A biobjective method for sample allocation in stratified sampling

Emilio Carrizosa and Dolores Romero Morales

Facultad de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain.
email: ecarrizosa@us.es.

Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom; e-mail: dolores.romero-morales@sbs.ox.ac.uk.

Abstract

The two main and contradicting criteria guiding sampling design are accuracy of estimators and sampling costs. In stratified random sampling, the sample size must be allocated to strata in order to optimize both objectives.

In this note we address, following a biobjective methodology, this allocation problem. A two-phase method is proposed to describe the set of Pareto-optimal solutions of this nonlinear integer biobjective problem. In the first phase, all supported Pareto-optimal solutions are described via a closed formula, which enables quick computation. Moreover, for the common case in which sampling costs are independent of the strata, all Pareto-optimal solutions are shown to be supported. For more general cost structures, the non-supported Pareto-optimal solutions are found by solving a parametric knapsack problem. Bounds on the criteria can also be imposed, directing the search towards implementable sampling plans. Our method provides a deeper insight into the problem than simply solving a scalarized version, whereas the computational burden is reasonable.

Keywords: Integer programming; stratified random sampling; sample allocation; biobjective integer program; parametric knapsack problem.

1 Introduction

The sample allocation problem for stratified simple random sampling is the following: we are given a population of size N divided into n groups (strata), with population sizes $N_1, \ldots, N_n$. Simple random samples without replacement of sizes $x_1, \ldots, x_n$, are to be drawn independently from the different strata. The sampling cost within each stratum is assumed to be linear in its sample size x_i, with unit sampling cost within stratum i equal to a positive integer c_i. The total sampling cost is the sum of the sampling costs within the strata.

The drawn sample is used to estimate some parameter of the variable under study Y. Throughout this paper, we assume that the parameter to be estimated is $\overline{Y}$, the average of the variable Y in the population. Then, the parameter $\overline{Y}$ will be estimated via its Horvitz-Thompson estimator $\hat{\overline{Y}}$,

$$\hat{\overline{Y}} = \sum_{i=1}^{n} \frac{N_i}{N} \overline{y}_i, \tag{1}$$

where $\overline{y}_i$ denotes the sample average within stratum i, see e.g. [5] for further statistical details on the problem considered.

Estimator $\hat{\overline{Y}}$ is unbiased, and its variance $var(\hat{\overline{Y}})$ is given by

$$\begin{array}{rcl} var(\hat{\overline{Y}}) & = & \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2 var(\overline{y}_i) \\ & = & \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2 \left(\frac{1}{x_i} - \frac{1}{N_i}\right) \sigma_{c,i}^2, \end{array} \tag{2}$$

where $\sigma_{c,i}^2$ is the quasivariance of Y within stratum i.

We assume, as customary in the literature, that the quasivariances $\sigma_{c,i}^2$ are either known from previous similar experiments, or replaced by known upper bounds. For instance, if Y_i, the values of variable Y within stratum i, is a Boolean variable, we can use the upper bound $\frac{N_i}{N_i - 1}\frac{1}{4}$, [5].

The goal is to determine sample sizes $x_1, \ldots, x_n$ minimizing simultaneously

- The total sampling cost
- The variance of the Horvitz-Thompson estimator $\hat{\overline{Y}}$.

Two types of constraints are imposed. On the one hand, box constraints are considered on the sample sizes x_i,

$$l_i \leq x_i \leq u_i, \tag{3}$$

for positive integers $l_i \leq u_i$, for all $i = 1, 2, \ldots, n$.

Constraints (3) are motivated as follows. First, at least one element must be sampled from each stratum, since, otherwise, the expression (2) is meaningless; moreover, since sampling is without replacement, no more than N_i individuals can be sampled from stratum i.

These trivial bounds $1 \leq x_i \leq N_i$ may not be sharp enough for practical purposes. Indeed, if we are not only concerned with the variance of the estimator $\widehat{\overline{Y}}$, but also with the variance of the estimators $\overline{y}_i$ within the strata, constraints of the form

$$var(\overline{y}_i) \leq \mu_i, \tag{4}$$

for $\mu_i > 0$ given, may be imposed. Constraint (4) can also be written as

$$x_i \geq \left\lceil \frac{\sigma_{c,i}^2 N_i}{N_i \mu_i + \sigma_{c,i}^2} \right\rceil,$$

which, as asserted, yields a constraint of type (3).

On the other hand, the aim of simultaneous minimization of cost and variance may lead to sampling plans in which one of the two objectives attains a low value at the expense of a very high value on the other. To avoid this, we include also in the model target constraints in the form

$$\begin{array}{rcl} \sum_{i=1}^n c_i x_i & \leq & K^* \\ var(\widehat{\overline{Y}}) & \leq & B \end{array} \tag{5}$$

for positive K^* and B, allowed also to take the value $+\infty$.

The problem under consideration is the following biobjective nonlinear integer program

$$\begin{array}{ll} \min & \left(\sum_{i=1}^n c_i x_i, var(\widehat{\overline{Y}}) \right) \\ \text{s.t.} & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & \sum_{i=1}^n c_i x_i \leq K^* \\ & var(\widehat{\overline{Y}}) \leq B \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n. \end{array} \tag{6}$$

Define the constants

$$\begin{array}{rcl} A_i & = & \sigma_{c,i}^2 \left(\frac{N_i}{N} \right)^2,\ i = 1, 2, \ldots, n \\ B^* & = & B + \sum_{i=1}^n \frac{A_i}{N_i}. \end{array}$$

Then, by (2), (6) yields after erasing additive constant terms

$$\begin{array}{ll} \min & \left(\sum_{i=1}^n c_i x_i, \sum_{i=1}^n \frac{A_i}{x_i} \right) \\ \text{s.t.} & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & \sum_{i=1}^n c_i x_i \leq K^* \\ & \sum_{i=1}^n \frac{A_i}{x_i} \leq B^* \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n. \end{array} \tag{P_{K^*,B^*}}$$

In particular, the monotonicity of the criteria implies that, for $K^* \geq \sum_{i=1}^n c_i u_i$ and $B^* \geq \sum_{i=1}^n \frac{A_i}{l_i}$, constraints (5) are redundant, and (P_{K^*,B^*}) reduces to

$$\begin{array}{ll} \min & \left(\sum_{i=1}^n c_i x_i, \sum_{i=1}^n \frac{A_i}{x_i} \right) \\ \text{s.t.} & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n. \end{array} \tag{$P_{\infty,\infty}$}$$

The set $\mathcal{P}_{K^*,B^*}$ of Pareto-optimal solutions of (6), or, equivalently, of (P_{K^*,B^*}), is sought. We recall that a feasible solution $x = (x_1, \ldots, x_n)$ will be Pareto-optimal for (P_{K^*,B^*}) iff no feasible x^* for this problem exists satisfying

$$\begin{aligned} \sum_{i=1}^{n} c_i x_i^* &\leq \sum_{i=1}^{n} c_i x_i \\ \sum_{i=1}^{n} \frac{A_i}{x_i^*} &\leq \sum_{i=1}^{n} \frac{A_i}{x_i}, \end{aligned}$$

with at least one of the two inequalities above strict. Alternatively, we could be interested in the set of Pareto outcomes, i.e.,

$$\left\{ \left(\sum_{i=1}^{n} c_i x_i, \sum_{i=1}^{n} \frac{A_i}{x_i} \right) : \, x \in \mathcal{P}_{K^*,B^*} \right\}.$$

See e.g. [14] for further details on Pareto-optimality in general settings and [7, 16, 18] and the references therein for results and applications to other combinatorial problems.

It immediately follows that $\mathcal{P}_{K^*,B^*}$ can be obtained from the set $\mathcal{P}_{\infty,\infty}$ of Pareto-optimal solutions to $(P_{\infty,\infty})$,

$$\mathcal{P}_{K^*,B^*} = \left\{ x \in \mathcal{P}_{\infty,\infty} : \, \sum_{i=1}^{n} c_i x_i \leq K^*, \sum_{i=1}^{n} \frac{A_i}{x_i} \leq B^* \right\}. \tag{7}$$

Hence, we can restrict ourselves to the study of $(P_{\infty,\infty})$.

Although a full description of the Pareto-optimal set for multiobjective integer problems is usually extremely hard, even in the linear case, [7, 16], it turns out that the structure of $(P_{\infty,\infty})$ is simple enough to allow us to obtain an easy characterization of $\mathcal{P}_{\infty,\infty}$, and, by (7), of $\mathcal{P}_{K^*,B^*}$, under certain conditions usually held in practice. Moreover, when such conditions are not fulfilled, standard Branch-and-Bound techniques can be customized to construct $\mathcal{P}_{K^*,B^*}$. As far as the authors are aware, this is the first time this sample allocation problem is directly addressed as a biobjective problem. See Section 3 and [2, 12] for references on related single-objective models. We will illustrate with a real-world database that finding $\mathcal{P}_{K^*,B^*}$ provides a deeper insight into the problem than simply solving a scalarized version, whereas the computational burden is reasonable.

In what follows we assume that the threshold values K^* and B^* are such that (P_{K^*,B^*}) is feasible. This can be tested by solving, e.g. with the technique described in Section 3.1, the problem

$$\begin{array}{ll} \min & \sum_{i=1}^{n} \frac{A_i}{x_i} \\ \text{s.t.} & \sum_{i=1}^{n} c_i x_i \leq K^* \\ & l_i \leq x_i \leq u_i, \; i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z}, \; i = 1, 2, \ldots, n, \end{array}$$

and checking whether its optimal value does not exceed B^*.

The remainder of the paper is structured as follows. In Section 2 we consider one of the most usual procedures for generating elements of $\mathcal{P}_{\infty,\infty}$, namely the weighting approach. It turns out that the set of optimal solutions of such problems, the set of supported solutions, can be easily characterized. Sections 3 and 4 address the problem of describing the non-supported Pareto-optimal solutions of (P_{K^*,B^*}). First arbitrary cost structures are considered, and a branch-and-bound algorithm is designed. Finally, the particular case in which the costs are independent of the strata is studied, showing that the supported solutions are the only Pareto-optimal solutions. Numerical experiments with a realworld database are presented in Section 5. The paper ends with a discussion on extensions and lines of further research.

2 Supported solutions for $(P_{\infty,\infty})$

A very popular scalarization strategy in Multiple-Objective optimization is the so-called *weighting method*, in which the objectives are linearly aggregated: ν, $0 < \nu < 1$ is given, and therefore $(P_{\infty,\infty})$ is replaced by the scalar problem

$$\begin{array}{ll} \min & (1-\nu)(\sum_{i=1}^{n} c_i x_i) + \nu \sum_{i=1}^{n} \frac{A_i}{x_i} \\ \text{s.t.} & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n, \end{array}$$

or, setting $\lambda := \frac{\nu}{1-\nu} \in (0, +\infty)$, by

$$\begin{array}{ll} \min & \sum_{i=1}^{n} c_i x_i + \lambda \sum_{i=1}^{n} \frac{A_i}{x_i} \\ \text{s.t.} & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n. \end{array} \tag{8}$$

By varying λ in the interval $(0, +\infty)$, the set of optimal solutions obtained this way would yield $\mathcal{S}_{\infty,\infty}$, the so-called set of *supported* solutions of $(P_{\infty,\infty})$.

Obtaining the full set $(P_{\infty,\infty})$ of supported solutions has important practical consequences. Indeed, if, as frequently done in multiple-objective problems, the final sampling allocation plan is chosen by minimizing a weighted average of the estimator variance and the cost, or, in other words, by solving a problem of type (8), we know that such a plan is not only Pareto-optimal, but also supported. Hence, if, as a preprocessing step, $(P_{\infty,\infty})$ is obtained, then an optimal solution to (8) can be obtained from $(P_{\infty,\infty})$ by complete enumeration.

On the other hand, when the full set of Pareto-optimal solutions is sought, obtaining first the supported solutions and later the remaining Pareto-optimal ones may lead to important savings in computing times, as described for other multiple-objective combinatorial problems e.g. in [17, 18], and shown for this problem in Section 5.

Moreover, as described in Section 4, in the important case in which all costs are equal, all Pareto-optimal solutions are supported, thus a description of the supported solutions yields a description of the full set of Pareto-optimal solutions.

Two well-known properties of the supported solutions, stated in the following theorem, will be used in the sequel.

Theorem 2.1 *Let $x^* \in \mathcal{S}_{\infty,\infty}$. One has:*

1. *$x^* \in \mathcal{P}_{K,B}$ for any $K \geq \sum_{i=1}^n c_i x_i^*$ and $B \geq \sum_{i=1}^n \frac{A_i}{x_i^*}$*
2. *x^* solves the scalar problem*

$$\begin{array}{ll} \min & \sum_{i=1}^n \frac{A_i}{x_i} \\ s.t. & \sum_{i=1}^n c_i x_i \leq \sum_{i=1}^n c_i x_i^* \\ & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n. \end{array}$$

Now we address the problem of describing $\mathcal{S}_{\infty,\infty}$, which, by Theorem 2.1, consists of Pareto-optimal solutions of $(P_{\infty,\infty})$. To obtain such description, we first explore the structure of (8) for a choice of $\lambda > 0$.

As already discussed some years back by Aggarwal [1], (8) is a nonlinear convex separable integer problem with just box constraints, which can be solved analytically.

Indeed, consider for each index i the convex univariate problem

$$\begin{array}{ll} \min & c_i x_i + \lambda \frac{A_i}{x_i} \\ \text{s.t.} & x_i \in \mathbf{R}^+, \end{array} \tag{9}$$

where $\mathbf{R}^+$ denotes the set of non-negative real numbers. Problem (9) has as unique optimal solution $x_i(\lambda)$,

$$x_i(\lambda) = \sqrt{\frac{\lambda A_i}{c_i}}, i = 1, 2, \ldots, n.$$

Hence, the optimal solutions of

$$\begin{array}{ll} \min & c_i x_i + \lambda \frac{A_i}{x_i} \\ \text{s.t.} & l_i \leq x_i \leq u_i \\ & x_i \in \mathbf{Z} \end{array} \tag{10}$$

are given either by the closest feasible point to $x_i(\lambda)$, in case $x_i(\lambda)$ is outside the range $[l_i, u_i]$, or, else, the point(s) in the set $\{\lfloor x_i(\lambda) \rfloor, \lceil x_i(\lambda) \rceil\}$, yielding the lowest objective value.

For any positive integer k, the objective value in (10) at $x_i = k$ is not greater than at $x_i = k + 1$ iff $k(k+1) \geq \frac{\lambda A_i}{c_i}$. In other words, the set $\mathcal{S}_i(\lambda)$ of optimal solutions for (10) will be of the form

$$\mathcal{S}_i(\lambda) = \begin{cases} \{l_i\}, & \text{if } x_i(\lambda) \leq l_i \\ \{u_i\}, & \text{if } x_i(\lambda) \geq u_i \\ \{\lfloor x_i(\lambda) \rfloor\}, & \text{if } l_i < x_i(\lambda) < u_i \text{ and } (\lfloor x_i(\lambda) \rfloor + 1)\lfloor x_i(\lambda) \rfloor > \frac{\lambda A_i}{c_i} \\ \{\lceil x_i(\lambda) \rceil\}, & \text{if } l_i < x_i(\lambda) < u_i \text{ and } (\lfloor x_i(\lambda) \rfloor + 1)\lfloor x_i(\lambda) \rfloor < \frac{\lambda A_i}{c_i} \\ \{\lfloor x_i(\lambda) \rfloor, \lceil x_i(\lambda) \rceil\}, & \text{if } l_i < x_i(\lambda) < u_i \text{ and } (\lfloor x_i(\lambda) \rfloor + 1)\lfloor x_i(\lambda) \rfloor = \frac{\lambda A_i}{c_i}. \end{cases} \tag{11}$$

Finally, the set $\mathcal{S}(\lambda)$ of optimal solutions to (8) is the Cartesian product of the sets $\mathcal{S}_i(\lambda)$ above.

Hence, as soon as the parameter λ in (8) is provided, a full description of the whole set $\mathcal{S}(\lambda)$ of optimal solutions is at hand. However, in practice, it is not easy to provide a precise value for λ (or ν), [3].

Nevertheless, it is straightforward from the discussion above to obtain a characterization of the set of supported solutions $\mathcal{S}_{\infty,\infty} = \cup_{\lambda>0}\mathcal{S}(\lambda)$. Indeed, one has

Theorem 2.2 *Given $x = (x_1, \ldots, x_n) \in \mathbf{Z}^n$, with $l_i \leq x_i \leq u_i \, \forall i$, define*

$$\begin{aligned} \underline{x} &= \max_{i:\, l_i < x_i \leq u_i} \frac{x_i(x_i-1)c_i}{A_i} \\ \overline{x} &= \min_{i:\, l_i \leq x_i < u_i} \frac{x_i(x_i+1)c_i}{A_i}. \end{aligned}$$

Then, $x \in \mathcal{S}_{\infty,\infty}$ iff $\overline{x} \geq \underline{x}$.

Proof: By definition, x is supported iff there exists $\lambda > 0$ such that x solves the corresponding problem (8). By (11), for each given i, one has:

$$\{\lambda : x_i \in \mathcal{S}_i(\lambda)\} = \begin{cases} (0, \frac{l_i(l_i+1)c_i}{A_i}], & \text{if } x_i = l_i \\ \left[\frac{u_i(u_i-1)c_i}{A_i}, +\infty\right), & \text{if } x_i = u_i \\ \left[\frac{x_i(x_i-1)c_i}{A_i}, \frac{x_i(x_i+1)c_i}{A_i}\right], & \text{if } l_i < x_i < u_i. \end{cases} \tag{12}$$

Hence, $\bigcap_{i=1}^n \{\lambda : x_i \in \mathcal{S}_i(\lambda)\} \neq \emptyset$ iff $\underline{x}$, the highest lower bound in (12), does not exceed $\overline{x}$, the smallest upper bound in (12), and the result follows. □

Expressions (11) and (12) enable us to give an algorithm to describe the set $\mathcal{S}_{\infty,\infty}$:

Theorem 2.3 *The set $\mathcal{S}_{\infty,\infty}$ of supported solutions of $(P_{\infty,\infty})$ can be written as the union of at most N sets of the form $S_1 \times S_2 \times \ldots \times S_n$, where each S_j is either a singleton or consists of two consecutive integers, $S_j = \{k_j, k_j + 1\}$.*

In particular, $\mathcal{S}_{\infty,\infty}$ can be described in $O(Nn)$ time.

Proof: By definition, for each λ, the set $\mathcal{S}(\lambda)$ can be written, as the Cartesian product

$$\mathcal{S}(\lambda) = \mathcal{S}_1(\lambda) \times \ldots \times \mathcal{S}_n(\lambda),$$

where each $\mathcal{S}_i(\lambda)$ is, by (11), either a singleton or consists of two consecutive integers, and can thus be described in $O(n)$ time.

By (12), each set $\mathcal{S}_i(\lambda)$ is constant when λ varies in an interval whose endpoints are consecutive elements of the set C_i of critical values of λ,

$$C_i = \left\{ \frac{l_i(l_i+1)c_i}{A_i}, \frac{(l_i+1)(l_i+2)c_i}{A_i}, \ldots, \frac{(u_i-1)u_ic_i}{A_i} \right\}. \tag{13}$$

The total number of such critical values for λ has as upper bound $\sum_{i=1}^{n}(u_i-1-l_i+1) \leq \sum_{i=1}^{n}(N_i-1) = N-n$. Hence, $O(N)$ sets of the form $\mathcal{S}(\lambda)$ need to be constructed, yielding an overall time complexity of $O(Nn)$. □

Remark 2.4 The proof of Theorem 2.3 gives a procedure to describe $\mathcal{S}_{\infty,\infty}$: Construct, for each $i = 1, 2, \ldots, n$, the set C_i as defined in (13); for each $\lambda \in \bigcup_{i=1}^{n} C_i$, obtain, following (11), $\mathcal{S}(\lambda)$. Then, $\bigcup_{i=1}^{n}\bigcup_{\lambda\in C_i}\mathcal{S}(\lambda) = \mathcal{S}_{\infty,\infty}$. Moreover, without increasing complexity, the output can be obtained sorted by the first or second criterion. Indeed, given $0 < \lambda_1 < \lambda_2$, and, for $j = 1, 2$, an optimal solution x^j for (8), with $\lambda = \lambda_j$, it then follows that $\sum_{i=1}^{n}\frac{A_i}{x_i^2} \leq \sum_{i=1}^{n}\frac{A_i}{x_i^1}$ and $\sum_{i=1}^{n} c_i x_i^1 \leq \sum_{i=1}^{n} c_i x_i^2$. Hence, if the list of breakpoints in $\bigcup_{i=1}^{n} C_i$ is scanned in increasing order of λ, the corresponding set of optimal solutions obtained will appear sorted in nondecreasing and nonincreasing order for the first and second criterion respectively.

Since, by construction, each of the n lists C_i in (13) is already sorted, a data structure such as a heap will allow to sort $\bigcup_{i=1}^{n} C_i$ in $O(N \log n)$ time. Since the description of $\mathcal{S}_{\infty,\infty}$ given in Theorem 2.3 requires $O(Nn)$, it turns out that a description of $\mathcal{S}_{\infty,\infty}$ with the images sorted by one of the two criteria is obtained in $O(Nn) + O(N \log n) = O(Nn)$ time. □

Example 2.5 As a simple illustration, we have considered data on number of employees by area of industrial activity in the region of Andalucía, Spain, for the year 2000 as available in the database Tempus of the Spanish National Statistics Bureau, INE, [15]. The total number of employees is $N = 231,334$, grouped into $n = 14$ industrial activities (strata). The number N_i of individuals per stratum is given in the first column of Table 1.

The allocation problem is considered under the commonly used assumption that population quasivariances are independent of the strata, and then chosen to be fixed at 1. The lower and upper bounds imposed are the trivial ones: $l_i = 1$ and $u_i = N_i$, for all $i = 1, \ldots, n$.

We have considered two scenarios for the costs: in the first scenario, we have assumed all the costs c_i to be equal (and then fixed at 1), and in the latter, different costs c_i are randomly associated with the strata, as depicted in the second column of Table 1.

By using Theorem 2.3, it turns out that for both cost scenarios $\mathcal{S}_{\infty,\infty}$ is a set with cardinality $231,321$, obtained in 0.45 seconds in a AMD Athlon XP 2400+ with 1Gb RAM, running Debian/GNU Linux 3.0 with kernel 2.4.18. The fact that for these two different cost structures the number of supported solutions is the same is due to the arbitrary size of the strata.

Figures 1 and 2 depict in the output space the part of $\mathcal{S}_{\infty,\infty}$ in which the first and second objectives do not exceed $K^* = 10,000$ and $B^* = 0.001$ for the two above-mentioned cost scenarios.

N_i	c_i (2nd scenario)
7068	1
53856	1
20450	4
9812	7
10835	9
8331	8
4808	7
20714	8
30142	5
8864	3
9010	4
17519	9
20665	9
9260	9

Table 1: Data of Example 2.5

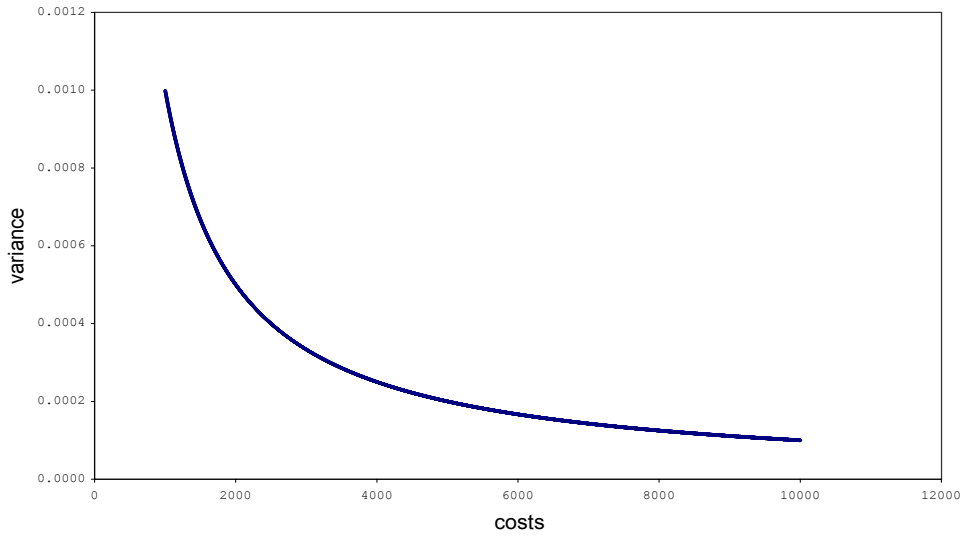

Figure 1: Output of $\mathcal{S}_{\infty,\infty}$ for equal costs

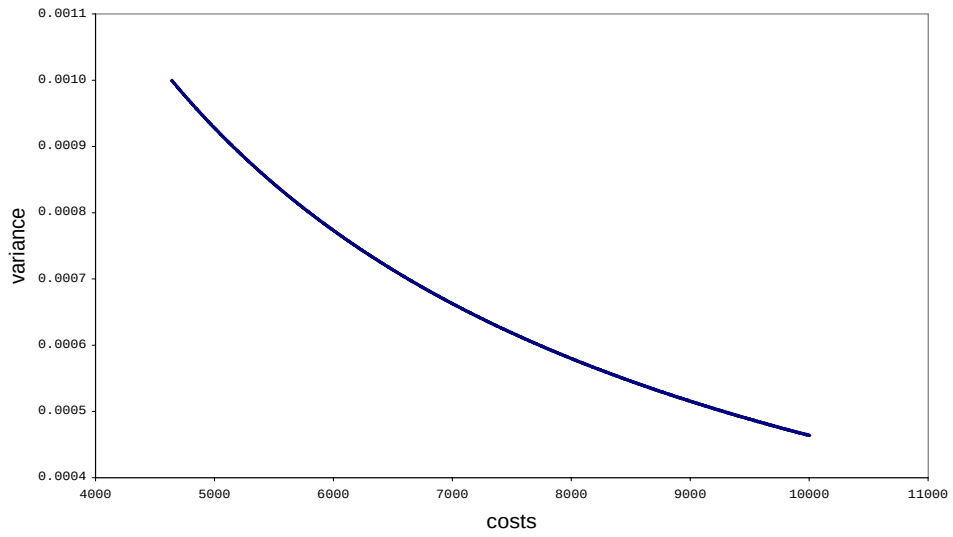

Figure 2: Output of $\mathcal{S}_{\infty,\infty}$ for non-equal costs

Remark 2.6 Although $\mathcal{S}_{\infty,\infty}$ can be described in $O(Nn)$ time, its cardinality can be exponential in n. Indeed, take for instance $l_1 = l_2 = \ldots = l_n$, $u_1 = u_2 = \ldots = u_n$, and $\frac{c_1}{A_1} = \frac{c_2}{A_2} = \ldots = \frac{c_n}{A_n}$.

In this particular case, the sets C_i in (13) coincide for all i. For the critical values of λ, each $\mathcal{S}_i(\lambda)$ has two elements, which means that each $\mathcal{S}(\lambda)$ has 2^n elements. □

3 Describing $\mathcal{P}_{K^*,B^*}$. The case of general costs

Although the set of supported solutions is contained in the set of Pareto-optimal solutions, it is usual in multiobjective combinatorial problems that such inclusion is strict, [16]. This is also the case of the problem under consideration, as shown in the following example.

Example 3.1 Consider a three-strata allocation problem, where the data are given in Table 2.

Representing all feasible points in the value space, we obtain the plot given in Figure 3.

N_i	$\sigma^2_{c,i}$	c_i	l_i	u_i
10	1	3	2	5
20	1	12	2	7
5	1	9	1	4

Table 2: Data to illustrate $\mathcal{S}_{\infty,\infty} \subsetneq \mathcal{P}_{\infty,\infty}$

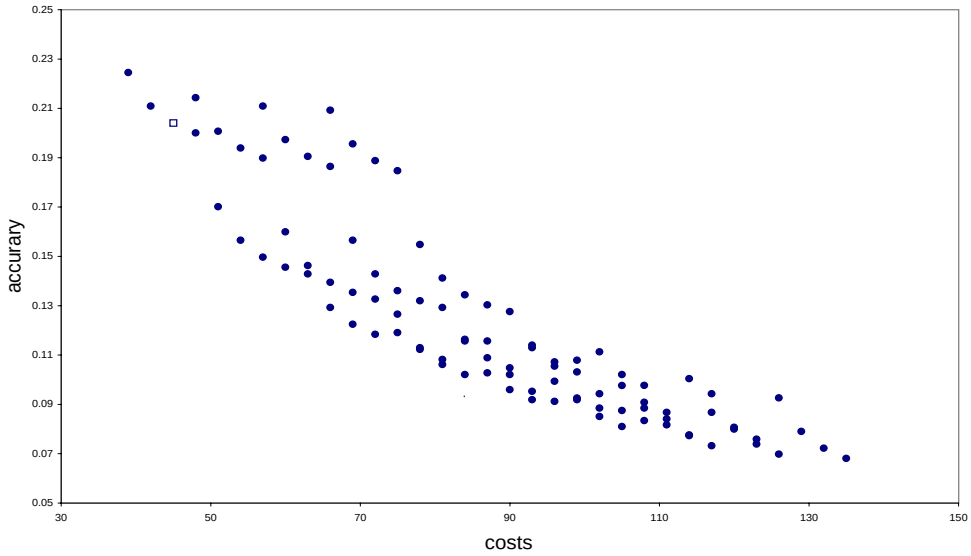

Figure 3: Value space for Example 3.1

Then, $x = (5, 2, 1)$ yields the point $(48, 0.2)$ in the value space, represented as an empty square in the figure, and $x \in \mathcal{P}_{\infty,\infty} \setminus \mathcal{S}_{\infty,\infty}$. Indeed, it is straightforward to check by complete enumeration (and evident from the picture) that $x \in \mathcal{P}_{\infty,\infty}$. However,

$$\begin{aligned} \underline{x} &= \frac{x_1(x_1-1)c_1}{A_1} = 735 \\ \overline{x} &= \min\left\{\frac{x_2(x_2+1)c_2}{A_2}, \frac{x_3(x_3+1)c_3}{A_3}\right\} = 220.5. \end{aligned}$$

Hence, $\overline{x} < \underline{x}$, thus, by Theorem 2.2, $x \notin \mathcal{S}_{\infty,\infty}$, as asserted. □

In this section we describe a procedure to obtain the set $\mathcal{P}_{K^*,B^*}$, under no assumptions on the costs excepting their integrality. To do that, we first consider a parametric class of scalar problems,

$$\begin{array}{ll} \min & \sum_{i=1}^n \frac{A_i}{x_i} \\ \text{s.t.} & \sum_{i=1}^n c_i x_i \le K \\ & l_i \le x_i \le u_i,\ i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n, \end{array} \tag{14}$$

which amounts to finding the sampling allocation of minimal variance whose cost does not exceed a threshold value K.

We discuss in Section 3.1 how to solve (14) for K *fixed*, and devote Section 3.2 to show how the resolution of (14) for different right-hand sides yields a description of $\mathcal{P}_{K^*,B^*}$. Moreover, we show, following ideas of [17, 18], how the knowledge of $\mathcal{S}_{\infty,\infty}$ can be exploited to alleviate the computational burden needed for constructing $\mathcal{P}_{K^*,B^*}$.

3.1 Solving the constrained problems

Since the pioneering work of Neyman [12] in 1934, different solution approaches, both exact or heuristics, have been proposed in the last 70 years for (14). The most popular method is a heuristic which yields a closed-formula: if both integrality and box constraints are dropped from (14), we come up with a separable strictly convex linearly-constrained problem, the optimal solution of which is given by x^*,

$$x^* = \frac{K}{\sum_{i=1}^n \sqrt{c_i A_i}} \left(\sqrt{\frac{A_1}{c_1}}, \ldots, \sqrt{\frac{A_n}{c_n}} \right),$$

see [5].

Although more efficient procedures could be used, see e.g. [9], an integer solution satisfying the cost constraint is found by rounding down fractional components. This heuristic solution is considered in textbooks as satisfactory, since, being usually the objective fairly flat around the optimum, [5], its gap is expected to be small, and box constraints will be automatically satisfied unless very small or very large strata have respectively high or small variances.

In our opinion, this argument is misleading. First, there is no guarantee that the gap will be small enough. Hence, optimal instead of suboptimal solutions should be provided, as soon as they can be obtained with reasonable computational effort.

Moreover, if the analyst does not care too much about the precise value of the variance (or the cost), instead of just using an allocation heuristic, a full parametric analysis, provided by the set of outcomes of $\mathcal{P}_{K^*,B^*}$, as described below, would be of much more use in order to find the right trade-off between variances and costs.

Several branch-and-bound methods have been suggested in the literature to solve exactly (14). These methods differ in the way they obtain the lower bounds and the way feasible (suboptimal) solutions are generated.

For instance, in [2], the continuous relaxation,

$$\begin{array}{rcll} z_1 & = & \min & \sum_{i=1}^n \frac{A_i}{x_i} \\ & & \text{s.t.} & \sum_{i=1}^n c_i x_i \leq K \\ & & & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n, \end{array} \tag{15}$$

is proposed. This is a convex problem with box and one linear constraints, and can be solved via its Lagrangean dual

$$\begin{array}{lll} \max_{\theta \geq 0} & \min & \sum_{i=1}^n \frac{A_i}{x_i} + \theta(\sum_{i=1}^n c_i x_i - K) \\ & \text{s.t.} & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n. \end{array} \tag{16}$$

A second bounding procedure directly follows from the reformulation as knapsack problems of convex separable integer problems, [8]. Introducing, for each $i = 1, 2, \ldots, n$ and each $j = 1, 2, \ldots, u_i - l_i$ the Boolean variables y_{ij} and coefficients $\eta_{ij} = A_i \left(\frac{1}{l_i + j} - \frac{1}{l_i + j - 1} \right)$,

(14) can be reformulated as the knapsack problem

$$\begin{array}{ll} \min & \sum_{i=1}^{n}\left(\frac{A_i}{l_i}+\sum_{j=1}^{u_i-l_i}\eta_{ij}y_{ij}\right) \\ \text{s.t.} & \sum_{i=1}^{n} c_i \sum_{j=1}^{u_i-l_i} y_{ij} \leq K \\ & y_{ij}\in\{0,1\},\ j=1,2,\ldots,u_i-l_i,\ i=1,2,\ldots,n. \end{array} \tag{17}$$

Let z_2 denote the optimal value of the linear programming (hereafter LP) relaxation of this knapsack problem,

$$\begin{array}{rcll} z_2 & = & \min & \sum_{i=1}^{n}\left(\frac{A_i}{l_i}+\sum_{j=1}^{u_i-l_i}\eta_{ij}y_{ij}\right) \\ & & \text{s.t.} & \sum_{i=1}^{n} c_i \sum_{j=1}^{u_i-l_i} y_{ij} \leq K \\ & & & 0\leq y_{ij}\leq 1,\ j=1,2,\ldots,u_i-l_i,\ i=1,\ldots,n. \end{array} \tag{18}$$

In the following result we show that this relaxation is at least as good as the continuous relaxation of the original problem, (15).

Theorem 3.2 *Let z denote the optimal value of (14), and z_1, z_2 as defined respectively in (15), (18). Then,*

$$z_1 \leq z_2 \leq z.$$

Proof: It is enough to show that $z_1 \leq z_2$. The value z_2 can be obtained by solving the Lagrangean dual of (18),

$$\begin{array}{lll} \max_{\theta\geq 0} & \min & \sum_{i=1}^{n}\left(\frac{A_i}{l_i}+\sum_{j=1}^{u_i-l_i}\eta_{ij}y_{ij}\right)+\theta(\sum_{i=1}^{n} c_i \sum_{j=1}^{u_i-l_i} y_{ij}-K) \\ & \text{s.t.} & 0\leq y_{ij}\leq 1,\ j=1,2,\ldots,u_i-l_i,\ i=1,\ldots,n. \end{array} \tag{19}$$

Problem (19) yields the same optimal value as its $0-1$ version, which is simply a $0-1$ reformulation of the problem

$$\begin{array}{lll} \max_{\theta\geq 0} & \min & \sum_{i=1}^{n}\frac{A_i}{x_i}+\theta(\sum_{i=1}^{n} c_i x_i-K) \\ & \text{s.t.} & l_i\leq x_i\leq u_i,\ i=1,2,\ldots,n, \\ & & x_i\in\mathbf{Z},\ i=1,2,\ldots,n. \end{array} \tag{20}$$

Moreover, it is an upper bound of its continuous relaxation, namely (16) and the result follows. □

Both inequalities above can be strict, as shown in the following example.

Example 3.3 Consider the data given in Table 2, with $K=46$. Then (14) can be solved by complete enumeration, yielding $(4,2,1)$ as optimal solution and $z=0.2041$ as optimal value. On the other hand, (15) has as optimal solution $x^1=(\frac{222}{90},\frac{222}{90},1)$, with optimal value $z_1=0.1859$, whereas (18) has $x^2=(3,\frac{7}{3},1)$ as optimal solution, and $z_2=0.1927$ as optimal value. Hence, $z_1<z_2<z$, as asserted. □

As pointed out in [8], (17) is a knapsack problem which can be solved rather efficiently by a number of procedures, see [11]. In particular, we have implemented the depth-search branch-and-bound algorithm of Horowitz and Sanhi, in which bounds are obtained by solving (18), the variable to branch is the one with the lowest cost per unit weight $\frac{\eta_{ij}}{c_i}$ among those which are not yet fixed, (observe that we are solving a minimization version of the knapsack problem), and the search continues from the node with the branching variable fixed to one. As already pointed out in [8], for a given optimal solution y^* of (17) and a stratum i, the convexity of (14) implies that if $y^*_{i\hat{j}} = 0$, then $y^*_{ij} = 0$ for all $j \geq \hat{j}$. We use this property of the knapsack formulation explicitly in the branching scheme to reduce the tree search.

Remark 3.4 The discussion in this section relies upon the fact that one of the two objectives, namely the cost, is put as constraint and the other one remains as objective. If instead it is the cost the one remaining as objective and the sampling variance is put as constraint, a similar analysis can be done, since problems of the form

$$\begin{array}{ll} \min & \sum_{i=1}^n c_i x_i \\ \text{s.t.} & \sum_{i=1}^n \frac{A_i}{x_i} \leq B \\ & l_i \leq x_i \leq u_i,\ i = 1, 2, \ldots, n \\ & x_i \in \mathbf{Z},\ i = 1, 2, \ldots, n, \end{array} \tag{21}$$

can also be reformulated as $0-1$ knapsack problems and solved via a branch-and-bound technique in a similar way. □

3.2 Enumerating the set $\mathcal{P}_{K^*,B^*}$

As detailed below, a full description of $\mathcal{P}_{K^*,B^*}$ can be obtained by solving a series of problems of type (14) and (21). Moreover, significant savings will be obtained when the information supplied by $\mathcal{S}_{\infty,\infty}$ is used.

For any x, feasible for (P_{K^*,B^*}), the constraints imply that

$$\underline{K} \leq \sum_{i=1}^n c_i x_i \leq \overline{K},$$

with

$$\begin{aligned} \overline{K} &= \min\left\{\sum_{i=1}^n c_i u_i, K^*\right\} \\ \underline{K} &= \min\left\{\sum_{i=1}^n c_i x_i : \sum_{i=1}^n \frac{A_i}{x_i} \leq B^*,\ x \in \mathbf{Z}^n_+, l_i \leq x_i \leq u_i\ \forall i\right\}. \end{aligned}$$

Observe that, since (P_{K^*,B^*}) is, by assumption, feasible, $\underline{K}$ is well defined.

Theorem 3.5 *One has*

1. *For any* $K \in \{\underline{K}, \underline{K}+1, \ldots, \overline{K}-1, \overline{K}\}$, *(14) is feasible. Moreover, any optimal solution for (14) is feasible for* (P_{K^*,B^*}).

2. *For any* $K \in \{\underline{K}, \underline{K}+1, \ldots, \overline{K}-1, \overline{K}\}$, *if there exists only one optimal solution to (14), say* x^K, *then* x^K *is Pareto-optimal for* (P_{K^*,B^*}). *Otherwise, any optimal solution to (21) with right-hand side coefficient* $\sum_{i=1}^n \frac{A_i}{x_i^K}$ *is Pareto-optimal for* (P_{K^*,B^*}).

3. *Any Pareto-optimal solution for* (P_{K^*,B^*}) *solves (14) for some integer* $K \in \{\underline{K}, \underline{K}+1, \ldots, \overline{K}-1, \overline{K}\}$.

Proof: Let $K \in \{\underline{K}, \underline{K}+1, \ldots, \overline{K}-1, \overline{K}\}$. Any x^*, optimal to

$$\min\left\{\sum_{i=1}^n c_i x_i : \sum_{i=1}^n \frac{A_i}{x_i} \leq B^*,\, x \in \mathbf{Z}_+^n, l_i \leq x_i \leq u_i \,\forall i\right\},$$

is feasible for (14) for right-hand side $\underline{K}$, and thus also for right-hand side K.

In particular, any x optimal for (14) with right-hand side K is feasible for (14), thus

$$\sum_{i=1}^n c_i x_i \leq K \leq \overline{K} \leq K^*,$$

and, since x^* is also feasible,

$$\sum_{i=1}^n \frac{A_i}{x_i} \leq \sum_{i=1}^n \frac{A_i}{x_i^*} \leq B^*,$$

showing that x is feasible for (P_{K^*,B^*}). Hence, part 1 holds. Part 2 follows trivially by the definition of the constrained problems. For part 3, observe that, by definition of Pareto-optimality, any $x \in \mathcal{P}_{K^*,B^*}$ solves (14) for right-hand side $K = \sum_{i=1}^n c_i x_i$. Since by assumption, the coefficients c_i are integer,

$$\sum_{i=1}^n c_i x_i \in \{\underline{K}, \underline{K}+1, \ldots, \overline{K}-1, \overline{K}\},$$

and the desired result follows. □

This yields the following procedure to describe $\mathcal{P}_{K^*,B^*}$. Initially, we calculate all the supported solutions, which will be the first Pareto-optimal solutions at hand. Then, we find the possible Pareto-optimal solutions associated with each value K in $\{\underline{K}, \underline{K}+1, \ldots, \overline{K}-1, \overline{K}\}$. We go through this list in decreasing order. Given K, we first check whether it corresponds to the sampling cost of any supported solution. If that is the case, by Theorem 2.1, the corresponding Pareto-optimal solution is supported and already found in the initialization. Otherwise, we solve (14). If there exists exactly one optimal solution to this problem, we add it to $\mathcal{P}_{K^*,B^*}$. Otherwise, we solve (21) with right-hand side coefficient $\sum_{i=1}^n \frac{A_i}{x_i^K}$. All the optimal solutions to this problem are Pareto-optimal.

Algorithm: Describing $\mathcal{P}_{K^*,B^*}$

Step 0. Set $\mathcal{P} := \{x \in \mathcal{S}_{\infty,\infty} : \sum_{i=1}^n c_i x_i \leq K^*, \sum_{i=1}^n \frac{A_i}{x_i} \leq B^*\}$ and set $K := \overline{K}$.

Step 1. If $K \in \left\{\sum_{i=1}^n c_i x_i : x \in \mathcal{S}_{\infty,\infty}, \sum_{i=1}^n \frac{A_i}{x_i} \leq B^*\right\}$, set $K := K - 1$ and GoTo Step 4.

Step 2. Find *one* optimal solution x^K to (14) with right-hand side coefficient K.

Step 3. If x^K is the unique optimal solution, then set $\mathcal{P} = \mathcal{P} \cup \{x^K\}$. Else, find the set $\mathcal{O}^K$ of *all* optimal solutions of (21) with right-hand side coefficient $\sum_{i=1}^n \frac{A_i}{x_i^K}$, set $\mathcal{P} := \mathcal{P} \cup \mathcal{O}^K$ and $K := \sum_{i=1}^n c_i x_i^K - 1$.

Step 4. If $K \geq \underline{K}$, then GoTo Step 2. Else set $\mathcal{P}_{K^*,B^*} := \mathcal{P}$ and STOP.

Some technical implementation issues follow. First, by Remark 2.4, one does not need to fully construct in Step 0 the set $S_{\infty,\infty}$: if, at some breakpoint λ an optimal solution x for (8) is obtained by (11) with $\sum_{i=1}^n c_i x_i > K^*$ (respectively $\sum_{i=1}^n \frac{A_i}{x_i} > B^*$) then no breakpoint $\lambda' > \lambda$ (respectively no $\lambda' < \lambda$) can yield solutions feasible for (P_{K^*,B^*}).

For those values of K for which the associated problem must be solved in Step 2, the computational burden can be alleviated by doing some simple preprocessing at Step 0 as well as by using some extra information from the cases already studied of this parametric problem. Indeed, (14) is reformulated in (17), as a knapsack problem, to be solved by a branch-and-bound algorithm, in which the LP-relaxation (18) is used as a bounding scheme. The first step performed to solve such LP-relaxation of the root node is to calculate all ratios, $\frac{\eta_{ij}}{c_i}$, and sort them in increasing order. This sorted list of ratios is independent of the parameter K, and can thus be already calculated in Step 0. Moreover, before solving (14), one may already have feasible solutions (obtained from $\mathcal{S}_{\infty,\infty}$) and upper bounds (the optimal values previously obtained in Step 2 for higher values of K).

Step 3 is the hardest part. First, we must check whether the x^K obtained in Step 2 is the unique optimal solution to (14). This can be tested in Step 2, by pruning in the branch-and-bound tree only those nodes whose lower bound is strictly worse than the best incumbent. Moreover, we have to keep all optimal solutions of (14) with minimal variance.

If instead of finding all optimal solutions to (21) with right-hand side coefficient $\sum_{i=1}^n \frac{A_i}{x_i^K}$, we just take one of its optimal solutions, the algorithm above will finally stop with a set $\mathcal{P} \subseteq \mathcal{P}_{K^*,B^*}$ describing the Pareto outcomes.

4 Describing $\mathcal{P}_{K^*,B^*}$. The case of equal costs

In this section we address the important particular case of (P_{K^*,B^*}) where the costs c_i are independent of the stratum, showing that, contrary to the general case discussed in Section 3, all Pareto optima are supported. By Theorem 2.3, this yields a closed characterization of $\mathcal{P}_{K^*,B^*}$ as well as an $O(Nn)$-time procedure.

We have

Theorem 4.1 *If $c_1 = c_2 = \ldots = c_n$, then $\mathcal{S}_{\infty,\infty} = \mathcal{P}_{\infty,\infty}$. Moreover,*

$$\mathcal{P}_{K^*,B^*} = \left\{ x \in \mathcal{S}_{\infty,\infty} : \sum_{i=1}^{n} c_i x_i \leq K^*, \sum_{i=1}^{n} \frac{A_i}{x_i} \leq B^* \right\}. \tag{22}$$

Proof: Any supported solution is Pareto-optimal, so we only need to show the converse. Given $x^* \in \mathcal{P}_{\infty,\infty}$, x^* solves (14) for $K := \sum_{i=1}^{n} c_i x_i^*$. The result can then be derived from Theorem 4.1.1. and Section 4.7 of [9]; however, for the sake of self-containedness, a complete proof is derived here.

Define, for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, u_i - l_i$, the Boolean variables y_{ij}^* as

$$y_{ij}^* = \begin{cases} 1, & \text{if } j \leq x_i^* - l_i \\ 0, & \text{else.} \end{cases}$$

Then, y^* is optimal for (17), which, since $c_1 = \ldots = c_n$, is equivalent to its continuous relaxation (18). In particular, $\theta^* \geq 0$ exists such that (θ^*, y^*) is a saddle-point pair for (19) and for its $0-1$ reformulation. Moreover, $\theta^* > 0$, since we have strong duality and the optimal value of (14) is strictly positive.

By construction of y^*, we have that (θ^*, x^*) is a saddle-point pair for (20), which shows in particular that

$$x^* \in \mathcal{S}(\theta^*) \subset \mathcal{S}_{\infty,\infty}.$$

Then, (22) follows from (7). □

Under the assumption of equal costs, the set $\mathcal{S}_{K^*,B^*}$ of supported solutions of (P_{K^*,B^*}) can also be obtained directly from the set $\mathcal{S}_{\infty,\infty} = \mathcal{P}_{\infty,\infty}$:

Corollary 4.2 *If $c_1 = c_2 = \ldots = c_n$, then*

$$\mathcal{S}_{K^*,B^*} = \left\{ x \in \mathcal{S}_{\infty,\infty} : \sum_{i=1}^{n} c_i x_i \leq K^*, \sum_{i=1}^{n} \frac{A_i}{x_i} \leq B^* \right\}. \tag{23}$$

Proof: We only need to show the inclusion,

$$\mathcal{S}_{K^*,B^*} \subseteq \left\{ x \in \mathcal{S}_{\infty,\infty} : \sum_{i=1}^{n} c_i x_i \leq K^*, \sum_{i=1}^{n} \frac{A_i}{x_i} \leq B^* \right\}.$$

Given $x \in \mathcal{S}_{K^*,B^*}$, x satisfies by construction the constraints, and is Pareto-optimal for (P_{K^*,B^*}). Hence $x \in \mathcal{P}_{\infty,\infty} = \mathcal{S}_{\infty,\infty}$, showing the result. □

5 Numerical results

In this section we illustrate the two-phase algorithm proposed in Section 3.2 to describe the set of Pareto-optimal solutions for the database presented in Example 2.5. We only consider here the second scenario of costs, since for the first one the set of Pareto-optimal solutions is equal to the set of supported solutions which has been already calculated in Example 2.5. We have imposed budget and accuracy constraints to the sample allocation problem (6). In particular, we have enumerated the set $\mathcal{P}_{K^*,B^*}$, where $K^* = 10,000$ and $B^* = 0.001$. In phase 1, 1394 supported solutions were found, which took 0.05 seconds. In phase 2, we found a total number of 3967 non-supported solutions and the computation time was equal to 741.33 seconds. The number of Pareto-optimal solutions is equal to 5361, and the two-phase algorithm took 741.38 seconds. To illustrate the savings reached by applying the knowledge of $\mathcal{S}_{\infty,\infty}$, we have calculated the same Pareto-optimal solutions without performing phase 1 of the algorithm. This took a total of 905.81 seconds, which incurs in a 22.18% increase in the computation time with respect to our two-phase algorithm.

We are also interested in extracting some information about the optimization of the linear knapsack problem (17). We may recall that this problem has been solved using a branch-and-bound algorithm. In Figure 4, we have plotted the error bound of the integer solution of (17) available at the root node of the branch-and-bound tree. We observe that this error bound decreases when the right-hand side of the cost constraint, K, increases. Next to the error bound, we have also plotted the computation time per knapsack problem solved, see Figure 5. From this plot we can see that the computation time per knapsack problem solved tends to increase with the parameter K, probably due to the fact that, when the parameter K increases, the number of variables with positive value in the LP-relaxation of (17) also increases and the minimum number of nodes which we should inspect to prove optimality in the branch-and-bound scheme of Horowitz and Sanhi increases. Finally, we have observed that the optimal solution of (17) is found in a very early stage of the branch-and-bound tree and most of the nodes are pruned because they are integer or they are not promising.

6 Further research

The results obtained in this paper directly apply also to allocation problems for other estimators or other estimation settings, see Carrizosa and Romero Morales [4].

A very challenging extension of the sample allocation model we have dealt with is considering Y as an ℓ-dimensional variable, $Y = (Y^1, \ldots, Y^\ell)$, thus the ℓ-dimensional parameter $\overline{Y} = (\overline{Y^1}, \overline{Y^2}, \ldots, \overline{Y^\ell})$ must be estimated from a single stratified random sample. The technique developed in this article is still applicable when, for instance, all ℓ variables are $0-1$ or under the very common assumption that the population quasivariance $\sigma^2_{c,Y^j,i}$ of the variables Y^j are independent of the strata, see Carrizosa and Romero Morales [4]. The case where none of these reductions can be made will be addressed in a forthcoming

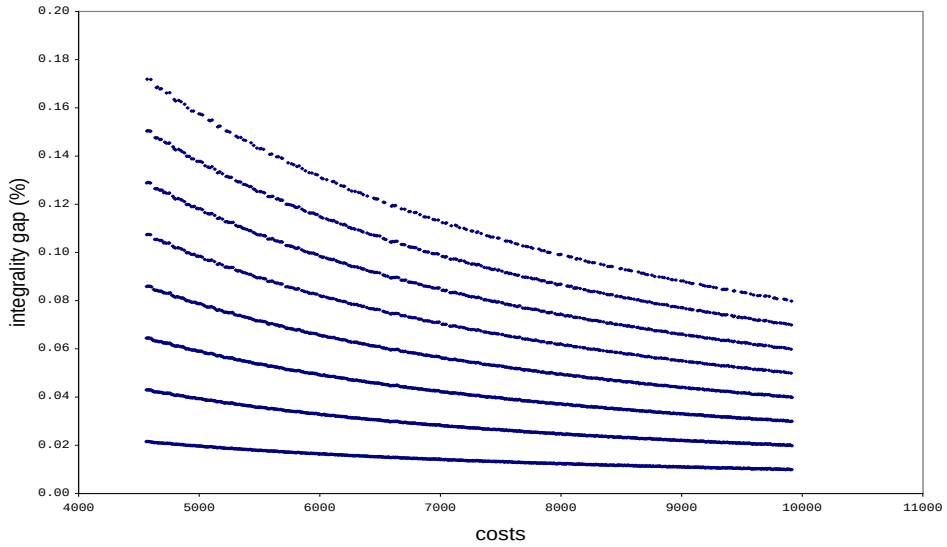

Figure 4: Plot of integrality gap for knapsack problem (17)

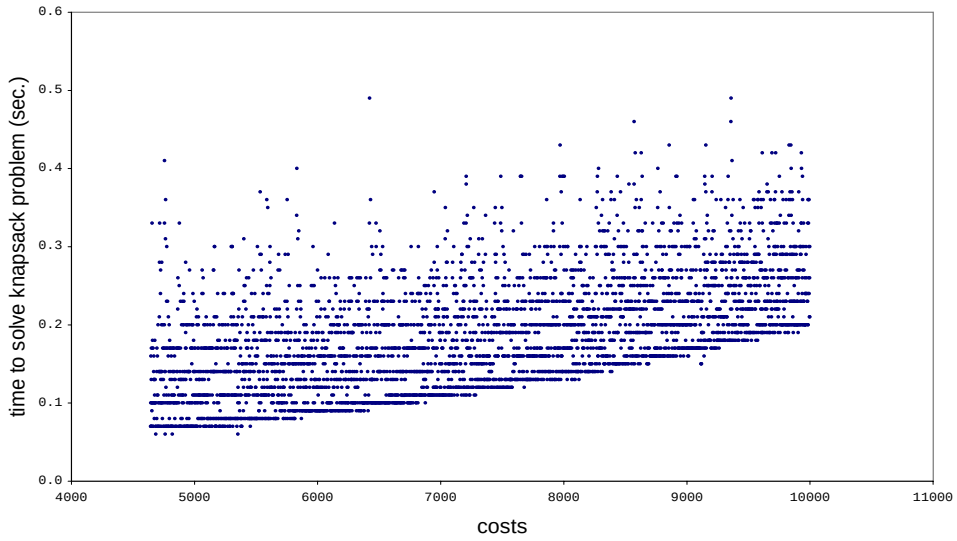

Figure 5: Plot of time to find non-supported solutions when $K^* = 10,000$ and $B^* = 0.001$

paper.

Acknowledgements

This research is supported by grant BFM2002-04525-C02-02 of Ministerio de Ciencia y Tecnología, Spain.

References

[1] Om P. Aggarwal. Bayes and minimax procedures in sampling from finite and infinite populations. *The Annals of Mathematical Statistics*, 30:206–218, 1959.

[2] K.M. Bretthauer, A. Ross, and B. Shetty. Nonlinear integer programming for optimal allocation in stratified sampling. *European Journal of Operational Research*, 116:667–680, 1999.

[3] E. Carrizosa, E. Conde, F.R. Fernández, and J. Puerto. An axiomatic approach to the centdian criterion. *Location Science*, 3:165–171, 1994.

[4] E. Carrizosa and D. Romero Morales. A biobjective method for sample allocation in stratified sampling. *METEOR Research Memorandum*, RM/03/019, 2003.

[5] W.G. Cochran. *Sampling techniques*. Wiley, New York, 3rd edition, 1977.

[6] M. Eben-Chaime. Parametric solution for linear bicriteria knapsack models. *Management Science*, 42:1565–1575, 1996.

[7] M. Ehrgott and X. Gandibleux. An annotated bibliography of multiobjective combinatorial optimization. *OR Spektrum*, 22:425–460, 2000.

[8] D. S. Hochbaum. A nonlinear knapsack problem. *Operations Research Letters*, 17:103–110, 1995.

[9] T. Ibaraki and N. Katoh. *Resource allocation problems. Algorithmic approaches*. MIT Press, Cambrigde, Massachusetts, 1988.

[10] L. Kish. Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, 139:80–95, 1976.

[11] S. Martello and P. Toth. *Knapsack problems, algorithms and computer implementations*. John Wiley & Sons, New York, 1990.

[12] J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606, 1934.

[13] C.J. Skinner, D.J. Holmes, and D. Holt. Multiple frame sampling for multivariate stratification. *ISI Review*, 62:333–347, 1994.

[14] R.E. Steuer. *Multiple criteria optimization: theory, computation, and application.* Wiley, New York, 1986.

[15] Tempus database. *Instituto Nacional de Estadística*. `http://www.ine.es/`.

[16] E.L. Ulungu and J. Teghem. Multi-objective combinatorial optimization problems: a survey. *Journal of Multi-Criteria Decision Analysis*, 3:83–104, 1994.

[17] E.L. Ulungu and J. Teghem. The two-phases method: An efficient procedure to solve biobjective combinatorial optimization problems. *Foundations of Computing and Decision Sciences*, 20:149–165, 1995.

[18] M. Visée, J. Teghem, M. Pirlot, and E.L. Ulungu. Two-phases method and branch and bound procedures to solve the bi-objective knapsack problem. *Journal of Global Optimization*, 12:139–155, 1998.