to extract information from the plots. In 10–15 minutes with encouragement to ask questions about the mechanics of the process, the group is comfortable with the information and able to consider the details of their own decision.

One reviewer noted that the particular set of choices will differ with each application. For example, the three clear winners in this case would not likely occur for other applications or even in subsequent years. Our goal in describing the DMRCS process is not to define a specific path to decision-making, but rather to give a framework to guide how teams approach their task to make informed decisions. If instead of choosing the top 4, the sponsor only had funds for the top 2, then the team would take a closer look at A3, B4, and A4 to determine which of these emerge as most critical. The process of eliminating non-contenders is helpful to make the problem manageable, but the discussion and choices in the subjective Combine and Select stages will be unique to each problem.

As discussed previously, several subjective choices were made in the Combine and Select steps, including the choice of scaling and the DF forms. A comparison of the additive DF results to using a multiplicative DF was performed with the same scaling [0, 10] → [0,1]. In the dynamic mixture plots at $w_C = 0.4$ and $w_C = 0.5$ (Figure SM1.2), we see that stockpiles A3, B4, and A4 remain clear winners. However, the choice of the fourth stockpile again requires additional investigation among B11, D5, D4, and B7. The proportion plot in Figure SM1.3(a) confirms that stockpiles A3, B4, and A4 are clearly among the four most critical stockpiles for nearly all weights. Stockpiles B11, D5, D4, and B7 compete for the fourth place, each of which is among the top 4 choices between 20–25% of all weights. Hence, the multiplicative DF also identifies the same stockpiles (although with minor frequency changes). Figure SM1.3(b) shows the proportion plot for the subset of weights with $OR \geq 20\%$, $OU \geq 10\%$, and $C \geq 30\%$. The same decision is made using the multiplicative DF as the additive DF: D4 is again the fourth ranked stockpile. This was helpful confirmation for members of the decision-making team about the robustness across both DF forms.

The team also investigated the impact of the choice of scaling on the final decision. One of the alternative scaling options is to use the range of criteria values found on the top N layered PFs. For example, for Overall Reliability, the minimum value found on the layered PFs is 6.25, which maps to 0 and the maximum value 9.5 maps to 1. The detailed graphical summaries for using the scaling based on the identified layered PFs using an additive or a multiplicative DF are available in Sections 2 and 3 of the SM. Even with the change of scaling, the top three choices remain the same (A3, B4, and A4), for both the additive and multiplicative DFs

(see Figure SM2.2 and SM3.2, for example). The team was reassured that their decision was robust to these subjective choices. The differences in the scaling and DF impacted the fourth most critical stockpile; however, when the more focused weight region is considered, the same set of top 4 choices (A3, B4, A4, and D4) is always identified.

In addition to the scaling and DF choices, the stockpile decision-making team also investigated the impacts of combining the reliability and urgency metrics in different ways. In the original analysis, the Overall Reliability and Overall Urgency criteria each combined two metrics equally. A mini-factorial of the possible weight combinations between the two reliability measures (1.a and 1.b) and the two urgency measures (2.a and 2.b) are explored. Table SM4.1 in the SM shows the four combinations of weight combinations for the reliability and urgency metrics that were investigated. The tables of the stockpiles identified on the top 4 PF layers and the graphical summaries using an additive DF and the scaling $[0, 10] \rightarrow [0, 1]$ for supporting informed decision-making are included in Section 4 in the SM. The number of stockpiles on the PF layers does change under these four different scenarios. In the original analysis, there were 26 stockpiles on the top 4 PF layers. Scenario 1, however has 29 stockpiles on the top 4 PF layers (Table SM4.2), while scenario 4 only has 19 (Table SM4.5).

The top 3 leading choices remain as A3, B4, and A4 across all four scenarios. The choice of the fourth stockpile, however, does change with different combinations of the reliability and urgency metrics. The team liked the consistency of the results for the top 3 stockpiles. The differences in the choices of the 4th stockpile did trigger animated discussion. However, the team members admitted that no a priori preference for a particular reweighting and they were primarily reacting to the results obtained. Again, the DMRCS process, which builds a solid foundation of focusing on how to make the decision before looking at results builds consensus and understanding of the true goals of the decision.

The final decision of the top 4 stockpiles depends on the region of weights for each criterion as well as the metrics chosen for quantifying the key characteristics. However, when the average reliability and urgency metrics were used, the final decision was more robust to the scaling and DF choices for the more focused priorities. Arriving at a consensus for the region of weights is therefore critical for the team to agree on the final decision of the most critical stockpiles.

Finally, we compare the original process for selecting these stockpiles with the new structured approach. The previous method determined the outcome in a single meeting where the managers presented arguments. In this case, the onus was on managers to develop convincing arguments, which often involved trying to sway the sponsor with an argument that highlighted the merits of choosing particular stockpiles. The metrics they chose were selected for their persuasiveness, rather than having a consistent objective. The new process using the DMRCS approach is data-driven, involves first determining which criteria will lead to an informed decision, and uses input from multiple experts. The sponsor was in a better position to make fair comparisons between alternatives.

A common alternative is to use a DF with a fixed set of weights, typically weighting the criteria equally. This is considerably better than the original method, in that it could encourage the careful a priori choice of metrics based on what is important and has a solid data-based foundation. However, this approach is weaker in that it becomes a "black box," where a single answer with a top 4 ranking of the stockpiles is obtained. The subjective choices made along this process are ignored and a single simplistic decision is made. There is also no exploration of the robustness of the results to weight choices. Often the discussion among the team on the relative importance of the criteria is a valuable part of the process and small differences may still lead to a common choice of best solution. Since there is little or no visualization of alternatives to understand the space of possibilities and robustness, a single choice of weights will either be hard to agree on or a default non-thoughtful choice made. This omission will likely limit the degree of buy-in in the final solution and the decision-making team is unlikely to be able to defend their choice effectively to others. The time to execute this method is similar to the new method, but results in a less satisfying decision.

While this new outlined process increases the time to make the decision over the original method, the stockpile managers understand how the decision has been made, and there is greater buy-in from all participants. Finally, choices made in the Define and Measure steps of the DMRCS process form the foundation of the process for subsequent years, when the choices made

will be reexamined, but likely continue to be the basis for the decision-making.

About the authors

Sarah E. Burke works for the Perduco Group as a statistician at the Scientific Test and Analysis Techniques Center of Excellence in Dayton, Ohio. She earned a doctorate in industrial engineering at Arizona State University. She is a member of the American Statistical Association and ASQ.

Christine M. Anderson-Cook is a Research Scientist in the Statistical Sciences Group at Los Alamos National Laboratory. She works on projects related to the design and analysis of experiments, reliability, non-proliferation, and quality control. She is a Fellow of the American Statistical Association and ASQ.

Lu Lu is an Assistant Professor in the Department of Mathematics and Statistics at the University of South Florida. She earned a doctorate in statistics from Iowa State University. She is a member of ASQ.

Douglas C. Montgomery is Regents' Professor of industrial engineering and statistics and Foundation Professor of Engineering at Arizona State University. His research and teaching interests are in industrial statistics. Professor Montgomery is a Fellow of the ASA, an honorary member of the ASQ, a Fellow of the RSS, a Fellow of IEE, a member of ISI, an academician of the IAQ, and has received several teaching and research awards.

Acknowledgments

The authors thank the anonymous reviewers and editor for their helpful suggestions.

References

Anderson-Cook, C. M. 2013. Balancing priorities: Making that elusive perfect decision: A primer. *Quality Progress* 46 (9):52–53.

Anderson-Cook, C. M., and L. Lu. 2015. Much-needed structure: A new 5-step decision-making process helps you evaluate, balance competing objectives. *Quality Progress* 48 (10):42–50

Anderson-Cook, C. M. 2017. Optimizing in a complex world: Statisticians' roles in decision-making" (with discussion and rejoinder). *Quality Engineering* 29 (1):27–41.

Anderson-Cook, C. M., L. Lu, G. Clark, S. P. DeHart, R. Hoerl, B. Jones, R. J. MacKay, D. C. Montgomery, P. A. Parker, J. Simpson, et al. 2012a. Statistical engineering – forming the foundations. *Quality Engineering* 24:110–132. doi: 10.1080/08982112.2012.641150.

Anderson-Cook, C. M., L. Lu, G. Clark, S. P. DeHart, R. Hoerl, B. Jones, R. J. MacKay, D. C. Montgomery, P. A. Parker, J. Simpson, et al. 2012b. Statistical engineering – roles for statisticians and the path forward. *Quality Engineering* 24:133–152. doi:10.1080/08982112.2012.641151.

Burke, S. E., C. M. Anderson-Cook, L. Lu, C. M. Borror, and D. C. Montgomery. 2016. Design of experiment selection using layered fronts. *Los Alamos National Laboratory Technical Report* LAUR-16-25643.

Burke, S. E., C. M. Anderson-Cook, and L. Lu. 2017. TopN-Pareto front search. *Los Alamos National Laboratory Computer Code* 17–002.

Collins, D. H., C. M. Anderson-Cook, and A. V. Huzurbazar. 2011. System health assessment. *Quality Engineering* 23:142–151. doi:10.1080/08982112.2010.529484.

Derringer, G., and R. Suich. 1980. Simultaneous optimization of several response variables. *Journal of Quality Technology* 12:214–219.

Hoerl, R. W., and R. D. Snee. 2010. Closing the gap: Statistical engineering links statistical thinking, methods, tools. *Quality Progress* 43 (5):52–53.

Hoerl, R. W., and R. D. Snee. 2012. *Statistical Thinking: Improving Business Performance*. 2nd ed. New York: Wiley.

JMP, Version 12. SAS Institute Inc., Cary, NC, 1989–2007.

Lu, L., and C. M. Anderson-Cook. 2011a. Prediction of reliability of an arbitrary system from a finite population. *Quality Engineering* 23:71–83.

Lu, L., and C. M. Anderson-Cook. 2011b. Using age and usage for prediction of reliability of an arbitrary system from a finite population. *Quality and Reliability Engineering International* 27:179–190. doi:10.1002/qre.1109.

Lu, L., and C. M. Anderson-Cook. 2012. Rethinking the optimal response surface design for a first-order model with two-factor interactions, when protecting against curvature. *Quality Engineering* 24 (3):404–421. doi:10.1080/08982112.2012.629940.

Lu, L., C. M. Anderson-Cook, and D. Lin. 2014. Optimal designed experiments using a Pareto front search for focused preference of multiple objectives. *Computational Statistics and Data Analysis* 71:1178–1192. doi:10.1016/j.csda.2013.04.008.

Lu, L., C. M. Anderson-Cook, and T. J. Robinson. 2011. Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier *Technometrics* 53:353–365. doi:10.1198/TECH.2011.10087.

Snee, R. D., and R. W. Hoerl. 2012. Inquiry on pedigree. *Quality Progress* 45 (12):66–68.