Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1476

Published online 10 December 2012 in Wiley Online Library

Balancing Multiple Criteria Incorporating Cost using Pareto Front Optimization for Split-Plot Designed Experiments

Lu Lu and Christine M. Anderson-Cook*†

Finding a D-optimal design for a split-plot experiment requires knowledge of the relative size of the whole plot (WP) and sub-plot error variances. Since this information is typically not known a priori, we propose an optimization strategy based on balancing performance across a range of plausible variance ratios. This approach provides protection against selecting a design which could be sub-optimal if a single initial guess is incorrect. In addition, options for incorporating experimental cost into design selection are explored. The method uses Pareto front multiple criteria optimization to balance these objectives and allows the experimenter to understand the trade-offs between several design choices and select one that best suits the goals of the experiment. We present new algorithms for populating the Pareto front for the split-plot situation when the number of WPs is either fixed or flexible. We illustrate the method with a case study and demonstrate how considering robustness across variance ratios offers improved performance. The Pareto approach identifies multiple promising designs, and allows the experimenter to understand trade-offs between alternatives and examining their robustness to different ways of combining the objectives. New graphical summaries for up to four criteria are developed to help guide improved decision-making.

Keywords: restricted randomization; whole plot to sub-plot error variance ratio; multiple design criteria; computer-generated designs; D-optimality; point exchange algorithm

1. Introduction

In many experiments, hard-to-change factors with expensive, difficult, or time-consuming levels are common. In these cases, a split-plot design (SPD) is typically run, where multiple combinations of other easy-to-change factors are run with fixed levels of the hard-to-change factors. The hard-to-change whole plot (WP) factor levels are applied to the larger experimental units called WPs. The easy-to-change sub-plot factor levels are applied to the smaller experimental units called sub-plots.

For situations with flexible or unusual size constraints and/or irregularly shaped design regions, computer-generated designs are common, because standard designs, such as full or fractional factorial designs, often require specific design sizes and regularly shaped regions. Standard computer-generated design construction requires the user to specify a model, the design size (i.e. the total number of sub-plot observations), and an optimality criterion. For SPDs, the relative size of the WP and sub-plot error variances is also required for optimization.[1] D-optimality is a common criterion for precise estimation of model parameters. If the primary interest is good prediction, then G- and I-optimality are popular choices. All of these metrics depend on the relative size of the WP and sub-plot variances. However, for good optimization, since this variance ratio is often not known a priori, it is desirable to take into account its associated uncertainty of its specified value.

In addition, many situations have some flexibility for the overall cost of the experiment. A practitioner may be willing to increase spending if the anticipated benefits of a larger experiment are justified. For the SPD case, both the design size and the number of WPs (#*WP*) typically influence cost, where the relative cost of the sub-plot and WP units vary for different applications. We advocate exploring different-sized experiments within the acceptable range to better understand whether changes in cost are likely to yield disproportionate changes in design performance. More details on generally desirable design characteristics are available in[2] (p. 282).

Consider a case study where the experimenter wants to conduct a designed experiment for studying the relationship between the response and three design factors, with two of them as hard-to-change. A first-order model with all two-factor interactions is assumed based on some scientific knowledge. The experimenter can collect between 12 and 16 observations with a flexible #*WP* at either high (+1) or low (−1) levels of the three factors. For our case, the cost of resetting the WP factor levels is assumed to be approximately the same as collecting and measuring a sub-plot observation. Due to very limited pilot study data, the estimate of the WP to sub-plot

Statistical Sciences Group, Los Alamos National Laboratory, USA
**Correspondence to: Christine M. Anderson-Cook, Statistical Sciences Group, Los Alamos National Laboratory, USA.*
†E-mail: icyemma@gmail.com; candcook@lanl.gov

variance ratio cannot be made more specific than between 0.1 and 10. Here, we use d_{lower} and d_{upper} to denote the lower and upper limits of the specified variance ratio, d. The goals of the experiment are to: (i) estimate model terms precisely; (ii) choose a design which is robust to uncertainty in the specified variance ratio and (iii) minimize the overall experimental cost (influenced by #*WP* and *N* for general SPD experiments).

Finding optimal SPDs has received considerable attention in the recent literature.[3–5] use *minimum aberration* criteria to determine optimal two-level fractional factorial SPDs for near-orthogonality.[6–8] use point exchange algorithms for finding D-optimal SPDs.[9] improve computational efficiency with coordinate exchange algorithms for D-optimal SPDs.[10] develop update formulas for improving the efficiency of point exchange algorithms with flexible constraints on #*WP*.[11] consider the G-optimality criterion.[12] construct SPDs with high D-efficiency and robustness to model mis-specification using a coordinate exchange algorithm.[13] develop a class of SPDs where the ordinary least squares estimates of model parameters are equivalent to the generalized least squares (GLS) estimates.[14] propose designs built by strata.[15] develop cost-penalized optimality criteria.[16,17] use graphical techniques to characterize SPD performance.

A good design produces desirable results under a variety of user-specified objectives.[18] With multiple dimensions needed to capture the 'goodness' of a design, it can be restrictive to limit optimization to a single criterion or a single combination of criteria. Rarely does a universally 'best' design exists which is superior for all criteria of interest. Hence, it is beneficial to evaluate trade-offs between the criteria and select a design with good performance for the priorities of the study.

The Pareto frontier approach for multiple criteria optimization in the SPD setting begins by objectively finding the set of designs not inferior to any other permissible design. All other designs can be eliminated from further consideration since at least one design on the Pareto front is superior. The second subjective step combines experimenter priorities with the Pareto optimal designs to trim the set of promising designs based on the relative importance of the criteria. Finally, a single best design is selected after considering design performance, trade-offs and robustness. A sensitivity study for a range of experiment priorities is easy given the set of Pareto optimal designs.

Due to constraints on randomization for SPDs, the Pareto Aggregating Point Exchange (PAPE) algorithm for populating the Pareto front in the completely randomized design (CRD) setting[19] and its enhancements[20] cannot be applied directly. We develop two variations of a new SPD search algorithm for scenarios with a fixed *N*, and either a fixed or flexible #*WP*. Once the Pareto front has been found, a rich set of graphical tools[19,21] offer the user quantitative and structured design selection strategies. Several approaches to incorporating cost are explored and compared. New extensions of the graphical summaries for four criteria are illustrated.

In the next section, we review the basics of the Pareto front approach for designed experiments and provide background for split-plot models and design criteria for addressing the uncertainty associated with specified variance ratio as well as cost concerns. Section 3 describes the new search algorithm for fixed *N* and #*WP*. An alternative algorithm for fixed *N* and flexible #*WP* is given in Appendix A. Section 4 illustrates the SPD decision-making process using the Pareto front approach for a simple scenario when the maximum allowable design size $N = 16$ is used. The results are compared to a relevant example from[1] (p. 213) for a few specific choices of variance ratios. Section 5 returns to the more general situation of our case study which considers a range of experiment sizes. Different ways for incorporating cost are examined. When *N* and #*WP* are treated as separate criteria related to experimental cost, a four-criteria design selection problem is discussed and new graphical tools are developed. Section 6 shows some other examples of SPDs and illustrates differing patterns of trade-offs between criteria. Section 7 gives some conclusions.

2. The Pareto approach and the D-Optimality criterion for SPDs

In this section, we provide background on the general Pareto front approach for multiple objective optimization, and the model and D-optimality criterion used for SPDs. Then, we give specifics on the design criteria used in our examples: the D-criteria using both the maximum and minimum possible variance components ratios and cost-based criteria (dependent on both *N* and #*WP*). The motivation for using two D-criteria based on different variance ratios comes from the unique challenge of needing to specify the variance ratio prior to running the experiment in order to find the best design. First, we summarize the Pareto approach by[19] for multiple criteria design optimization in the CRD setting. The special adaptations required for SPDs are described in Section 3.

2.1. The Pareto approach

Pareto multiple objective optimization has been broadly used in other disciplines, before being introduced as a structured decision-making process in the design of experiment paradigm.[19] The method consists of two stages: (i) objective Pareto optimization, to assemble a set of non-dominated candidate designs, and (ii) subjective Pareto decision analysis, to select a subset of designs from the Pareto front closest to an 'ideal' solution across a spectrum of weight combinations. A final design is then chosen based on individual performance, trade-offs and robustness to different weightings using a set of graphical methods. By separating the objective and subjective steps, we can first see the complete set of choices, before imposing any subjective considerations. This first step allows the experimenter to be aware of all options before making a subjective final choice.

A main competitor to the Pareto front approach is the desirability function (DF) approach of,[22] which does not directly consider trade-offs between criteria, but rather a best design is identified with a 'black box' algorithm based on the pre-determined relative weighting of scaled individual criteria in a combined summary. It specifies a single best choice and does not show how other competing choices compare to the optimum and hence often does not match real life complex decision-making. It can lead to results sensitive to the user-specified weightings, expert guesses of ranges of criteria values specified prior to the search, scaling schemes (for converting the criteria values to be on a 0–1 scale) and DF forms (for additively or multiplicatively aggregating multiple criteria into a single

summary). A sensitivity analysis to explore the impact of these subjective choices is computationally expensive with the DF approach, since each new setup requires a separate and frequently time-consuming search for a solution.

For a CRD, one way to construct the Pareto front for a moderate-sized design is with the PAPE algorithm.[19] The algorithm uses a user-specified candidate set and multiple random starts to build new designs, where multiple designs can be added to the Pareto front for a given random start. For the second stage, an adapted Utopia point approach explores all weight combinations to reduce the promising candidates to a more manageable size, where every member is optimal for at least one combination of weights. The typically unattainable Utopia point has best values for all criteria, and commonly used to select designs closest to this 'ideal' target. Let Ω^* denote the Pareto set of designs, $f_j(\boldsymbol{\xi})$ denote the scaled objective function for the j-th criterion for design ξ and f_j^0 denote the Utopia point value for the j-th criterion. A L_1-norm[23] is formulated as $\underset{\boldsymbol{\xi}\in\Omega^*}{\text{Min}} \sum_{j=1}^{C} \left| w_j\left(f_j(\boldsymbol{\xi}) - f_j^0\right)\right|$, where the w_j in $j \in \{1,2,\cdots,C\}$ are the weights assigned to each criterion. This formulation produces identical solutions to the additive DF with matching weights and scale.[19] When the multiplicative DF is of interest, the L_1-norm on the log scale can be used with a logarithm transformation for the f_j's.

Designs which are best for at least one combination of weights are further evaluated on three aspects: (i) optimal choices for all user specified desired region of weights, (ii) robustness across ranges of weights close to the region of interest, and (iii) performance as measured by synthesized efficiency[21] relative to the best possible design for different weighting choices. The user makes a final decision based on the priorities of the study after considering the above information. The Pareto approach allows an informed decision by providing a rich set of information with manageable computational effort.

Next, we consider the split-plot model and quantitative criteria for a 'good' design in our particular case study.

2.2. *The split-plot model and the D-optimality criterion*

The linear mixed model for the response, denoted by an $N \times 1$ vector, $\mathbf{y}$, in a SPD, is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{X}$ is the $N \times p$ design matrix expanded to the model form for the p fixed effects model parameters, $\boldsymbol{\beta}$, including both the WP and sub-plot variables; $\mathbf{Z}$ is an $N \times \#WP$ matrix of ones and zeroes with the entry in i-th row and j-th column being 1 if the i^{th} observation ($i = 1,\ldots,N$) is in the j^{th} WP ($j = 1,\ldots,\#WP$), and 0 otherwise; $\delta \sim^{i.i.d.} N\left(\mathbf{0}, \sigma_\delta^2 \mathbf{I}_{\#WP\times\#WP}\right)$ is a $\#WP \times 1$ vector of WP random effects where σ_δ^2 is the WP variance; $\boldsymbol{\varepsilon} \sim^{i.i.d.} N\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{N\times N}\right)$ is the $N \times 1$ vector of random errors with σ_ε^2 as the sub-plot variance. The $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are assumed independent. The design matrix can be written as $\mathbf{X} = [\mathbf{W}|\mathbf{S}]$, where $\mathbf{W}$ contains model terms involving only WP factors (and the intercept if applicable), and $\mathbf{S}$ contains all model terms involving sub-plot factors.

Given the split-plot model in (1), the variance-covariance matrix of the response, $\mathbf{y}$, is

$$\begin{aligned}\text{Var}(\mathbf{y}) = \boldsymbol{\Sigma} &= \sigma_\delta^2 \mathbf{Z}\mathbf{Z}' + \sigma_\varepsilon^2 \mathbf{I}_{N\times N} \\ &= \sigma_\varepsilon^2\left(d\mathbf{Z}\mathbf{Z}' + \mathbf{I}_{N\times N}\right)\end{aligned}$$

where $d = \sigma_\delta^2/\sigma_\varepsilon^2$ represents the WP to sub-plot variance ratio. If the entries in $\mathbf{y}$ are ordered by WPs, then $\mathbf{Z} = diag\{\mathbf{1}_{n_1}\cdots, \mathbf{1}_{n_{\#WP}}\}$, where $\mathbf{1}_{n_j}$ is an $n_j \times 1$ vector of one's, and n_j is the number of sub-plot observations in the j^{th} WP. The variance-covariance matrix can be written as a block diagonal, $\boldsymbol{\Sigma} = diag\{\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_{\#WP}\}$, where $\boldsymbol{\Sigma}_j$ is an $n_j \times n_j$ variance-covariance matrix for the j^{th} WP:

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_\delta^2 & \cdots & \sigma_\delta^2 \\ \vdots & \ddots & \vdots \\ \sigma_\delta^2 & \cdots & \sigma_\varepsilon^2 + \sigma_\delta^2 \end{bmatrix} = \sigma_\varepsilon^2 \begin{bmatrix} 1+d & \cdots & d \\ \vdots & \ddots & \vdots \\ d & \cdots & 1+d \end{bmatrix}.$$

Note that the variance of an individual observation is $\sigma_\varepsilon^2 + \sigma_\delta^2$, and the covariance between observations in a common WP is $\sigma_\delta^2 = d\sigma_\varepsilon^2$. Observations from different WPs have covariance of zero. A popular method for estimating the variance components, σ_δ^2 and σ_ε^2, is to use restricted maximum likelihood.[24]

Using the generalized least squares (GLS) estimator, the parameters, $\boldsymbol{\beta}$, can be estimated by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$, with the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ as $Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}$, which is the inverse of the split-plot information matrix, $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})$. The D-efficiency for design ξ can be scaled by $N/\left(\sigma_\delta^2 + \sigma_\varepsilon^2\right)$ to give

$$|\mathbf{I}(\xi)|^{\frac{1}{p}} = \left|\left(\sigma_\delta^2 + \sigma_\varepsilon^2\right)\left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)/N\right|^{\frac{1}{p}} = \left|\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\right|^{\frac{1}{p}}/N \tag{2}$$

In (2), $\mathbf{R} = \frac{\boldsymbol{\Sigma}}{\sigma_\delta^2+\sigma_\varepsilon^2} = \frac{\boldsymbol{\Sigma}/\sigma_\varepsilon^2}{d+1}$ is the observational correlation matrix, which is dependent on d and the matrix $\mathbf{Z}$. The exponent adjusts the quantity to the scale of the response, and larger D-efficiency values indicate more precise model parameters estimates. In our examples, we use $|\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}|^{1/p}$ alone for quantifying D-efficiency with cost as a separate criterion.

Typically, software for constructing D-optimal SPDs requires the user to specify a model, d, N and perhaps, $\#WP$. If the variance ratio d is unknown, finding the optimal design relies on arbitrarily choosing a single value guess and can result in diminished performance if it is later discovered that the specified ratio is incorrect. In our example, the practitioner has limited a priori

knowledge of d (here, the best guess is a range of 0.1 to 10). Hence, not restricting the optimization search prematurely and taking this uncertainty into account during design construction is an important consideration for the SPD scenario. We denote the D-efficiency criterion as $D(d)$ to distinguish design performance based on different d values.

Experimental cost is often an important issue in design construction. Traditional design selection incorporates cost by building it into a single criterion metric such cost-adjusted D-efficiency in (2) by using cost in the denominator, which can entangle different aspects of design characteristics and prevent separate evaluation of their effects. On the other hand, the Pareto front approach allows the experiment to consider multiple aspects of design performance simultaneously and separately and hence provides more information about how competing objectives affect design selection as well as their interrelationship. The special structure for SPDs suggests two contributors to cost.[25,15] discuss potential cost formulations for SPDs. A popular choice is $Cost = \#WP + c_r N$, where c_r is the cost ratio of obtaining a sub-plot observation to a WP observation. Smaller values of c_r suggest that it is cheap to obtain more observations relative to another WP. This formulation is useful when $\#WP$ and N both influence the same aspect of cost (e.g. financially) and their relative contributions are known. There are other situations when N and $\#WP$ influence different aspects of cost (e.g. financial cost versus time), and so we also consider the more general case where $\#WP$ and design size, N, are treated as two separate criteria. This also enables us to compare the impact from different approaches to incorporate experimental cost.

Note that N and $\#WP$ are important decision variables for design construction and controlling the searching process. Typically, decision variables are separate from the criteria used to select the most suitable design. However, cost is an exception and can play both roles of decision variable and optimization criterion. We often have choices about how much to spend on collecting the data, but differences in cost should also influence our impression of the 'goodness' of a design. For most SPDs, both N and $\#WP$ influence the experimental cost, but also define the parameters of the search algorithm for the problem of interest. This unique characteristic of having cost as both decision variable and optimization criterion has some advantages in speeding up the Pareto searching process. The algorithm and how to leverage these advantages are discussed in the following section.

3. The PAPESPD algorithm

Given the criteria identified in Section 2, the Pareto front approach can be used to determine a best design for their objectives of interest. The success of the Pareto front approach replies on efficiently populating the complete Pareto front.[19] use the PAPE algorithm to improve the computational efficiency of the traditional point exchange algorithm by updating the Pareto front based on all designs from the search process. This substantially reduces the computational time to fully populate the front. However, the original PAPE algorithm uses a greedy search approach for guiding the searching direction, which updates a current design only if the new one strictly improves at least one of the criteria without making others worse. This constrains the search to move in limited directions and consequently restricts the region on the front attainable from a random starting design.[20] propose two alternatives to improve search performance by conducting parallel searches in diverse directions. For each random start, multiple searches are conducted simultaneously using either a set of pre-specified fixed weight combinations covering the weighting space or a set of stratified random weights generated at each updating step. The new updating mechanisms improve completeness of the identified front, efficiency finding the front, as well as reducing process variation from random starts. To conduct efficient Pareto front searches, the algorithms in[20] need to be adapted to include the randomization constraints of SPDs. An exchanged point can either form a separate WP, or be aggregated into an existing one. Hence, in each updating step, all choices need to be fully explored.

In this section, we propose a new algorithm, the PAPESPD using a set of fixed weighting choices. At each exchanging step, the algorithm employs the fixed weights for aggregating the different criteria to determine which alternative to pursue. More specifically, for each random start, parallel searches are conducted for each pre-specified weight combinations. For each of these searches, when multiple choices are available at an updating step, we select the choice which maximally improves the combined metric based on the particular set of weights. The diverse weightings spread the exploration and prevent the search from getting trapped in local optima. Meanwhile, the Pareto front is populated using all designs explored from all the searches.

Cost is a commonly considered criterion for design selection. Having the cost dependent on both N and $\#WP$ introduces a special attribute of the Pareto front for SPDs. Both N and $\#WP$ are discrete, monotonic and often have a moderate number of possible levels. To explore different ways of incorporating cost, we first find separate Pareto fronts for each combination of N and $\#WP$, and then construct the overall Pareto front from the collection of results after incorporating cost. The overall Pareto fronts for more criteria are formed using simpler fronts as building blocks. In addition, $\#WP$ and N, as decision variables, are easy to fix in the search algorithm, and hence we can adapt the search to leverage this. More specifically for our examples, we find separate Pareto fronts using the two D-criteria, $D(d_{lower})$ and $D(d_{upper})$, for multiple combinations of fixed N and $\#WP$, and then combine these into a single three or four criteria Pareto front based on different ways of incorporating cost (discussed more in Section 5). Note that when some decision variables have discrete values, we can break the higher dimensional search into multiple lower dimensional searches to reduce computational complexity and time, as well as leverage the benefits of parallelization. Often, the resources required to fully populate the front grows exponentially as the number of criteria increases.

Next, we give a detailed description of the PAPESPD for the scenario with a fixed N and $\#WP$. For applications with a flexible $\#WP$, the front can be generated from the combined collection of designs for different possible $\#WP$ values. An alternative search algorithm that allows $\#WP$ to change during the search with fixed N is in Appendix A.

3.1. The PAPESPD with fixed N and #WP based on fixed weightings

Let u denote the number of WP terms in the model (i.e. terms involving only WP factors plus an intercept if it is included). The PAPESPD algorithm for this case consists of steps described below:

Step 0: Determine the optimization criteria. The PAPESPD can accommodate any quantitative choice.
Step 1: Randomly generate a non-singular starting SPD with the required #*WP* and *N*. Each generated design is represented by a matrix with the first column containing the WP indexing number and the remaining columns containing the WP and sub-plot factor levels.
1.a) Generate u WPs with different factor level combinations from the candidate set of WP allocations using simple random sampling without replacement. This assures at least u different WP factor combinations in the selected design to increase the likelihood that all WP term parameters are estimable.
1.b) Generate the remaining ($\#WP - u$) WPs from the candidate set using simple random sampling with replacement. Randomly assign WP index numbers ($1, \ldots, \#WP$) to the WPs from 1.a) &1.b).
1.c) Randomly assign the N sub-plots to the WPs with at least one observation for each WP index number.
1.d) Calculate the criteria values for the generated design: $D(d_{lower})$, $D(d_{upper})$, N and #*WP*. Verify that the information matrix determinant for both criteria $D(d_{lower})$ and $D(d_{upper})$ is above a pre-determined small threshold value (say 0.01, to assure stability of parameter estimates). Repeat 1.a)–1.d) until an acceptable SPD is generated.

Step 2: Construct a Pareto set of non-dominated designs from the random design generated in Step 1: For the generated random start, multiple searches are conducted simultaneously based on a set of pre-specified weight combinations. For a particular set of weights, for every row in the design matrix, explore the set of all alternative candidate locations:
2.a) For every candidate location, all new designs are generated by replacing the *current row* with the *new row* conditional on keeping #*WP* the same. The scenarios for creating all possible new designs are:

- Scenario 1: The current row is in a separate WP with only one sub-plot observation. The new design replaces the current row with a new row with the same WP index number. There are no restrictions on the factor levels for this new WP.
- Scenario 2: The WP containing the current row has at least two sub-plot observations with the new and current rows with the same WP factor levels. New designs can be created by: (i) replacing the current row with the new row in the same WP, or (ii) if there are other WPs with the same WP factor levels, add the row to one of these. The number of possible new designs is #WP with the same WP factor levels as the new row.
- Scenario 3: The WP containing the current row has at least two sub-plot observations with the new and current rows with different WP factor levels. If there are other WPs with the same WP factor levels as the new row, then add the new row to one of the other WPs with the same factor levels. If there are no WPs in the current design with the same WP factor levels as the new row, then no new design can be created to keep #WP the same. The number of possible new designs is the #WP with the same WP factor levels as the new row.

2.b) Calculate the criteria values for all new designs. Create an overall score for each design by combining the criteria based on the particular set of weights. If the best new design is better than the current design based on the combined summary index, then update the current design.

2.c) For each weight combination, repeat 2.a) and 2.b) for all the rows in the current design and all candidates until no further change can be made. As designs are created, their criteria values are evaluated to update the current Pareto front. For more details, see[19] (Section 4).

2.d) Repeat 2.a)–2.c) for all specified weight combinations. The Pareto front is updated for each fixed weight search with an overall front identified based on multiple independent searches from the same random starting design.

Step 3: Repeat Steps 1 and 2 for multiple random starts. Fronts are generated for each random start, and a combined Pareto front is constructed by merging them.

The new PAPESPD algorithm uses fixed weightings with good coverage of the weighting space to guarantee exploration of all regions of the Pareto front. An alternative can be the adaptation from the enhanced PAPE algorithm using stratified random weights at every updating step[20] for the SPD setting.

4. Example using the maximum available design size

To illustrate the decision-making process for SPD selection, we initially consider a simple example related to our example in the Introduction, which selects D-optimal SPDs with two WP factors and one sub-plot factor, two levels for each factor and a flexible #*WP* for a fixed design size, $N = 16$. The decision to focus on the maximum allowable design size is a common one, as it is known that more observations improve the precision of estimation and achieve better D-efficiency. In the Section 5, we return to the original problem and evaluate how considering a smaller experiment may yield adequate precision and allow resources to be saved for future data collection stages. Due to the limited pilot study data, the variance components ratio, d, cannot be specified more precisely than

[0.1, 10]. The experiment objectives include precise estimation of the model coefficients while minimizing overall cost. For fixed $N = 16$, the experimental cost is determined only by the #*WP*. Hence, we seek a best SPD based on simultaneously balancing three criteria: D(0.1), D(10), and #*WP*. Here, #*WP* is both decision variable and optimization criterion and can have values between 5 and 16, where the 16-WP SPD is the CRD. We select the lower bound of 5 WPs because at least 4 degrees of freedom are needed to estimate the WP terms and we include at least one extra degree of freedom for estimating the WP error variance. Due to the uncertainty associated with *d*, we want a robust SPD that performs well over the range *d* in [0.1, 10]. Hence, we seek the best SPD that simultaneously maximizes the D-efficiency at the end points of the variance ratio interval, i.e. D(0.l) and D(10), while minimizing experimental cost, i.e. #*WP*.

A Pareto optimization search was conducted for $N = 16$ based on D(0.1) and D(10), using a set of parallel searches with the PAPESPD algorithm for every fixed #WP value between 5 and 16. Each search for a given #*WP* takes less than 10 s on average on a standard desktop, and the Pareto front based on D(0.1) and D(10) for a given #*WP* consists of a single design. Combining the two-criteria Pareto fronts for different #*WP* forms a single three-criteria Pareto front (Figure 1). There is a trade-off between D(0.1) and D(10) for designs with 8 and more WPs. However, with little sacrifice in D(0.1) disproportionate D(10) gains are possible with fewer WPs. The D(10) criterion starts to decrease as #*WP* drops below 8. Therefore, if interest lies in cheaper designs (no more than 8 *WP*s), there is only a trade-off between D-efficiency and #*WP* regardless of the variance ratio. However, if more WPs are allowed (more than 8), then there is additional trade-off between the D-efficiencies based on different variance ratios.

Using the Pareto front solutions, we can also identify the D-optimal designs for all ratios in [0.1,10] in Figure 2. Separate searches for the D-optimal designs for each value in a fine mesh of equally spaced variance components ratios from 0.1 to 10 were performed, and all of these D-optimal designs were found by the PAPESPD algorithm based on D(0.1) and D(10). Figure 2 shows the relative

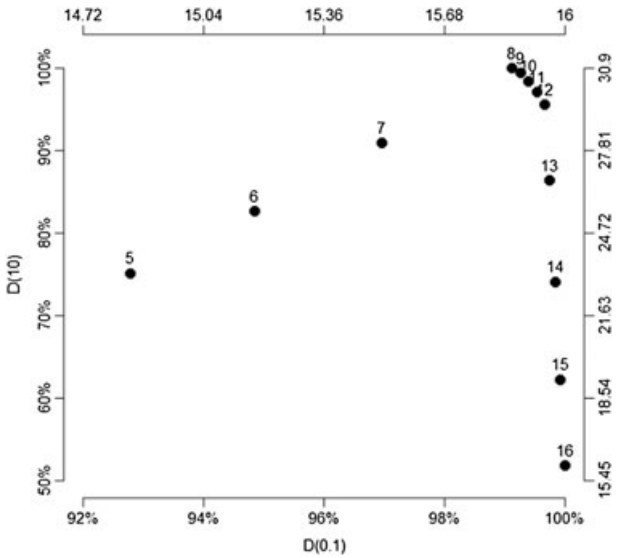


Figure 1. Pareto front for 16-run SPDs based on D(0.1), D(10) and #*WP*. Labels for the designs are the #*WP*. The bottom and left axes are the relative D-efficiencies compared to the optimum. The top and right axes are the actual D-efficiency values

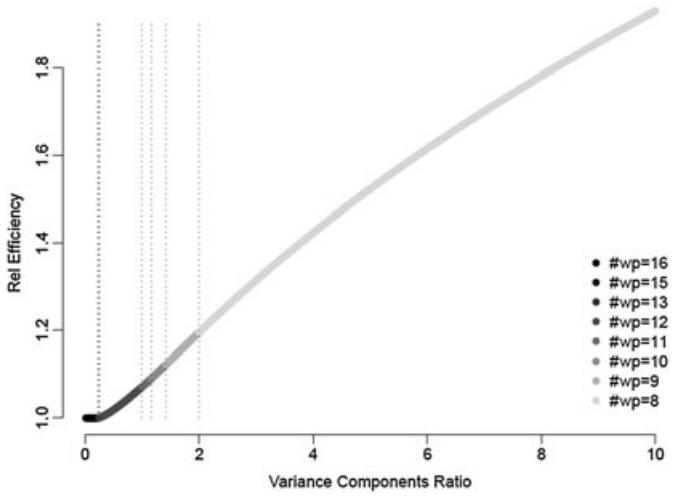


Figure 2. Relative efficiency of the D-optimal 16-run SPDs compared to the 16-run CRD for different variance components ratios in [0.1, 10]. Different shades of gray distinguish the designs with different #*WP*

efficiency of the optimal designs compared to the 16-run CRD for different variance ratios in [0.1,10]. The 16-run CRD is optimal for the region with small variance ratios (less than 0.23). Some designs are only optimal for narrow intervals (e.g. 15 and 13 WPs), while other designs are more robust across wider ranges like the 8-WP design for [2,10]. The 12-WP, 11-WP, 10-WP and 9-WP designs are D-optimal for $d \in [0.25, 1)$, $d \in [1,1.15]$, $d \in (1.15, 1.41]$ and $d \in (1.41, 2]$, respectively. Since all the D-optimal designs for various variance ratios are found using the end point-based Pareto search, Figure 2 can also be obtained from only the Pareto front designs, which is quick and straightforward to do.[1] (p. 212 Table 8.4) considers a similar problem which seeks 16-run D-optimal SPD for a few fixed ratios. The results obtained match those of[1] for $d = 0.1, 0.25, 0.5, 0.75, 1$ and 2, except that for $d = 0.25$ both the 12 and 13WPs are D-optimal. However, the Pareto front approach found all the promising designs more efficiently than conducting independent searches on a fine grid of variance ratios.

Given the Pareto front for the three criteria D(0.1), D(10) and #*WP* in Figure 1, we illustrate how to select a single best design from the set based on the particular goal of a study. Recall that the experimenter will only be able to run a single design, and so a strategy is needed to reduce the choices on the Pareto front to a final choice. A smaller set of designs can be identified as best for at least some weighting of the three criteria by aggregating the criteria into a single summary based on different weightings of the criteria using a user-specified scaling scheme and DF form. To scale the criteria, we set the worst relative D-efficiency from either D(0.1) and D(10) to 0, making the scaled version of the D-criteria more directly comparable. The cost-based criterion, i.e. #WP, is scaled using the standard scaling (worst as 0 and best as 1). Both additive and multiplicative DF forms are explored. By using the adapted Utopia point approach,[19] 7 designs are identified best for at least one set of weights from a fine mesh of weighting using the L_1-norm (additive DF) and 10 designs for the L_1-norm on the log scale (multiplicative DF). Because experimenters often find it hard to be precise about how to weigh the different criteria, the mixture plots[26,19] in Figure 3 show which designs are best for different weight combinations. The criteria values for designs in Figure 3 and their area of weight combinations are shown in Table I. Designs are labeled with their #*WP*s in the corresponding regions. Each point within the three-component simplex corresponds to a weight combination with the three entries summing to one. The vertices correspond to optimizing on a single criterion and the edges optimize using just two criteria. Unless the weights of interest happen to be close to the edge of a region, designs with bigger regions are more robust to uncertainty associated with weight specification.

If there is some interest in each criterion, we think it is unlikely that a practitioner would assign a weight of less than 10% to any single criterion. Also, since designs that are sensitive to small weight changes are generally undesirable, we focus on designs with at

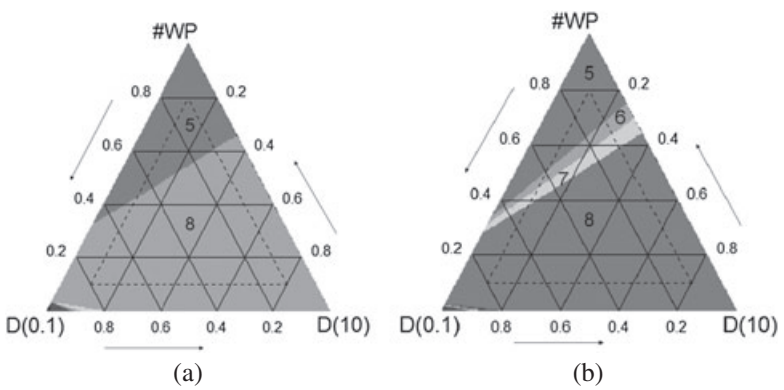


Figure 3. Mixture plots of selected designs for different weightings of the three criteria: D(0.1), D(10) and #*WP*, based on using (a) the L_1-norm and (b) the L_1-norm on the log scale for 16-run SPD example. Designs with at least 10% weight for each criterion and at least 1% of weighting area are labeled with #*WP* shown in corresponding regions

Table I. The criteria values and the area of weight combinations in Figure 3 for the optimal designs selected using the L_1-norm and the L_1-norm on the log scale based on D(0.1), D(10) and #*WP* for the 16-run SPD example

			Area (%)	
#WP	Rel. D(0.1)-eff. (raw)	Rel. D(10)-eff. (raw)	L_1-norm	L_1-norm on log scale
5	92.78% (14.84)	75.07% (23.20)	23.67	16.94
6	94.85% (15.18)	82.60% (25.52)	---	4.33
7	96.96% (15.51)	90.88% (28.08)	---	4.61
8	99.12% (15.86)	100% (30.90)	75.86	73.60
9	99.25% (15.88)	99.40% (30.71)	0.25	0.23
10	99.39% (15.90)	98.38% (30.40)	0.08	0.09
11	99.53% (15.92)	97.06% (29.99)	0.05	0.05
12	99.66% (15.95)	95.54% (29.52)	0.26	0.12
13	99.75% (15.96)	86.34% (26.68)	---	0.01
14	99.84% (15.97)	74.01% (22.87)	---	---
16	100% (16)	51.78% (16)	0.03	0.00

least 1% of the weighting area. When using the L_1-norm, the 5 and 8 WP designs are selected with the 5-WP design being cheapest and hence optimal when low cost is valued most. The design with 8 WPs is optimal for more than 75% of weight combinations. This design is the best when cost is weighted between 3% and 33% regardless of the variance ratio. When the weight for cost becomes higher (between 33% and 65%), this design is best when D(10) is weighted more. This highlights that more expensive designs (with more *WPs)* can lead to different level of improvements in D-efficiency for different variance ratios. This reinforces the importance of incorporating the variance ratio uncertainty into the decision-making process, as an imprecisely specified variance ratio could lead to a sub-optimal design choice.

When the L_1-norm on the log scale is used, the 8-WP design is best for 73.6% of the weight combinations. When minimizing cost is weighted more, cheaper designs with 7, 6 and 5 WPs are selected. This is intuitive since the L_1-norm on the log scale penalizes poor values more severely and selects designs that perform moderately well for all criteria. The choice of how to combine the multiple criteria into a single summary is an important one and should be made by the practitioners based on experiment priorities. The subjective user's choice of scaling scheme also impacts which designs are selected and their corresponding weights. A sensitivity analysis to assess this can be quickly conducted with little extra computational cost.

Figure 4 shows the trade-offs[19] between criteria among the designs selected by the adapted Utopia point approach for both DF forms. The designs are sorted by #*WP* (smaller #*WP* on the right). Designs selected by the L_1-norm (5 and 8 WPs) are shown with black symbols, while those selected by the L_1 norm on the log scale (5, 6, 7 and 8 WPs) have bigger symbols. The 8-WP design does extremely well for both D(0.1) and D(10) and moderately well on cost. The 5-WP design has low cost but is worst for D(0.1) and D(10). Designs with 6 and 7 WPs both reduce the cost with relatively small sacrifice in D(0.1) but substantial loss in D(10). Compared with the 8-WP design, designs with 9 to 12 WPs have little gains in D(0.1) and small sacrifice in D(10) by increasing the overall cost. Designs with 13 and 16 WPs have negligible gains in D(0.1) but big sacrifices in D(10) and cost. Hence, designs with 9 and more WPs are unlikely choices unless D(0.1) has extremely high weight. For the remainder of the paper, we assume that the practitioner wants to protect against very poor performance on any criteria, and hence focuses on the L_1-norm on the log scale with the scaling method described above.

Since the experimenter can only run a single design, it is helpful to understand individual design performance relative to the best possible choice across the entire range of different weights. Figure 5 shows the synthesized efficiency plot[21] for designs with 5 to 8 WPs based on the L_1-norm on the log scale. The relative efficiency when the three criteria are synthesized into an overall desirability is compared with the best value possible from any allowable design. Different shades of white-gray-black represent high to low synthesized efficiency. The chosen shading has 20 gradations, so each shade represents a 5% range of synthesized efficiency. For a white region, the design is at least 95% efficient relative to the best possible choice. The lightest gray corresponds to 90% to 95% efficiency. In Figure 5, the 5-WP design has above 85% efficiency for half of the possible weights, especially when #*WP* is given more than half of the total weight. However, its efficiency is below 75% when D(10) is weighted more than 50%. The 7-WP design has above 75% efficiency for all weight combinations. The 8-WP design is at least 95% efficient for most of possible weights except for a small region when the cost (#*WP*) is weighted extremely high. Unless cost is the dominating objective, the 8-WP design is highly efficient for almost all weights and is quite robust to different variance ratios, and therefore is the leading choice.

From this example, it is straightforward and computationally efficient to conduct a thorough study of how the uncertainty associated with the variance components ratio affects our final solution by using the Pareto front approach. Taking into account variance ratio uncertainty makes a difference to which design is preferred, and the Pareto front allows the experimenter to reach an informed and justifiable decision tailored to the particular problem of interest.

5. Example with flexible design size and different approaches for incorporating cost

In many applications, there is some flexibility with the overall experimental cost, with an understanding that spending more on an experiment with more data typically leads to improvement in performance. Now, we return to our example where the overall design size is allowed to be ranging from 12 to 16 observations. The previous section explores a simpler example when the experimenter

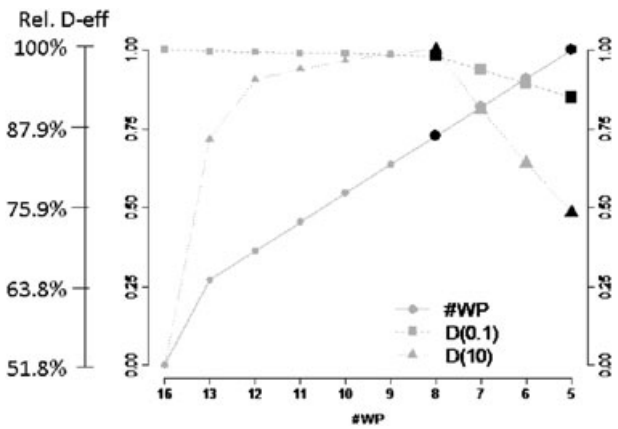


Figure 4. The trade-off plot of selected designs based on D(0.1), D(10) and #*WP* using both the L_1-norm and the L_1-norm on the log scale for the 16-run SPD example. The designs labeled in Figures 3(a) and 3(b) are highlighted in black and bigger symbols, respectively

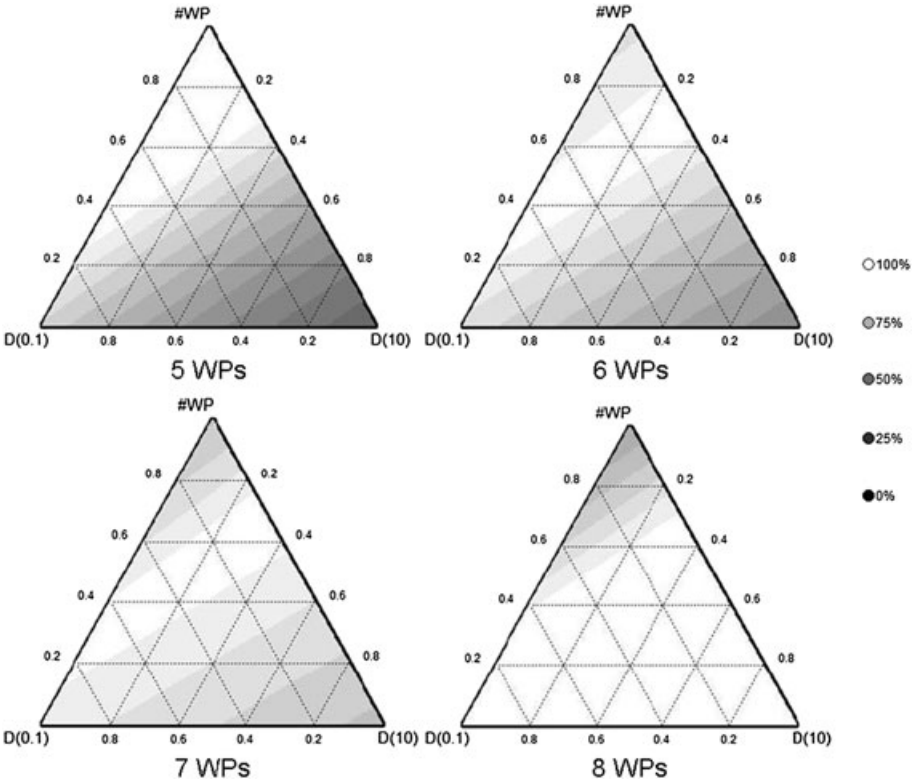


Figure 5. The synthesized efficiency plot of some selected designs (with 5–8 WPs) for different weightings of the three criteria: D(0.1), D(10) and #*WP* based on the L_1-norm on the log scale for the 16-run SPD example

simply decides to use all available resources and collect the maximal number of observations. However, for many sequential experiments, if conducting a smaller design in the early stage can obtain adequate estimates, the experimenter may decide to save some resources for later stages. This section considers the general case where a flexible design size is allowable. The goal of the experiment is still precise estimation of the specified model when considering the uncertainty with the specified variance ratio while balancing cost effectiveness. For SPDs, since both *N* and #*WP* may affect the experimental cost to different degrees depending on the cost structure, we consider different ways of incorporating cost into the design criteria and assessing their impacts on the final decision.

As previously discussed, the precision of model estimation taking into account variance ratio uncertainty is evaluated based on D-efficiency at the maximum and minimum of the possible values, i.e. D(0.1) and D(10). However, cost can be summarized in different ways. A popular cost formula by[25,15] for SPDs is given by Cost = $\#WP + c_r N$, where c_r is the cost ratio of obtaining a sub-plot observation relative to a WP observation, which is useful when both *N* and #*WP* focus on the same aspect of cost (such as financial) and c_r is known. Depending on the value of c_r for a particular experiment, #*WP* and *N* may suggest different impacts on the total cost and thus the selected design. However, when #WP and *N* affect different aspects of cost (e.g. time and money), it is not suitable to combine them, and it makes more sense to consider them separately in decision-making. Below, we consider two different ways of incorporating cost, which correspond to design selection using three or four criteria. The commonly used cost-adjusted criterion based on a per unit cost summary is not considered here since combining different design characteristics is against the general spirit of examining trade-offs with the Pareto front approach.

To construct the overall Pareto front, we initially build Pareto fronts for each combination of *N* and #*WP*. Since both *N* and #*WP* are discrete, the Pareto front based on D(0.1) and D(10) can be constructed from the set of Pareto fronts for all possible combinations of *N* and #*WP*. The advantage of building a higher dimensional front from a collection of lower dimensional fronts is reduced computational complexity and time. It also provides the opportunity to use parallelization to further reduce computing time. Pareto fronts based on three and four criteria can be obtained from combinations of *N* and #*WP* 2-criteria fronts.

Figure 6(a) shows the paired D(0.1) and D(10) values for Pareto fronts for all combinations of *N* and #*WP*. It contains 64 designs in total, and all are on the four criteria Pareto front for D(0.1), D(10), *N* and #*WP*. In Figure 6(a), the bottom and left axes show the relative D(0.1) and D(10) efficiencies compared to their optimal values, respectively. The top and right axes show the raw values. Line types represent different *N* values, and symbols represent different #*WP* values. Several patterns emerge from Figure 6(a). First, for each fixed *N*, the Pareto fronts for the remaining three criteria have a consistent pattern, where both D(0.1) and D(10) increase as #*WP* increases to a certain point, with further increases in #*WP* leading to only small improvement in D(0.1) and substantial deterioration in D(10) as we approach the CRD. When #*WP* is small there is a simple trade-off between #*WP* and D-efficiency regardless the variance ratio. As #*WP* increases, there is an extra trade-off between D(0.1) and D(10). Second, as *N* increases, the Pareto front shows substantial improvement in both D-efficiencies. This suggests strong benefits for considering larger *N* values, as the trade-off between *N* and other criteria is quite pronounced. Third, the range of efficiencies on the Pareto front differs substantially for the high and low ratios:

the relative D(0.1)-efficiency ranges from just below 70% to 100%, while the worst relative D(10)-efficiency is below 40%. This indicates a bigger impact on D(10) than D(0.1) from changing the design size N and #WP.

Now we consider the first scenario when both N and #WP affect the total cost through the form Cost = #$WP + c_r N$. For our example, adding a WP costs the same as collecting an additional sub-plot observation, and hence the cost ratio is $c_r = 1$. The Pareto front using the three criteria D(0.1), D(10) and total cost consists of 24 designs from the original 64 designs in Figure 6(a). All 16-run designs in Figure 6(a) are on the 3-criteria front. For $N = 12–15$, only designs with no more than 8 WPs are included. Then 19 designs are further selected from the front by using the adapted Utopia point approach based on the L_1-norm on the log scale with the user-specified scaling with 0 as the worst of the both D(0.1) and D(10)-efficiency values. Figure 7(a) shows the mixture plot for the 19 designs with 8 designs being best for at least 1% of weights labeled using the N(#WP) notation. For these 8 more robust designs which cover more than 95% of the total weighting area, their criteria values and proportion of areas are summarized in Table II. The 8-WP 16-run design is again the dominant choice as it is optimal for more than 70% of the possible weights. It is the best design when cost is weighted less than 30% regardless of the variance ratio. It is also optimal when all three criteria weighted equally (cost, D(0.1), D(10)) = (1/3,1/3,1/3), or when the cost and D-criterion are valued equally and the weights are split evenly between the two different variance ratio D-efficiencies (1/2,1/4,1/4). Figure 7(b) shows the trade-off plot for the 19 designs in Figure 7(a), with the labeled designs in larger black symbols. The designs are sorted from left to right for increasing cost. Generally, higher cost is associated with the better D(0.1) efficiency. However, for the larger variance ratio, cost has a more complicated impact on the D-efficiency depending on N and #WP. Figure 7(c) shows the synthesized efficiency plot for the 16-run 8-WP design. The 16-run 8-WP design has above 95% efficiency for 82% of weights and is at least 50% efficient for all possible weights. Hence, unless the cost is weighted more than 60%, the 16-run design with 8 WPs has best D-efficiency regardless of the variance ratio.

Next, we consider an alternate way of handling the cost, with N and #WP as two separate criteria. This is useful for situations when N and #WP affect different aspects of cost and are not suitable to be combined. For example, the financial cost of running the experiment may come mainly from collecting data from individual observations (N), while #WP mainly affects the time to run

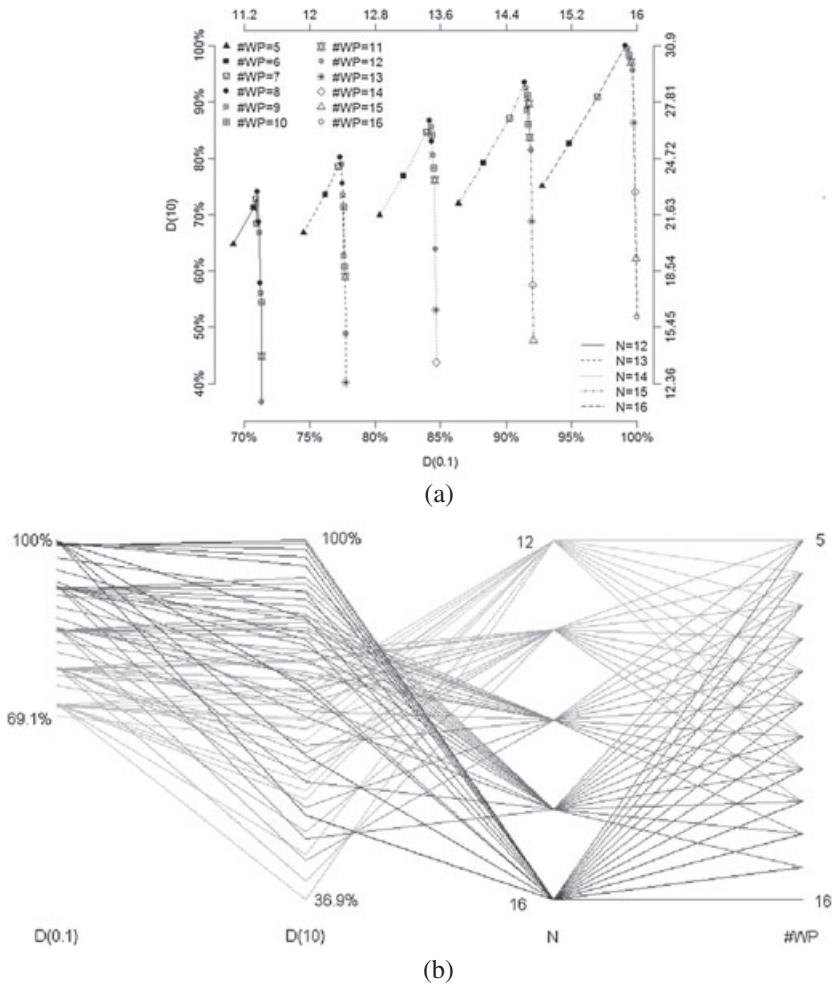
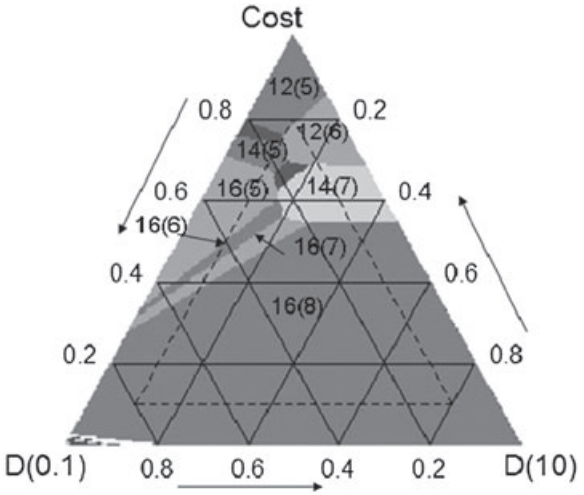
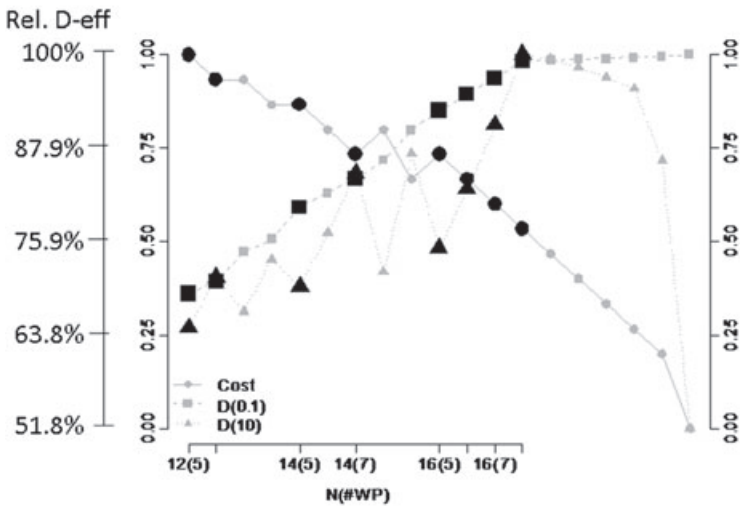


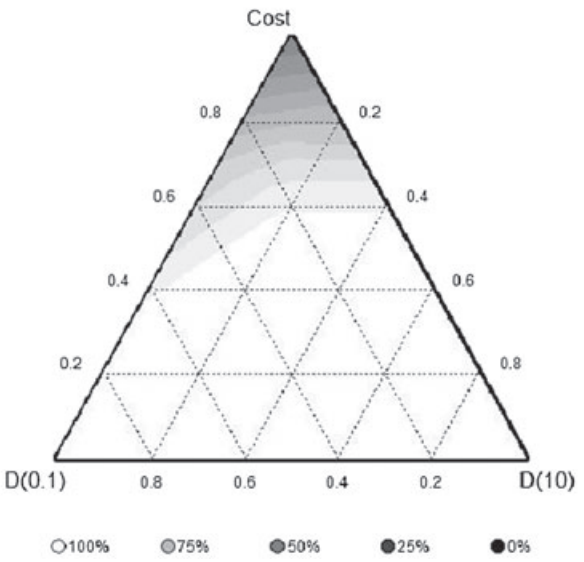
Figure 6. (a) Paired values of (D(0.1) and D(10)) for all designs on the Pareto fronts for fixed combinations of N and #WP for the example with flexible #WP and N ranging from 12 to 16. Note that all 64 designs are on the Pareto front based on all four criteria: D(0.1), D(10), N and #WP. (b) Parallel plot for the 64 designs with best values for all criteria shown at the top. The amount of crossing of the lines indicates the amount of trade-off between the adjacent criteria. The black-to-gray shading shows different level based on D(0.1) with darker gray representing higher efficiency



(a)



(b)



(c)

Figure 7. (a) The mixture plot for designs selected based on using the L_1-norm on the log scale for the three criteria: D(0.1), D(10) and total cost when $c_r = 1$ for the flexible design size example. Designs with at least 10% weight for all three criteria and at least 1% of the total simplex area are highlighted with N(#*WP*) shown in corresponding regions. (b) The trade-off plot with black symbols for highlighted designs for designs shown in (a). (c) The synthesized efficiency plot for the leading choice, 16-run with 8 WPs

Table II. The criteria values and the area/volume of weight combinations for the optimal designs selected using the L_1-norm on the log scale based on 3 criteria (D(0.1), D(10), and total cost) and 4 criteria (D(0.1), D(10), N, #*WP*) for the flexible size SPD example. The designs are sorted based on their corresponding volume based on 4 criteria. The designs corresponding to at least 1% of the total area based on 3 criteria (containing more than 95% area in total) are included in the table

Design						
N	#WP	Cost ($c_r = 1$)	Rel. D(0.1)-Efficiency (raw)	Rel. D(10)-Efficiency (raw)	Area based on 3-Criteria (%)	Volume based on 4-Criteria (%)
12	6	18	70.68% (11.31)	71.31% (22.03)	3.51	21.47
15	8	23	91.43% (14.63)	93.38% (28.86)	<1	18.00
12	5	17	69.14% (11.06)	64.81% (20.03)	4.18	14.63
14	7	21	83.94% (13.43)	84.69% (26.17)	5.41	8.91
12	8	20	70.98% (11.36)	74.06% (22.88)	<1	7.74
14	8	22	84.16% (13.47)	86.74% (26.80)	<1	5.41
13	8	21	77.34% (12.37)	80.26% (24.80)	<1	3.67
13	7	20	77.16% (12.35)	78.65% (24.30)	<1	3.37
15	5	20	86.37% (13.82)	71.97% (22.24)	<1	3.02
14	5	19	80.33% (12.85)	69.95% (21.62)	1.56	2.68
15	7	22	90.26% (14.44)	87.12% (26.92)	<1	2.61
16	8	24	99.12% (15.86)	100% (30.90)	71.82	1.93
12	7	19	70.83% (11.33)	72.76% (22.48)	<1	1.61
13	5	18	74.54% (11.96)	66.85% (20.66)	<1	1.51
16	5	21	92.78% (14.84)	75.07% (23.20)	5.90	<1
16	7	23	96.96% (15.51)	90.88% (28.08)	2.61	<1
16	6	22	80.32% (12.85)	69.95% (21.62)	2.21	<1

the experiment. In this case, it would be hard to equate these two different aspects and makes more sense to treat them as separate criteria for decision-making. The 4-criteria Pareto front shown in Figure 6(a) can be complemented with the parallel plot in Figure 6(b). The best values for each criterion are placed on the top, and the range of each criterion is labeled. The scaling for D(0.1) and D(10) is based on the worst of the relative D-efficiency from either criteria. How much crossing of lines between adjacent criteria indicates the amount of trade-off. Generally, the criteria are ordered from left to right with more important criteria on the left. To further track connections between designs across all of the criteria, grayscale is used based on D(0.1) with darker color for higher values. The *N* and #*WP* have a bigger impact on D(10) with the range of relative D-efficiency values twice as wide as for D(0.1). The sharpest trade-off is between D-efficiencies and *N*, with the best designs for D-efficiency corresponding to the largest design size. There is also considerable crossing between D(0.1) and D(10), indicating the importance of considering variance ratio uncertainty for SPD optimization. The parallel plot scales well to higher dimensions.

To use the adapted Utopia point approach for the 4-critiera case, some changes to the two and three criteria graphical methods are needed. Figure 8(a) illustrates the tetrahedron of all possible 4-criteria weight combinations, with every point corresponding to a weight combination with the sum of all entries equal to one. The vertices optimize based on a single criterion, the edges consider pairs of criteria and the faces simultaneously consider three criteria. The two sample slices in Figure 8(a) from left to right represent possible weight combinations for a fixed weight of 0.2 and 0.6 for *N*, respectively. When the weight for *N* equals 0.2, the sum of the other three criteria weights is 1–0.2 = 0.8. When the weight for *N* is 0.6, the sum of the weight for the other three criteria is 0.4. Hence, the total area of the slice with a fixed weight on one criterion becomes proportionately smaller as its weight increases.

For the L_1-norm on the log scale, 49 designs are identified as optimal for at least one set of weights. The 14 designs with at least 1% of the total volume of the tetrahedron are identified and shown in Table II in decreasing order of volume. Five of the 14 designs are also labeled in Figure 7(a) and are optimal for at least 1% of weights when using three criteria, D(0.1), D(10) and the total cost. The remaining three labeled designs from Figure 7(a) are shown at the bottom of Table II. With two criteria focused on cost-related aspects, a larger number of cheaper designs (with smaller *N*) are selected. The three highest volume designs correspond to 54% of the total weights. Two of these are 12-run designs with 5 and 6 WPs. The second biggest area corresponds to a 15-run 8-WP design, which is optimal for 18% of weight combinations. Only one 16-run design (#*WP* = 8) is selected among the 14 designs, which is also a leading choice for the three criteria scenario. Recall the L_1-norm on the log scale metric severely penalizes design with worst performance on at least one of the criteria, and hence large designs are less appealing. This is a subjective choice of the experimenter and needs to be tailored to particular priorities of the study. A sensitivity analysis can be helpful to see the impact from different scaling and metric choices.

To understand where designs are best, Figure 8(b) shows the optimal designs for different slices of weight combinations with a fixed weight for *N* = 0, 0.2, 0.4 and 0.6 from the left to the right. If the experimenter has particular weights for *N* in mind, the figure is easily adapted to highlight these slices. The designs in these four slices with at least 1% of total volume are labeled with *N*(#*WP*). When bigger designs are not penalized (weight for *N* equals 0), the optimal designs all have *N* = 16. Hence, this slice in Figure 8(b) is the same as Figure 3 (b). When #*WP* is weighted at least 40%, only 12-run designs are selected with #*WP* = 5–8. Note that the 15-run 8-WP design as the second

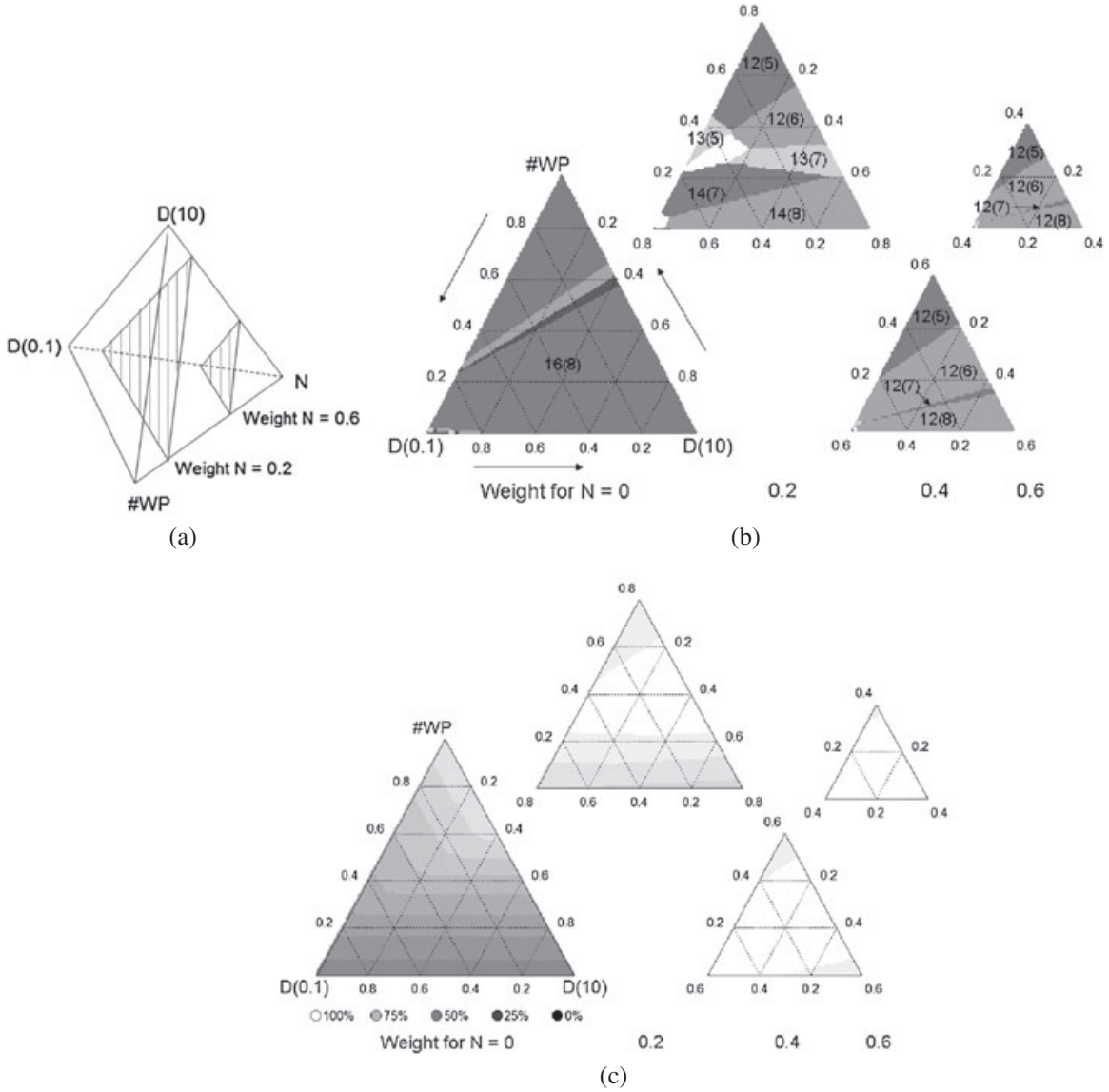


Figure 8. (a) Illustration of the tetrahedron of all possible weight combinations for 4 criteria scenario, with sample slices shown for 3 criteria with a fixed weighting of 0.2 and 0.6 for N. (b) The optimal designs for different weight combinations with fixed weight for N as 0, 0.2, 0.4 and 0.6 from the left to the right based on using the L_1-norm on the log scale for the flexible design size example. Designs corresponding to at least 1% of the total tetrahedron volume are labeled by N(*#WP*) in corresponding regions. (c) The synthesized efficiency plot for the 12-run 6-WP design with fixed weight for N as 0, 0.2, 0.4 and 0.6

biggest region in the tetrahedron is not shown in Figure 8(b). This is due to the particular choice of slices selected. This design is optimal for weight combinations with weights for *N* between 0 and 0.2.

The adaptation of the synthesized efficiency plot to four criteria is similar to the mixture plot, and is illustrated in Figure 8(c) using the 12-run 6-WP design. It is above 80% synthesized efficient when *N* is weighted at 20% and has above 90% efficiency when *N* is weighted more. Efficiency plots for a few other designs in Table II are provided in Appendix B. The 15-run 8-WP design has high efficiency for some region when *N* weighted between 0 and 0.2, but relatively lower efficiency when *N* is weighted more than 0.2. The 14-run 7-WP design is highly efficient when the weight for *N* is around 20% and *#WP* is weighted no more than 40%. The 13-run 7-WP design is at least 70% efficient for weight combinations with no more than 40% weight for *N*. The leading choice for the 3-criteria case, the 16-run 8-WP design (not show in Appendix B), only has high synthesized efficiency for all weight combinations with zero weight for *N*, but has extremely low efficiency for weightings with higher weight for *N*. Overall, when *N* is not strongly down-weighted, the 12-run 6-WP design has high synthesized efficiency for the largest region of weights and hence is a leading choice for many cases.

Figure 9 shows the geometric representation of the leading designs for three and four criteria. Each design is represented by a cube with 8 candidate locations represented by the circles. Each circle represents one observation with the WP index number shown in the circle. The 16-run 8-WP design (a leading choice for three criteria) is quite symmetric with each WP having two observations with both high and low levels for the sub-plot factor and each of the four level combinations for the two WP factors having observations split evenly between two WPs. The 12-run 6-WP design (good for four criteria) also has two observations with both high and low levels for the sub-plot factor for each of the WP. The 6 WPs are split evenly for only one of the two WP factors.

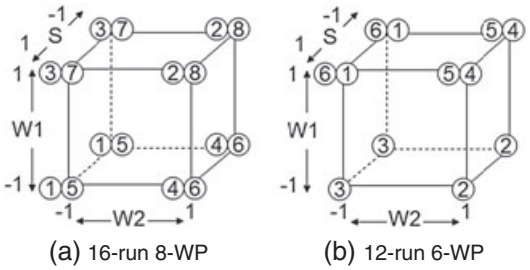
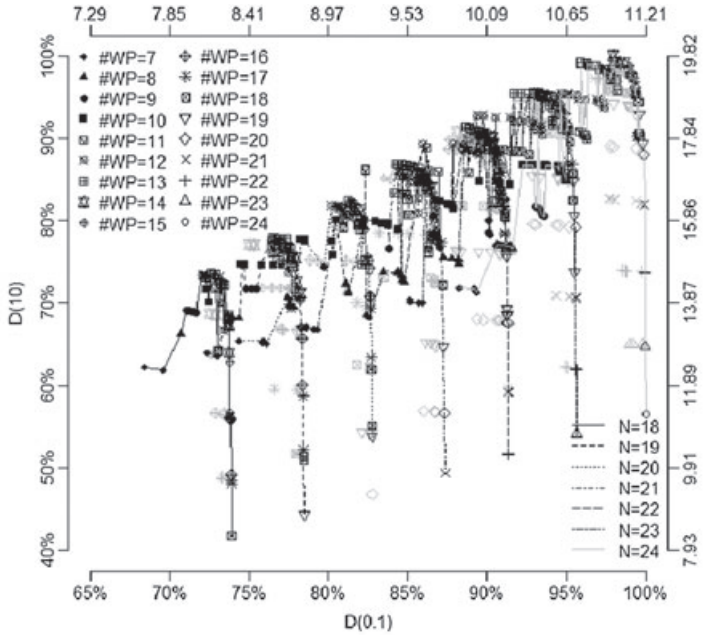
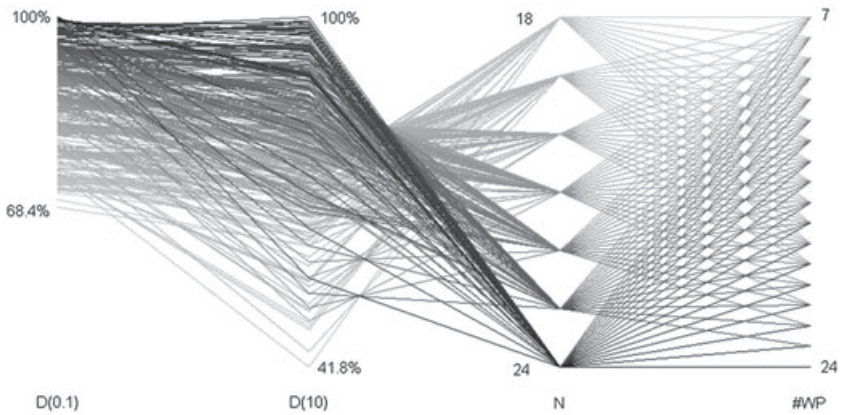


Figure 9. Geometrical representations of the leading design choices, (a) 16-run 8-WP and (b) 12-run 6-WP based on considering 3 and 4 criteria, respectively. Each design is represented by a cube with 8 candidate locations represented by the circles. Each circle represents one observation with the WP index number shown in the circle

From the above example, we see how *N* and #*WP* affect the design performance, and how different ways of incorporating cost in the formulation of design criteria can lead to different understanding of the relative advantages of the choices. We believe the decision of which criteria combination to select should be made based on the goal of the study, and how much flexibility there is to change overall cost. By considering different approaches, the experimenter can better understand the impact of subjective



(a)



(b)

Figure 10. (a) The paired values of (D(0.1) and D(10)) for all designs on the Pareto fronts of fixed combinations of N and #*WP* for the example with N = 18–24 and quadratic model. Note that all 486 designs on the Pareto front based on four criteria: D(0.1), D(10), N and #*WP* are displayed in black, and the remaining 297 designs not on the front are shown in gray. (b) Parallel plot for the 486 designs on the Pareto front based on the four criteria: D(0.1), D(10), N and #*WP*

choices and sources of uncertainty. The Pareto front method enhanced with graphical summaries provides efficient tools for deriving a comprehensive, defensible and tailored solution.

6. Other scenarios with different patterns of trade-offs

To illustrate that different relationships and amounts of trade-off can exist across variance ratios, we explore a few other examples focusing on D(d_{lower}) and D(d_{upper}) criteria. For a 2 WP and 1 sub-plot SPD with $N = 18–24$ and #*WP* =7-*N* based on a quadratic model with the variance ratio between 0.1 and 10, Figure 10(a) shows the paired D(0.1) and D(10) values for the collection of designs on the Pareto fronts of fixed combinations of *N* and #*WP*. Note the minimal #*WP* is chosen to be 7 to ensure all six WP model terms and WP variance are estimable. There are 783 designs on the set of all 2-D Pareto fronts based for combinations of *N* and #*WP*. From these designs, the 486 designs on the Pareto front based on all four criteria: D(0.1), D(10), *N* and #*WP*, are displayed in black. Various line types and symbols show different values of *N* and #*WP*, respectively. Compared with Figure 6(a), an obvious distinction between the two fronts is the increased complexity. This front also shows more trade-off between *N* and D-efficiency, with different amounts of trade-off between D(0.1) and D(10) for small and large #*WP*. The worst D-efficiency for d = 0.1 and 10 also varies considerably. There is also more overlapping of the fronts from different combinations of *N* and #*WP* in Figure 10(a) and more crossing of lines between criteria in the parallel plot in Figure 10(b).

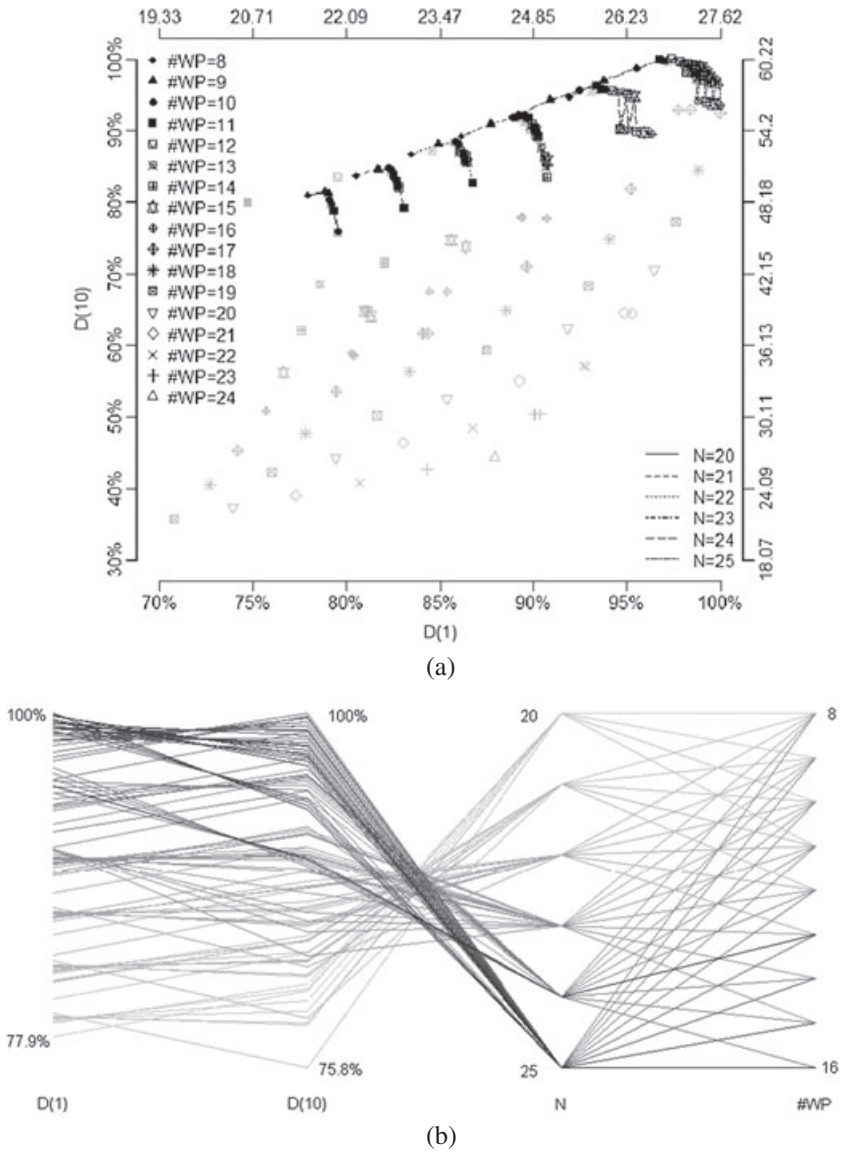


Figure 11. (a) The paired values of (D(1) and D(10)) for all designs on the Pareto fronts of fixed combinations of N and #*WP* for the continuous five-factor example with N = 20–25 and linear model with all two-factor interactions. Note that all 102 designs on the Pareto front based on four criteria: D(1), D(10), N and #*WP* are displayed in black, and the remaining designs not on the front are shown in gray. (b) Parallel plot for the 102 designs on the Pareto front based on the four criteria: D(1), D(10), N and #*WP*

Figure 11 shows the collection of 2 criteria (D(1) and D(10)) Pareto fronts for a third example for a SPD with 3 WP and 2 sub-plot factors with flexible $N = 20–25$ and $\#WP = 8\text{-}N$ for a first-order linear model with all two-factor interactions given the d between 1 and 10. Designs on the 4-criteria front are displayed in black. A distinct change in pattern is that for each fixed N, as $\#WP$ increases after a certain point, both D(1) and D(10)-efficiency deteriorate as we approach the CRD of the same design size. Hence, the 4-criteria Pareto front consists only of designs at the top right corner for each fixed N and is simpler with less trade-off between criteria. Consequently, Figure 11(b) shows less crossing between criteria, similar ranges of relative efficiency for D(1) and D(10) and fewer choices of $\#WP$ among the contending designs. A complete set of analysis results is available in.[27]

Comparing the three examples shows that different amount of trade-offs are present between the D(d_{lower}), D(d_{upper}) and the cost-related criteria for different SPD scenarios. A priori, it is difficult to anticipate what pattern of relationship exists between the criteria and how much trade-off there is between D-efficiencies for a range of variance ratios. The nature of the relationship can be affected by many factors such as constraints on the size, complexity and region of the design, the specified model and the selected range of variance ratios. Hence, we do not anticipate a general one-size-fits-all solution for all SPD optimization problems, and it is strongly recommended to incorporate variance ratio uncertainty into design selection and evaluate its impact for a particular problem of interest.

7. Conclusions and discussion

SPDs are widely used in design of experiment applications when there are logistical or economic restrictions on randomization. Currently, standard practice focuses on finding computer-generated SPDs based on a single optimality criterion. Recently, there has been some research on simultaneously considering multiple design optimality criteria for SPDs.[28,12] However, these are problem specific and not general for any set of criteria. This paper proposes a Pareto front approach for constructing optimal SPDs based on multiple objectives and develops a new search algorithm, PAPESPD, adaptable for any user-specified design criteria.

With standard software, finding a D-optimal SPD depends on specifying the relative size of the WP to sub-plot variances. However because of a lack of available information, the exact variance component ratio is typically unknown prior to conducting the experiment. Hence, it is more realistic for experimenters to supply a range of possible values. Being forced to select a single variance ratio and optimizing based on it can lead to a sub-optimal design choice, lack of understanding about the impact of this choice and diminished performance. Also, the degree of uncertainty associated with the variance ratio may have different impacts in different scenarios. Therefore, it is important to consider and actively evaluate this uncertainty when selecting a design.

This paper proposes using the Pareto front approach based on D-optimality for the maximal and minimal possible variance ratios. We demonstrate that this method can efficiently find designs robust to the variance ratio uncertainty and have more balanced performance across the entire range of possible values for design characteristics of interest.

In addition to the variance ratio uncertainty, experimental cost is often another important aspect for design selection. We encourage evaluating cost and other design characteristics separately instead of combining them to form cost-adjusted criteria where it is difficult to disentangle their contributions. For SPDs, there are two aspects of cost (design size N and the $\#WP$) that can have a complex impact on performance. Different ways of handling cost may be appropriate in different applications. Exploring possibilities and understanding their impact on the final solution place the experimenter in a better position to make an appropriate decision based on study requirements.

The Pareto approach uses the PAPESPD algorithm to populate the Pareto front of non-dominated designs, and then reduces the front to a smaller manageable set using the adapted Utopia point approach, and finally selects a single best design based on graphical summaries for evaluating performance, trade-offs and robustness. Interpreting these comparisons should be guided by the priorities of the study and lead to an informed and justifiable decision. A major advantage of the Pareto approach over classic DF methods is the greater flexibility and computational efficiency afforded in identifying optimal designs for different weightings of criteria. It also helps quantify the robustness of these designs to different choices of weighting, scaling and metric schemes.

The PAPESPD algorithm accommodates the special constraints on randomization for SPDs and can be applied to any set of design criteria. Because of the important role of the $\#WP$ in design construction and cost evaluation, two variants of the algorithm for fixed or flexible $\#WP$ are developed. The utilization of a set of fixed weights or stratified random weights in each updating step can improve the coverage and computational efficiency of any general Pareto search algorithm.

Graphical tools summarize important design features, trade-offs between competing criteria and individual design performance relative to the best possible, and can be helpful to guide improved decision-making. The scatterplot provides an overview of the Pareto front and its location in the criteria space. For larger numbers of criteria, the parallel plot provides an efficient way of displaying the amount of trade-off between criteria. The mixture plot identifies optimal designs for different possible weight choices and shows robustness across weight combinations. The trade-off plot allows the user to make direct comparisons between several promising designs. The synthesized efficiency plot shows how individual designs perform relative to the best possible for different weightings of the criteria. In addition, new adaptations of the graphical summaries for considering four criteria are developed.

The proposed method and algorithm are demonstrated for different design sizes and cost structures. Both a fixed design size and a more general scenario where a range of design sizes is allowed were considered. With the second scenario, the experimenter has the flexibility to decide if some of the resources can be saved for later data collection stages if the required design performance can be achieved with a smaller design. In addition, it explores different ways to incorporate cost into the design selection process. When the design size and $\#WP$ are evaluated separately as different aspects of cost, the decision-making uses the Pareto approach for four criteria. By allowing some flexibility in the overall design size and cost, we can explore the potential gain in design performance with increased cost and can help the experimenters to understand the potential benefits of additional investment. While the examples are based on specific stated criteria, the methodology and tools can be easily adapted to other quantitative criteria of interest.

References

1. Goos P. The Optimal Design of Blocked and Split-Plot Experiments. Springer: New York, 2002.
2. Myers RH, Montgomery DC, Anderson-Cook CM. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley: New York, 3rd ed. 2009.
3. Huang P, Chen D, Voelkel JO. Minimum-Aberration Two-Level Split-Plot Designs. *Technometrics* 1998; **40**:314–326.
4. Bingham D, Sitter RR. Minimum-Aberration Two-Level Fractional Factorial Split-Plot Designs. *Technometrics* 1999; **41**:62–70.
5. Bingham D, Sitter RR. Design Issues for Fractional Factorial Experiments. *Journal of Quality Technology* 2001; **33**:2–15.
6. Goos P, Vandebroek M. Optimal Split-Plot Designs. *Journal of Quality Technology*2001; **33**(4):436–450.
7. Goos P, Vandebroek M. D-Optimal Split-Plot Designs with Given Numbers and Sizes of Whole Plots. *Technometrics* 2003; **45**:235–245.
8. Goos P, Vandebroek M. Outperforming Completely Randomized Designs. *Journal of Quality Technology* 2004; **36**(1):12–26.
9. Jones B, Goos P. A Candidate-Set-Free Algorithm for Generating D-Optimal Split-Plot Designs. *Applied Statistics* 2007; **56**:347–364.
10. Arnouts H, Goos P. Update Formulas for Split-Plot and Block-Designs. *Computational Statistics and Data Analysis* 2010; **54**:3381–3391.
11. Anbari FT, Lucas JM. Super-Efficient Designs: How to Run Your Experiment for Higher Efficiency and Lower Cost. *ASQC Technical Conference Transactions* 1994; 852–863.
12. Smucker BJ, del Castillo E, Rosenberger JL. Model-Robust Designs for Split-Plot Experiments. *Computational Statistics and Data Analysis* 2012; **56**:4111–4121.
13. Parker AP, Kowalski SM, Vining GG. Classes of Split-Plot Response Surface Designs for Equivalent Estimation. *Quality and Reliability Engineering International* 2006; **22**:291–305.
14. Trinca LA, Gilmour SG. Multi-Stratum Response Surface Designs. *Technometrics* 2001; **43**:25–33.
15. Liang L, Anderson-Cook CM, Robinson TJ. Cost-Penalized Estimation and Prediction Evaluation for Split-Plot Designs. *Quality and Reliability Engineering International* 2007; **23**:577–596.
16. Liang L, Anderson-Cook CM, Robinson T, Myers RH. Three-Dimensional Variance Dispersion Graphs for Split-Plot Designs. *Journal of Computational and Graphical Statistics* 2006; **15**:757–778.
17. Liang L, Anderson-Cook CM, Robinson TJ. Fraction of Design Space Plots for Split-Plot Designs", *Quality and Reliability Engineering International* 2006; **22**:275–289.
18. Robinson TJ, Anderson-Cook CM. A Closer Look at D-Optimality for Screening Designs. *Quality Engineering* 2011; **23**:1–14.
19. Lu L, Anderson-Cook CM, Robinson TJ. Optimization of Designed Experiments based on Multiple Criteria Utilizing Pareto Frontier. *Technometrics* 2011; **53**:353–365.
20. Lu L, Anderson-Cook CM. Adapting the Hypervolume Quality Indicator to Quantify Trade-Offs and Search Efficiency for Multiple Criteria Decision-Making Using Pareto Fronts. *Quality and Reliability Engineering International* 2012. http://onlinelibrary.wiley.com/doi/10.1002/qre.1464/pdf
21. 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.
22. Derringer G, Suich R. Simultaneous Optimization of Several Response Variable. *Journal of Quality Technology* 1980; **12**:214–219.
23. Eschenauer H, Koski J, Osyczka AE. Multicriteria Design Optimization: Procedures and Applications. Berlin: Springer, 1990.
24. Searle SR, Casella G, McCulloch CE. Variance Components. John Wiley & Sons, Inc: New York, 1992.
25. Bisgaard S. The Design and Analysis of 2^{k-p} x 2^{q-r} Split-Plot Experiments. *Journal of Quality Technology* 2000; **32**:39–56.
26. Cornell J. Experiments with Mixtures: Design, Models, and the Analysis of Mixture Data. Wiley: New York, 3rd ed. 2002.
27. Lu L, Anderson-Cook CM. Balancing Multiple Criteria with a Pareto Front for Optimal Split-Plot Designs, Los Alamos National Laboratory Technical Report, LAUR 11–06834. 2011.
28. Parker PA, Anderson-Cook CM, Robinson TJ, Liang L. Robust Split-Plot Designs. *Quality and Reliability Engineering International* 2008; **24**:107–121.

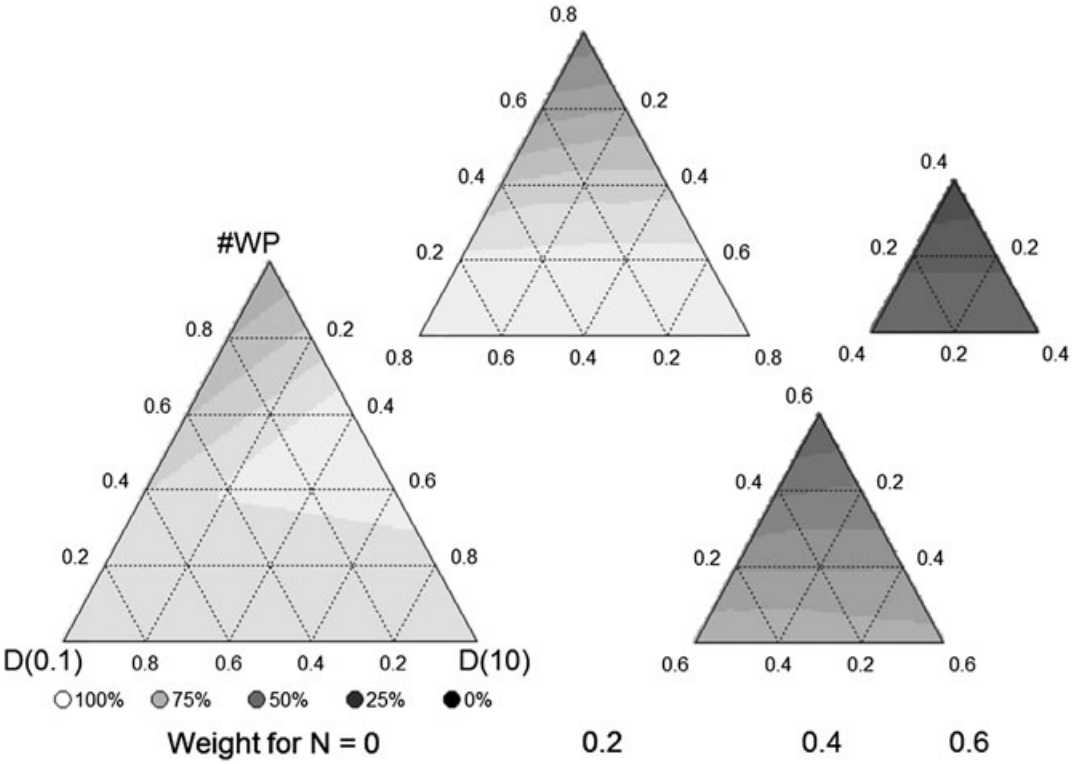
Appendix A: The PAPESPD algorithm for situations with a fixed *N* and a flexible #*WP*

The major difference between this algorithm and the fixed #*WP* in Section 3 is that it considers more possibilities for new designs with varied #*WP* in Step 2.a). All possibilities in each updating step are summarized in the following three scenarios:

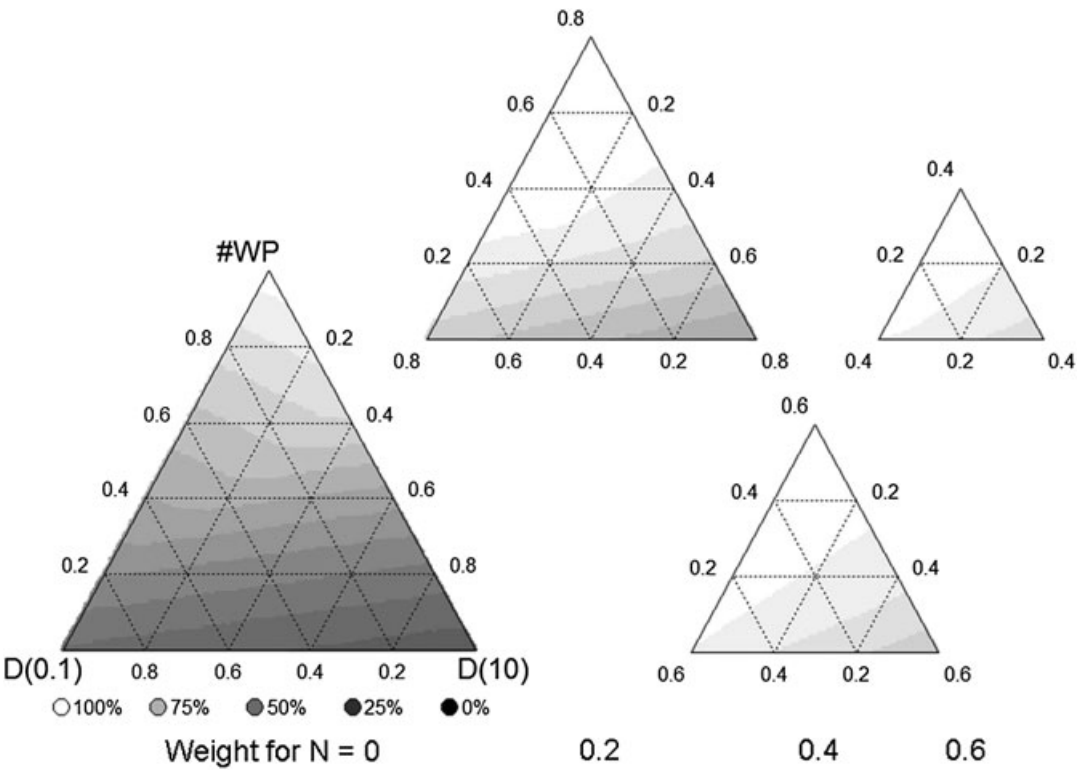
- Scenario 1: The current row is in a separate WP with only one sub-plot observation. New designs can be generated by: (i) replacing the current row with the new row (this maintains the same #*WP*), or (ii) if there are other WPs in the current design with the same WP factor levels, then add the new row to that WP (decreases #*WP* by one).
- Scenario 2: The WP containing the current row has at least two sub-plot observations and the new and current rows have the same WP factor levels. New designs which can be created by: (i) replacing the current row with the new row and keep it in the same WP (keeps same #*WP*), or (ii) if there are other WPs with the same WP factor levels as the new row, then swapping the two rows and changing the WP index to another WPs with the same factor levels (keeps same #*WP*), or (iii) removing the current row and adding the new row as a new separate WP with only a single sub-plot observation (adds one WP).
- Scenario 3: The WP containing the current row has at least two sub-plot observations and the new and current rows have different WP factor levels. New designs can be created by: (i) if there are other WPs with the same WP factor levels as the new row, then swap the two rows and change the WP index number to one of the candidate WPs (keeps same #*WP*), (ii) if there is no WP in the current design with the same WP factor levels as the new row, then create a new WP with just one observation (adds one additional WP).

Note that if #*WP* changes, it is necessary to check if the new #*WP* is allowable and discard the design if the constraint is not satisfied. Another change for this flexible #*WP* algorithm is that the random starting design in Step 1 is generated with a random number of whole plots within the range of interest.

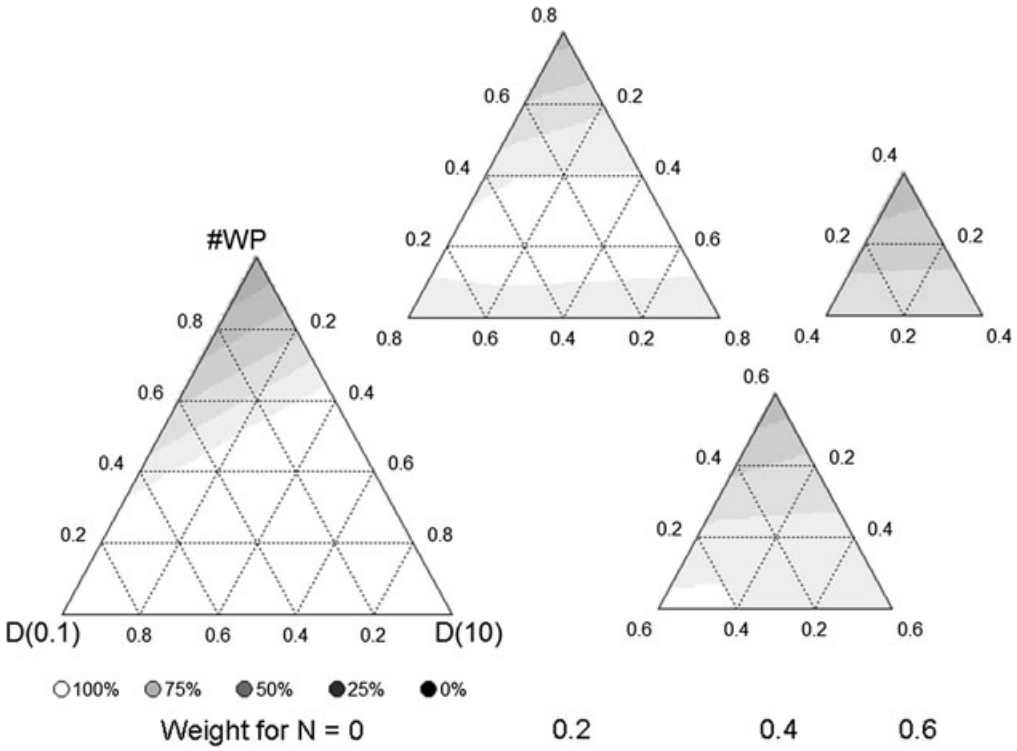
Appendix B: Synthesized efficiency plots for top designs in Table II



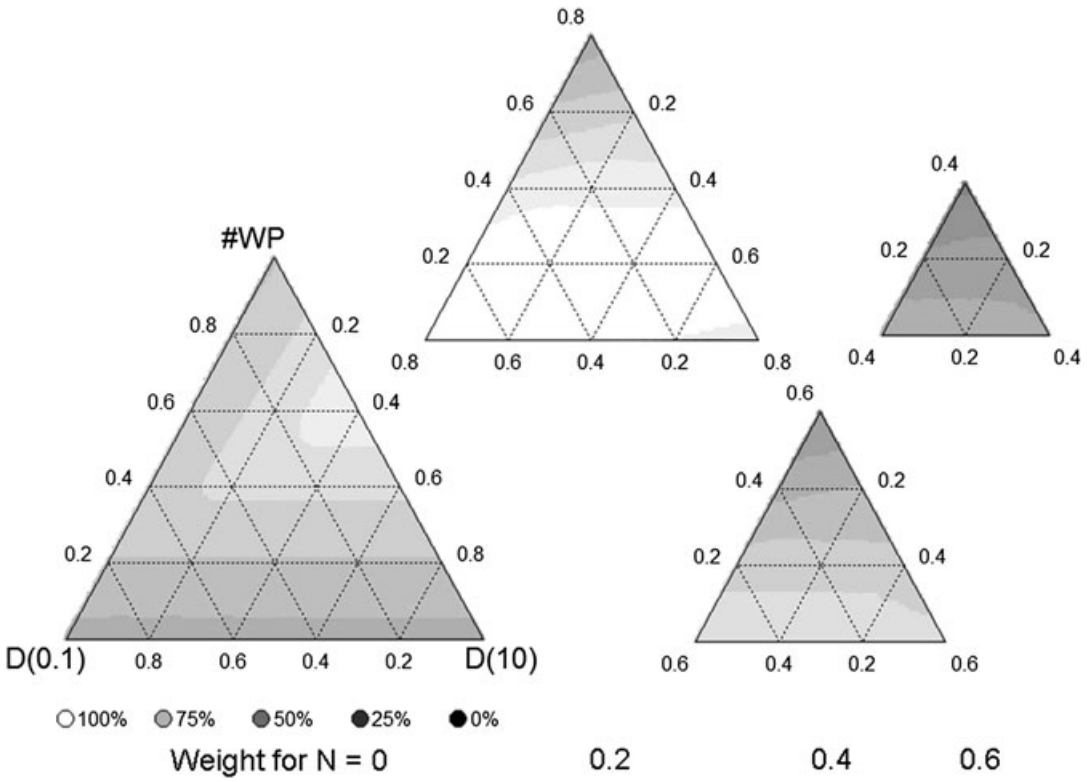
a) 15-run 8-WP design



b) 12-run 5-WP design



c) 13-run 7-WP design



d) 14-run 7-WP design