Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Integer programming formulations applied to optimal allocation in stratified sampling

by José André de Moura Brito, Pedro Luis do Nascimento Silva, Gustavo Silva Semaan and Nelson Maculan

Release date: December 17, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 427-442
Statistics Canada, Catalogue No. 12-001-X

Integer programming formulations applied to optimal allocation in stratified sampling

José André de Moura Brito, Pedro Luis do Nascimento Silva, Gustavo Silva Semaan and Nelson Maculan[1]

Abstract

The problem of optimal allocation of samples in surveys using a stratified sampling plan was first discussed by Neyman in 1934. Since then, many researchers have studied the problem of the sample allocation in multivariate surveys and several methods have been proposed. Basically, these methods are divided into two classes: The first class comprises methods that seek an allocation which minimizes survey costs while keeping the coefficients of variation of estimators of totals below specified thresholds for all survey variables of interest. The second aims to minimize a weighted average of the relative variances of the estimators of totals given a maximum overall sample size or a maximum cost. This paper proposes a new optimization approach for the sample allocation problem in multivariate surveys. This approach is based on a binary integer programming formulation. Several numerical experiments showed that the proposed approach provides efficient solutions to this problem, which improve upon a 'textbook algorithm' and can be more efficient than the algorithm by Bethel (1985, 1989).

Key Words: Stratification; Allocation; Integer programming; Multivariate survey.

1 Introduction

A large part of the statistics produced by official statistics agencies in many countries come from sample surveys. Such surveys have a well-defined survey population to be covered, including the geographic location and other eligibility criteria, use appropriate frames to guide the sample selection, and apply some well-specified sample selection procedures. The use of 'standard' probability sampling procedures enables producing estimates for the target population parameters with controlled precision while having data from typically small samples of the populations, at a fraction of the cost of corresponding censuses.

When designing the sampling strategy, the survey planner often seeks to optimize precision for the most important survey estimates given an available survey budget. Stratification is an important tool that enables exploring prior auxiliary information available for all the population units by forming groups of homogeneous units, and then sampling independently from within such groups. Thus stratification is very frequently used in a wide range of sample surveys.

Here we focus on element sampling designs (Särndal, Swensson and Wretman 1992) where the frame consists of one record per population unit, and besides identification and location information, some auxiliary information is also available for each population unit. Stratified sampling involves dividing the N units in a population U into H homogeneous groups, called strata. These groups are formed considering one (or more) stratification variable(s), and such that variance within groups is small (the stratum formation problem).

1. José André de Moura Brito and Pedro Luis do Nascimento Silva, Escola Nacional de Ciências Estatísticas (ENCE/IBGE), R.André Cavalcanti, 106, sala 403, Centro, Rio de Janeiro/RJ. E-mail: jambrito@gmail.com and pedronsilva@gmail.com; Gustavo Silva Semaan, Instituto do Noroeste Fluminense de Educação Superior, Universidade Federal Fluminense - INFES/UFF, Av. João Jasbick, s/n, Bairro Aeroporto - Santo Antônio de Pádua - RJ - CEP 28470-000. E-mail: gustavosemaan@gmail.com; Nelson Maculan, Universidade Federal do Rio de Janeiro (COPPE/UFRJ), Endereço: Av. Horácio Macedo, 2030 - CT, Bloco H, sl. 319 - Cidade Universitária, Ilha do Fundão - Rio de Janeiro, RJ - CEP 21941-914. E-mail: nelson.maculan@gmail.com.

Given a sample size n, once the strata are defined the next problem consists of specifying how many sample units should be selected in each stratum such that the variance of a specified estimator is minimized (the optimal sample allocation problem). When interest is restricted to estimating the population total (or mean) for a single survey variable, the well-known Neyman allocation (see e.g., Cochran 1977) may be used to decide on the sample allocation. Although surveys which have a single target variable are rare, Neyman's simple allocation formula may still be useful because the allocation which is optimal for a target variable may still be reasonable for other survey variables which are positively correlated with the one used to drive the optimal allocation.

When a survey must produce estimates with specified levels of precision for a number of survey variables, and these variables are not strongly correlated, a method of sample allocation that enables producing estimates with the required precision for all the survey variables is needed. In this case, we have a problem of multivariate optimal sample allocation.

According to the literature, in such cases the allocation of the overall sample size n to the strata may seek one of the following goals:

(i) the total variable survey cost C is minimized, subject to having Coefficients of Variation (CVs) for the estimates of totals of the m survey variables below specified thresholds; or
(ii) a weighted sum of variances (or relative variances) of the estimates of totals for the m survey variables is minimized.

Note that the CV is simply the square root of the relative variance.

This paper presents a new approach based on developing and applying two binary integer programming formulations that satisfy each of these two goals, while ensuring that the resulting allocation provides the global optimum. The paper is divided as follows. Section 2 reviews some key stratified sampling concepts and definitions. Section 3 describes the new approach proposed here. Section 4 provides results for a subset of numerical experiments carried out to test the proposed approach using selected population datasets. Section 5 gives some final remarks and concludes the paper. Appendix A provides information about three populations used in the numerical experiments presented in Section 4.

2 Stratified sampling and the optimal allocation problem

In stratified sampling (Cochran 1977; Lohr 2010) a population U formed by N units is divided into H strata $U_1, U_2, \ldots, U_H$ having $N_1, N_2, \ldots, N_H$ units respectively. These strata do not overlap (2.1) and together form the entire population (2.2) such that:

$$U_h \bigcap U_k = \varnothing, \quad h \neq k \tag{2.1}$$

$$\bigcup_{h=1}^{H} U_h = U \tag{2.2}$$

$$N_1 + N_2 + \ldots + N_H = \sum_{h=1}^{H} N_h = N. \tag{2.3}$$

Once the strata are defined, and given an overall sample size n, an independent sample of size n_h is selected from the N_h units in stratum U_h $(h = 1, \ldots, H)$ such that $n_{\min} \leq n_h \leq N_h \; \forall h$, where $n_{\min}$ is the smallest possible sample size in any stratum, and $n_1 + n_2 + \ldots + n_H = \sum_{h=1}^{H} n_h = n$.

A minimum sample size per stratum of $n_{\min} = 2$ is considered here, but this value may be changed as needed to accommodate specific survey requirements. A minimum sample size of one per stratum is not recommended because this might lead to solutions that require using approximate methods for variance estimation whenever the allocated sample sizes reach this minimum. In practice, it may even be wise to use $n_{\min}$ larger than 2, because of nonresponse or for other practical reasons.

Assuming full response, the data are collected for all units in the selected sample and used to produce estimates (of totals, say) for a set of m survey variables. Let $y_1, y_2, \ldots, y_m$ denote the survey variables. The variance of variable y_j in stratum h is defined as:

$$S_{hj}^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} \left(y_{ij} - \bar{Y}_{hj} \right)^2 \tag{2.4}$$

where y_{ij} is the value of y_j for the i^{th} population unit, and $\bar{Y}_{hj}$ is the population mean for y_j in stratum h, given by

$$\bar{Y}_{hj} = \frac{1}{N_h} \sum_{i \in U_h} y_{ij} = Y_{hj} / N_h \tag{2.5}$$

for $h = 1, \ldots, H$ and $j = 1, \ldots, m$. The population total Y_j for the j^{th} survey variable is $Y_j = \sum_{h=1}^{H} \sum_{i \in U_h} y_{ij} = \sum_{h=1}^{H} Y_{hj}$.

Under stratified simple random sampling (STSRS), the variance of the Horvitz-Thompson (HT) estimator t_j of the total for the j^{th} survey variable (Cochran 1977) is given by:

$$V\left(t_j\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hj}^2 \tag{2.6}$$

where $t_j = \sum_{h=1}^{H} N_h / n_h \sum_{i \in s_h} y_{ij} = \sum_{h=1}^{H} N_h \bar{y}_{hj}$, $s_h \subset U_h$ is the set of labels of the n_h units sampled in stratum h, and $\bar{y}_{hj}$ is the sample mean in stratum h.

Because the values of N_h and S_{hj}^2 are fixed after the strata have been defined, the variance of the HT estimator t_j of the total for the j^{th} survey variable in (2.6) depends only on the sample sizes n_h allocated to the strata. This allocation is important, because it is what enables the survey designer to control the precision of the survey estimates.

In general, when performing the allocation, the survey planner seeks a balance between achieving the desired precision for each of the survey variables of interest and the cost of the survey. The importance and computational complexity of this problem have motivated many contributions, which consider one of the two goals of the allocation problem, as described in Section 1. See for example Kokan (1963), Folks and Antle (1965), Kokan and Khan (1967), Huddleston, Claypool and Hocking (1970), Kish (1976), Bethel (1985, 1989), Chromy (1987), Valliant and Gentle (1997), Khan and Ahsan (2003), García and

Cortez (2006), Kozak (2006), Day (2010), Khan, Ali and Ahmad (2011), Ismail, Nasser and Ahmad (2011), Khan, Ali, Raghav and Bari (2012).

All of the above apply methods based on linear programming theory, convex programming, dynamic programming, multi-objective programming and heuristics to try and solve the multivariate optimal allocation problem. Here we propose two integer programming formulations to tackle the problem.

Formulation A

$$\text{Minimize} \quad \sum_{h=1}^{H} c_h n_h \tag{2.7}$$

$$\text{s.t. } n_{\min} \leq n_h \leq N_h, \; h = 1, \ldots, H \tag{2.8}$$

$$\sqrt{V\left(t_j\right)} \big/ Y_j \leq \mathrm{CV}_j \quad j = 1, \ldots, m \tag{2.9}$$

$$n_h \in Z_+ \quad h = 1, \ldots, H \tag{2.10}$$

where c_h represents the unit level survey cost for sampling from stratum h.

In this formulation, the objective function to be minimized (2.7) corresponds to the overall variable cost budget for the survey (which we denote by C). If the unit level survey costs for sampling from the various strata are unknown or are assumed to be the same, then c_h may all be set to one and the alternative objective function to minimize is $n = \sum_{h=1}^{H} n_h$, namely the overall sample size.

Constraint (2.8) ensures that at least $n_{\min}$ units are allocated to each stratum, and that the sample size will not exceed the population size for the stratum.

Constraint (2.9) ensures that the CV of the HT estimator of total for each survey variable is below a pre-specified threshold CV_j $(j = 1, \ldots, m)$ called target CV. Finally, constraint (2.10) ensures that all the allocated sample sizes are integers.

Note that the constraints (2.9) may be rewritten as:

$$\frac{V(t_j)}{Y_j^{\,2}\ \mathrm{CV}_j^2} \leq 1\ , j = 1, \ldots, m. \tag{2.11}$$

Now replacing the numerator in (2.11) by equation (2.6), leads to:

$$\sum_{h=1}^{H} \left(\frac{N_h^2 S_{hj}^2}{n_h Y_j^2 \mathrm{CV}_j^2} - \frac{N_h S_{hj}^2}{Y_j^2 \mathrm{CV}_j^2} \right) \leq 1, \quad j = 1, \ldots, m. \tag{2.12}$$

Defining

$$p_{hj} = \frac{N_h\ S_{hj}^2}{Y_j^2\ \mathrm{CV}_j^2} \tag{2.13}$$

the constraints (2.12) may be written as:

$$\sum_{h=1}^{H}\left(\frac{N_h\ p_{hj}}{n_h} - p_{hj}\right) \leq 1,\quad j = 1,\ldots,m. \tag{2.14}$$

Formulation B

$$\text{Minimize } \sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left[\sum_{h=1}^{H} N_h^2\ \left(\frac{1}{n_h} - \frac{1}{N_h}\right)\ S_{hj}^2\right] \tag{2.15}$$

$$\text{s.t. } n_{\min} \leq n_h \leq N_h,\quad h = 1,\ldots,H \tag{2.16}$$

$$\sum_{h=1}^{H} c_h n_h \leq C \tag{2.17}$$

$$n_h \in Z_+\quad h = 1,\ldots,H \tag{2.18}$$

$$0 < w_j < 1\ \ \forall j\ \text{ and }\ \sum_{j=1}^{m} w_j = 1 \tag{2.19}$$

where w_j are variable-specific weights, set a priori to represent the relative importance of the survey variables. The variable-specific weights w_j are set by subject matter experts or the survey designers. If they are not specified, equal relative weights could be assigned to all the survey variables considered.

In this formulation, the objective function (2.15) to be minimized corresponds to a weighted sum of the relative variances of the estimates of total for the m survey variables. We use relative variances because different survey variables may be measured in different units, and thus summing variances is not meaningful. Examining (2.15) it is clear that its minimum is achieved when

$$\sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left[\sum_{h=1}^{H} N_h^2\ \left(\frac{1}{n_h}\right)\ S_{hj}^2\right]$$

is minimum, since the last term

$$\sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left[\sum_{h=1}^{H} N_h^2\ \left(-\frac{1}{N_h}\right)\ S_{hj}^2\right]$$

does not depend on the stratum sample sizes. Hence the objective function (2.15) may be rewritten:

$$\text{Minimize } \sum_{j=1}^{m} w_j \frac{1}{Y_j^2}\left(\sum_{h=1}^{H} \frac{N_h^2}{n_h}\ S_{hj}^2\right). \tag{2.20}$$

Constraint (2.16) is the same as constraint (2.8) applied in Formulation A. Constraint (2.17) ensures that the total variable cost of the survey will not exceed the allocated budget C. Like constraint (2.10) in Formulation A, constraint (2.18) ensures that all the allocated sample sizes are integers. Constraint (2.19) ensures that the importance weights are adequate for aggregating the relative variances of the estimated totals for each of the survey variables.

When the unit level survey costs c_h per stratum are not known or may be assumed to be equal, constraint (2.17) may be replaced by $\sum_{h=1}^{H} n_h \leq n$ where n is the (maximum) overall sample size.

Both formulations A and B present non-linearity: constraint (2.9) or (2.14) in Formulation A, and the objective function in Formulation B. Therefore a first alternative one could use to resolve the non-linearity problem in these two Formulations would be one of the methods of non-linear programming or convex programming (Bazaraa, Sheralli and Shetty 2006; Luenberger and Ye 2008) that can deal with constraints, as for example penalty based methods or multiplier methods, amongst others. Nevertheless, application of such methods tends to produce solutions (sets of samples sizes to allocate in the strata) that, in general, are non-integers. In addition, when such solutions are rounded to obtain feasible sample sizes, there's no guarantee to obtain a global optimum (Wolsey 1998) in terms of minimizing the corresponding objective functions.

Alternatively, given that the solutions (sample sizes) must be integers, one could consider applying integer programming methods, such as *Branch and Bound* (Land and Doig 1960; Wolsey 1998; Wolsey and Nemhauser 1999). However, the non-linearity present in both formulations prevents the immediate application of such methods.

With these issues in mind, in the next section we propose two new formulations for integer programming that circumvent these problems and are equivalent to the Formulation A, defined jointly by (2.7), (2.8), (2.9) and (2.10), and Formulation B, defined jointly by (2.20), (2.16), (2.17), (2.18) and (2.19). More specifically, from the resolution of these new formulations it is possible to obtain integer sample sizes (n_h) for the sample allocation which satisfy the constraints established for each problem and also lead to a global optimum (Wolsey 1998) either for the objective function defined in (2.7), or for the objective function defined in (2.20), respectively.

3 Proposed formulations

From an optimization point of view, solving the problems defined by $(2.7)-(2.10)$ or by $(2.16)-(2.20)$ consists of determining $n_1, n_2, \ldots n_H$ chosen from the sets defined by $A_h = \{n_{\min}, \ldots, N_h\}$, $h = 1, \ldots, H$, that the constraints in each of these problems are satisfied and the corresponding objective function is minimized. As already indicated, a standard minimum sample size per stratum of $n_{\min} = 2$ is considered here to define the sets A_h, but this value may be changed as needed to accommodate specific survey requirements.

Taking this approach, a new formulation may be considered where the decision variables are indicator variables of which elements of the sets A_h $(h = 1, \ldots, H)$ will be chosen. For this purpose, we introduce the binary variable x_{hk} taking the value 1 if the sample size $k \in A_h$ is allocated to stratum h, and value 0 if this sample size is not allocated to stratum $h, h = 1, \ldots, H$.

Considering the formulations previously presented and these new binary variables, we may write two integer programming formulations where the decision variables (i.e., the unknowns to be determined) are of the $0-1$ type, therefore configuring a binary integer programming problem (Wolsey and Nemhauser 1999). The formulation equivalent to Formulation A is given by:

Formulation C

$$\text{Minimize } \sum_{h=1}^{H} c_h \left(\sum_{k=n_{\min}}^{N_h} k\, x_{hk} \right) \tag{3.1}$$

$$\text{s.t. } \sum_{k=n_{\min}}^{N_h} x_{hk} = 1 \;\forall h = 1, \ldots, H \tag{3.2}$$

$$\sum_{h=1}^{H} N_h\, p_{hj} \left(\sum_{k=n_{\min}}^{N_h} \frac{x_{hk}}{k} \right) - \sum_{h=1}^{H} p_{hj} \leq 1, \quad j = 1, \ldots, m \tag{3.3}$$

$$x_{hk} \in \{0,1\}, \quad k = n_{\min}, \ldots, N_h, h = 1, \ldots, H. \tag{3.4}$$

In Formulation C, constraint (3.2) ensures that, for each of the strata, there will be exactly one x_{hk} variable taking the value one. This is equivalent to ensuring the choice of only one value k (the sample size) from each set A_h $(h = 1, \ldots, H)$. Constraint (3.3) is equivalent to constraint (2.9) or its equivalent (2.14) in Formulation A. This formulation contemplates potentially varying unit survey costs for the various strata. If this is not necessary, the objective function in (3.1) may be redefined as

$$\text{Minimize } \sum_{h=1}^{H} \sum_{k=n_{\min}}^{N_h} k\, x_{hk}. \tag{3.5}$$

In order to help with the understanding of the proposed formulation, consider the following example.

Example 1: Suppose that there are three population strata $(H = 3)$ with $N_1 = 3, N_2 = 5$ and $N_3 = 4$, that the unit survey costs are the same across strata (say $c_h = 1\ \forall h$) and only one survey variable $(m = 1)$. Formulation C would then look like:

$$\text{Minimize } 1\,x_{11} + 2\,x_{12} + 3\,x_{13} + 1\,x_{21} + 2\,x_{22} + 3\,x_{23} + 4\,x_{24} + 5\,x_{25} + 1\,x_{31} + 2\,x_{32} + 3\,x_{33} + 4\,x_{34} \tag{3.6}$$

$$\text{s.t. } x_{11} + x_{12} + x_{13} = 1 \tag{3.7}$$

$$x_{21} + x_{22} + x_{23} + x_{24} + x_{25} = 1 \tag{3.8}$$

$$x_{31} + x_{32} + x_{33} + x_{34} = 1 \tag{3.9}$$

$$N_1\, p_{11} \left(1\, x_{11} + \frac{1}{2}\, x_{12} + \frac{1}{3}\, x_{13}\right) - p_{11} + N_2\, p_{21} \left(1\, x_{21} + \frac{1}{2}\, x_{22} + \frac{1}{3}\, x_{23} + \frac{1}{4}\, x_{24} + \frac{1}{5}\, x_{25}\right) - p_{21} + N_3\, p_{31} \left(1\, x_{31} + \frac{1}{2}\, x_{32} + \frac{1}{3}\, x_{33} + \frac{1}{4}\, x_{34}\right) - p_{31} \leq 1 \tag{3.10}$$

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{31}, x_{32}, x_{33}, x_{34} \in \{0,1\}. \tag{3.11}$$

Formulation B may also be translated to this new approach of using the binary variables as follows.

Formulation D

$$\text{Minimize } \sum_{j=1}^{m} w_j \frac{1}{Y_j^2} \sum_{h=1}^{H} \left(\sum_{k=n_{\min}}^{N_h} \frac{x_{hk}}{k} \right) N_h^2 S_{hj}^2 \tag{3.12}$$

$$\text{s.t. } \sum_{k=n_{\min}}^{N_h} x_{hk} = 1 \quad \forall h = 1, \ldots, H \tag{3.13}$$

$$\sum_{h=1}^{H} c_h \left(\sum_{k=n_{\min}}^{N_h} k \, x_{hk} \right) \leq C \tag{3.14}$$

$$x_{hk} \in \{0,1\}, \; k = n_{\min}, \ldots, N_h, h = 1, \ldots, H. \tag{3.15}$$

In Formulation D the objective function (3.12) is equivalent to the objective function (2.20). Constraint (3.13) is equivalent to constraint (2.16). Constraint (3.14) is equivalent to constraint (2.17) and ensures that the total variable cost of the survey will not exceed the allocated budget C. In case we do not have information on unit survey costs per strata, or wish to consider that they are the same across the strata, we replace constraint (3.14) by

$$\sum_{h=1}^{H} \sum_{k=n_{\min}}^{N_h} k \, x_{hk} \leq n. \tag{3.16}$$

In order to illustrate the proposed formulation, consider the following example.

Example 2: Suppose that there are two population strata $(H = 2)$ with $N_1 = 3$ and $N_2 = 4$, with two survey variables $(m = 2)$, equal unit survey costs for both strata, importance weights w_j equal to ½ for both survey variables and a total sample size of $n = 5$. Formulation D would then look like:

$$\begin{aligned} \text{Minimize } & \left[x_{11} \frac{N_1^2}{1} + x_{12} \frac{N_1^2}{2} + x_{13} \frac{N_1^2}{3} \right] \frac{1}{2} \left(\frac{S_{11}^2}{Y_1^2} + \frac{S_{12}^2}{Y_2^2} \right) + \\ & \left[x_{21} \frac{N_2^2}{1} + x_{22} \frac{N_2^2}{2} + x_{23} \frac{N_2^2}{3} + x_{24} \frac{N_2^2}{4} \right] \frac{1}{2} \left(\frac{S_{21}^2}{Y_1^2} + \frac{S_{22}^2}{Y_2^2} \right) \end{aligned} \tag{3.17}$$

$$\text{s.t. } x_{11} + x_{12} + x_{13} = 1 \tag{3.18}$$

$$x_{21} + x_{22} + x_{23} + x_{24} = 1 \tag{3.19}$$

$$1 x_{11} + 2 x_{12} + 3 x_{13} + 1 x_{21} + 2 x_{22} + 3 x_{23} + 4 x_{24} \leq 5 \tag{3.20}$$

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24} \in \{0,1\}. \tag{3.21}$$

In this paper, these two formulations were resolved applying a method of implicit enumeration called *Branch and Bound*. *Branch and Bound* (Wolsey 1998, Wolsey and Nemhauser 1999) methods obtain the optimal solution for binary integer programming problems efficiently, by considering the

resolution of a subset of problems associated with the feasible region for the problem. These methods were developed from the pioneer work of Land and Doig (1960).

The solutions for both Formulations C and D of the approach, labelled BSSM (the initials for Brito, Silva, Semaan and Maculan), were obtained using the R package *Rglpk*. The R code we developed is available on request. The package *Rglpk* contains a set of procedures that can be applied for solving linear and integer programming problems.

For comparison purposes in the case of our Formulation C, we also considered in our numerical illustrations an algorithm proposed by (Bethel 1985 and 1989) which is available in the R package *SamplingStrata*. This algorithm relies on the Kuhn-Tucker Theorem, and uses the Lagrange multipliers (Bazaraa, et al. 2006). In the case of our Formulation D, we compared our approach with a 'textbook method' proposed in Cochran (1977, Section 5.A.4), as suggested by the Associate Editor.

4 Numerical results

This section provides results for the application of the selected multivariate optimum allocation approaches to a set of population datasets. The approaches considered include:

- The BSSM algorithms developed to solve Formulations C and D provided in Section 3;
- An improved version of Bethel's algorithm (Bethel 1989) developed by Ballin and Barcarolli (2008);
- The textbook method proposed in Cochran (1977, Section 5.A.4).

Eleven population datasets were used for the numerical illustration, but for space considerations, here we report only the results for three of these populations. The three selected populations are described in tables A1 through A6 in Appendix A. Table A1 provides a brief description of each survey population and provides the list of the corresponding survey variables. Table A2 provides information about how each population was stratified prior to determining the optimum allocation. In particular, for the survey dataset called *MunicSw* the strata had been previously defined. The other two populations were stratified using a stratification algorithm available in the R package *stratification* or a classic $k-$means clustering method available in the *base* R package.

Table A3 presents the number of population strata (H), the number of survey variables (m), and the population size (N) for each of the populations considered. Tables A4 through A6 provide the population counts, means, and standard deviations per stratum for the survey variables considered in each of the three survey populations considered.

The results of all the numerical experiments reported here were obtained using the R packages and functions mentioned, and using a Windows 7 desktop computer with 24GB of RAM and with eight i7 processors of 3.40GHz. Processing time ranged from miliseconds (for the relatively small *MunicSw* population) to less than 4 seconds (for the larger *SchoolsNortheast* population, under formulation C). This demonstrates that the proposed formulations provide a feasible and efficient alternative for multivariate optimum allocation problems of small and medium size, for populations of sizes (N) in thousands and even tens of thousands.

Tables 4.1 through 4.3 provide the target coefficients of variation (CV_j) for each of the survey variables, the sample sizes obtained using the algorithm to solve proposed Formulation C (n_{BSSM}) and Bethel's algorithm (n_{Bethel}), and the achieved coefficients of variation for the estimators of totals of the survey variables considered in each population under the two algorithms compared.

Table 4.1
Results for the *CoffeeFarms* population

	Algorithm for Formulation C				Bethel's Algorithm			
CV_j (%)	n_{BSSM}	$\mathrm{CV}(t_1)$ (%)	$\mathrm{CV}(t_2)$ (%)	$\mathrm{CV}(t_3)$ (%)	n_{Bethel}	$\mathrm{CV}(t_1)$ (%)	$\mathrm{CV}(t_2)$ (%)	$\mathrm{CV}(t_3)$ (%)
5	**2,545**	1.24	5.00	2.92	2,546	1.23	5.00	2.91
10	**754**	3.30	10.00	7.01	755	3.30	9.99	7.07
15	**347**	5.21	15.00	11.01	349	5.11	14.95	10.85

Table 4.2
Results for the *SchoolsNortheast* population

	Algorithm for Formulation C			Bethel's Algorithm		
CV_j (%)	n_{BSSM}	$\mathrm{CV}(t_1)$ (%)	$\mathrm{CV}(t_2)$ (%)	n_{Bethel}	$\mathrm{CV}(t_1)$ (%)	$\mathrm{CV}(t_2)$ (%)
2	**1,624**	2.00	1.79	1,628	2.00	1.78
5	**294**	5.00	4.31	299	4.96	4.23
10	**80**	9.93	8.24	83	9.72	8.13

Table 4.3
Results for the *MunicSw* population

	Algorithm for Formulation C					Bethel's Algorithm				
CV_j (%)	n_{BSSM}	$\mathrm{CV}(t_1)$ (%)	$\mathrm{CV}(t_2)$ (%)	$\mathrm{CV}(t_3)$ (%)	$\mathrm{CV}(t_4)$ (%)	n_{Bethel}	$\mathrm{CV}(t_1)$ (%)	$\mathrm{CV}(t_2)$ (%)	$\mathrm{CV}(t_3)$ (%)	$\mathrm{CV}(t_4)$ (%)
5	**1,527**	2.01	3.88	5.00	4.41	1,529	2.00	3.88	4.99	4.40
10	**761**	3.61	7.27	9.99	8.77	763	3.60	7.25	9.97	8.75
15	**439**	5.01	10.22	14.98	13.07	441	4.95	10.16	14.94	13.03

As expected, in all cases the sample sizes obtained by solving Formulation C were smaller than (bold) or equal to those obtained using Bethel's algorithm. However, the improvements were generally not substantial. Nevertheless the proposed algorithm managed to improve upon the current best method in the nine scenarios considered (three populations times three levels for the target CVs). The improvements appeared to be a bit larger for the *SchoolsNortheast* Population, where the number of strata is also larger. Similar results (not shown here for conciseness but available from the authors on request) were obtained for the other eight populations considered in an initial version of the paper.

Tables 4.4 to 4.6 provide the results of applying Formulation D and the textbook method proposed in Cochran (1977, Section 5.A.4) to the same three survey populations. Now the goal is to minimize the weighted relative variance of the HT estimates of total, while keeping the overall sample size or cost. The first line in each of these tables contains the total sample sizes considered for the allocation. These sample sizes correspond to sampling fractions of 10%, 20% and 30% of the corresponding population sizes (N)

respectively, as indicated in the second line in each of the tables. The subsequent lines provide the allocation of the total sample into the strata, the coefficients of variation achieved for the HT estimates of totals of the survey variables considering the allocation, and the sum of the coefficients of variation $(\Sigma \mathrm{CV}(t_i))$, which is a summary measure of efficiency across all survey variables.

The importance weights were taken as equal across all survey variables, and the unit survey costs were taken as equal across all strata, in each population, for these applications.

Table 4.4
Results for the *CoffeeFarms* population

n	**2.047**		**4.094**		**6.142**	
Sampling fraction	**10%**		**20%**		**30%**	
Result	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**
n_1	1,174	1,124	2,483	2,340	3,792	3,625
n_2	662	737	1,400	1,544	2,139	2,306
n_3	211	186	211	210	211	211
$\mathrm{CV}(t_1)$	**1.02**	1.14	**0.62**	0.62	**0.42**	0.42
$\mathrm{CV}(t_2)$	**5.78**	5.79	**3.62**	3.65	**2.61**	2.63
$\mathrm{CV}(t_3)$	**2.86**	2.98	**1.73**	1.73	1.19	**1.17**
$\Sigma \mathrm{CV}(t_i)$	**9.66**	9.91	**5.97**	6.00	**4.22**	4.22

Table 4.5
Results for the *SchoolsNortheast* population

n	**7,508**		**15,017**		**22,525**	
Sampling fraction	**10%**		**20%**		**30%**	
Result	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**
n_1	82	58	82	60	82	66
n_2	36	33	62	53	53	62
n_3	7	6	7	6	7	6
n_4	206	214	465	433	771	611
n_5	1,083	1,000	2,091	1,962	2,671	2,121
n_6	447	452	891	914	1,428	1,436
n_7	361	371	711	750	1,182	1,175
n_8	2,995	2,989	5,963	6,055	9,088	9,634
n_9	976	1,023	1,965	2,069	3,078	3,229
n_{10}	399	419	800	849	1,331	1,338
n_{11}	797	813	1,742	1,647	2,596	2,612
n_{12}	119	130	238	219	238	235
$\mathrm{CV}(t_1)$	**0.86**	0.98	**0.54**	0.69	**0.39**	0.54
$\mathrm{CV}(t_2)$	0.73	**0.72**	**0.47**	0.47	0.35	**0.34**
$\Sigma \mathrm{CV}(t_i)$	**1.59**	1.70	**1.01**	1.16	**0.74**	0.88

Table 4.6
Results of formulation D for the *MunicSw* population

n	**290**		**579**		**869**	
Sampling fraction	**10%**		**20%**		**30%**	
Result	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**
n_1	67	59	134	118	202	182
n_2	68	77	136	153	206	233
n_3	40	35	80	70	120	107
n_4	58	47	116	93	171	128
n_5	32	43	65	85	97	129
n_6	16	21	31	43	47	65
n_7	9	8	17	17	26	25
$\mathbf{CV}(t_1)$	5.93	**5.40**	4.01	**3.61**	3.10	**2.75**
$\mathbf{CV}(t_2)$	12.53	**12.24**	8.36	**8.12**	6.36	**6.14**
$\mathbf{CV}(t_3)$	**19.49**	20.19	**12.46**	13.01	**8.95**	9.56
$\mathbf{CV}(t_4)$	**16.91**	17.45	**10.85**	11.27	**7.84**	8.30
$\Sigma\mathbf{CV}(t_i)$	**54.86**	55.28	**35.68**	36.01	**26.25**	26.75

As expected, in all three cases the sum of the coefficients of variation obtained by solving Formulation D were smaller than (bold) those obtained using the textbook algorithm. However, the textbook algorithm provided smaller CVs for some of the survey variables, in particular for the *MunicSw* population. The improvements were generally not very large, but again were slightly larger for the *SchoolsNortheast* population. In this comparison, however, the allocations are quite different between the two methods.

5 Final remarks

In this paper we provided two new formulations leading to the achievement of the global minimum in multivariate optimum allocation problems. These exact integer programming formulations can be efficiently implemented using off the shelf free software (namely the *Rglpk* R package). In addition, the proposed formulations enable the definition of minimum sample sizes per strata, something which is clearly of interest in practice to avoid allocations with sample sizes less than 2, for example, which would lead to difficulties regarding variance estimation. Such minimum sample sizes may be set at larger values (say 5, 10, 30 or some other number) to ensure that the samples are large enough to tolerate some nonresponse or to ensure estimation is feasible for each stratum, if the strata are used as estimation domains.

The proposed approach improves upon the existing methods by tackling the allocation problem directly, and dealing with the non-linearity of either the objective function or the constraints, as well as the requirement that the solution provides only integer sample sizes for the strata. In the literature, previously

existing methods tackle the problem with approaches which are not guaranteed to reach the global optimum, or that produce real-valued allocations that must be rounded to integer-values.

In practice, finding real-valued allocations is not a big problem, unless the stratum population sizes N_h are very small or when there is a very large number of strata. In the first case, sampling one unit more, or less, can make a big change in the sampling fractions, which can cause some large impacts in the variances. In the second case, rounding the allocated sample sizes can make a difference in the total sample size n. When all the stratum population sizes N_h are relatively large, and the number of strata is reasonable, rounding non-integer sample sizes will not create a problem.

In this paper we carried out some limited numerical work, aimed essentially at demonstrating the feasibility of the proposed approach. The results obtained using Formulation C of the proposed approach are comparable to those achieved using the Bethel method, while providing integer-valued allocations that correspond to the global optimum. But given that only little differences were found between the two methods (BSSM and Bethel) in the applications considered, there may be little incentive to move to the BSSM method. The results obtained under Formulation D showed modest improvements over the textbook method used in the comparison.

Further research is needed to test the approach for larger problems and to assess its merits compared to other methods under other practical scenarios. An important advantage of the proposed approach is that both formulations can be implemented using off the shelf software, as indicated.

Acknowledgements

This research was supported by FAPERJ. Research Grant E-26/111.947/2012.

Appendix A

Description of the survey populations considered in the numerical experiment

Table A1
Description of the populations

Population	Description	Survey Variables (y)
CoffeeFarms	Coffee farms in the state of Paraná, Brazil, from 1996 Agricultural Census.	Number of Coffee Trees Total Farm Area Coffee Production
SchoolsNortheast	Data from the 2012 census of schools, by school, for schools in the Northeast region of Brazil.	Number of classrooms Number of employees
MunicSw	Information about Swiss municipalities from the package *SamplingStrata*.	Area of Farming Industrial Area Number of Households Population

Table A2
Stratification of the populations

Population	Stratification
CoffeeFarms	Stratified considering the Number of Coffee Trees variable, using the Kozak algorithm available in the *Stratification* package.
SchoolsNortheast	Twelve strata were formed considering: school type (4 classes), and school size - number of students (3 classes). School size stratification was performed using $k-$means clustering algorithm within each school type.
MunicSw	This population is available from the *SamplingStrata* package and the strata correspond to regions of Switzerland.

Table A3
Number of strata, number of survey variables and total size for the survey populations considered

Population	H	m	N
CoffeeFarms	3	3	20,472
SchoolsNortheast	12	2	75,084
MunicSw	7	4	2,896

Table A4
Population summaries per stratum $-$ *CoffeeFarms*

Summary	Stratum		
	$h=1$	$h=2$	$h=3$
N_h	17,821	2,440	211
$\bar{Y}_{1h}$	4,291	26,688	218,712
$\bar{Y}_{2h}$	22	84	488
$\bar{Y}_{3h}$	2,671	13,204	129,033
S_{h1}	2,873	15,541	193,366
S_{h2}	69	262	583
S_{h3}	4,611	24,704	200,447

Table A5
Population summaries per stratum $-$ *SchoolsNortheast*

Stratum	N_h	$\bar{Y}_{1h}$	$\bar{Y}_{2h}$	S_{h1}	S_{h2}
$h=1$	82	45.1	54.0	309.2	24.9
$h=2$	63	23.9	146.3	14.4	92.6
$h=3$	7	80.9	700.4	29	342.5
$h=4$	783	16.2	95.7	6.4	49.5
$h=5$	2,676	10.9	57.7	21.6	23.7
$h=6$	3,958	6.1	26.7	4.2	17.9
$h=7$	2,172	13.6	76.8	5.7	27.9
$h=8$	45,243	2.5	9.3	3	8.8
$h=9$	9,674	7.7	38.0	3.2	17.9
$h=10$	1,743	17.3	49.1	9.2	36.7
$h=11$	8,445	7.3	15.3	4.1	13.5
$h=12$	238	37.7	140.8	18.4	88.9

Table A6
Population summaries per stratum – *MunicSw*

Summary	Statum						
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$
N_h	589	913	321	171	471	186	245
$\bar{Y}_{1h}$	262.5	367.2	262.7	438.0	429.5	668.9	47.0
$\bar{Y}_{2h}$	5.5	5.3	9.7	13.3	7.9	11.0	4.1
$\bar{Y}_{3h}$	963.9	782.1	1,345.2	3,319.1	906.0	1,465.2	550.7
$\bar{Y}_{4h}$	2,252.5	1,839.4	3,099.5	7,297.7	2,226.0	3,675.8	1,252.4
S_{h1}	220.5	342.4	173.2	290.2	414.2	568.7	65.3
S_{h2}	15.1	13.0	19.4	29.7	14.9	15.5	8.2
S_{h3}	4,600.9	2,794.7	5,003.5	14,610.0	2,178.6	2,802.1	1,197.5
S_{h4}	9,540.3	5,621.6	9,764.5	28,589.4	4,759.4	5,914.5	2,514.9

References

Ballin, M., and Barcaroli, G. (2008). Optimal stratification of sampling frames in a multivariate and multidomain sample design. *Contributi ISTAT*, 10.

Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (2006). *Nonlinear Programming: Theory and Algorithms*. New York: John Wiley & Sons, Inc, Third Edition.

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57.

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. Third Edition-Wiley.

Day, C.D. (2010). A multi-objective evolutionary algorithm for multivariate optimal allocation. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Folks, J.L., and Antle, C.E. (1965). Optimum allocation of sampling units to strata when there are R responses of interest. *Journal of the American Statistical Association*, 60 (309), 225-233.

García, J.A.D., and Cortez, L.U. (2006). Optimum allocation in multivariate stratified sampling: Multi-objective programming. *Comunicaciones Del Cimat*, no I-06-07/28-03-2006.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming. *Journal of the Royal Statistical Society, Series C*, 19 (3).

Ismail, M.V., Nasser, K. and Ahmad, Q.S. (2011). Solution of a multivariate stratified sampling problem through Chebyshev's Goal programming. *Pakistan Journal of Statistics and Operation Research*, vol. vii, 1, 101-108.

Khan, M.G.M., and Ahsan, M.J. (2003). A note on optimum allocation in multivariate stratified sampling. *The South Pacific Journal of Natural Science*, 21, 91-95.

Khan, M.F., Ali, I. and Ahmad, Q.S. (2011). Chebyshev approximate solution to allocation problem in multiple objective surveys with random costs. *American Journal of Computational Mathematics*, 1, 247-251.

Khan, M.F., Ali, I., Raghav, Y.S. and Bari, A. (2012). Allocation in multivariate stratified surveys with non-linear random cost function. *American Journal of Operations Research*, 2, 100-105.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Series A*, 139 (1), 80-95.

Kokan, A.R. (1963). Optimum allocation in multivariate surveys. *Journal of the Royal Statistical Society, Series A*, 126 (4), 557-565.

Kokan, A.R., and Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society, Series B*, 29 (1), 115-125.

Kozak, M. (2006). Multivariate sample allocation: Application of random search method. *Statistics in Transition*, 7 (4), 889-900.

Land, A.H., and Doig, A.G. (1960). An Automatic method for solving discrete programming problems. *Econometrica*, 28 (3), 497-520.

Lohr, S.L. (2010). *Sampling: Design and Analysis*, Second edition. Brooks/Cole, Cengage Learning.

Luenberger, D.G., and Ye, Y. (2008*). Linear and Non-Linear Programming*, Third Edition. Springer.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Valliant, R., and Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25, 337-360.

Wolsey, L.A. (1998). *Integer Programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization.

Wolsey, L.A., and Nemhauser, G.L. (1999). *Integer and Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization.