An Optimization Approach
Applied to Optimal Allocation in Stratified Sampling

José André de Moura Brito
Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Rio de Janeiro, RJ, Brazil
e-mail: jambrito@gmail.com

Pedro Luis do Nascimento Silva
Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Rio de Janeiro, RJ, Brazil
e-mail: pedronsilva@gmail.com

Augusto Fadel*
Fundação Instituto Brasileiro de Geografia e Estatística – IBGE
Rio de Janeiro, RJ, Brazil
e-mail: augustofadel@gmail.com

Gustavo Silva Semaan
Universidade Federal Fluminense – INFES/UFF
Santo Antônio de Pádua, RJ, Brazil
e-mail: gustavosemaan@gmail.com

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 sample allocation in multivariate surveys and several methods have been proposed. Basically, those methods are divided into two classes: The first one involves forming a weighted average of the stratum relative variances and finding the optimal allocation for the average relative variance. The second class consists of allocation methods leading to coefficients of variation of estimators below specified thresholds for all survey variables of interest. This paper proposes a new optimization approach for the second problem. This approach is based on a non-deterministic optimization method. Several experiments showed the proposed approach provides efficient solutions to this problem.

Keywords: allocation; stratification; optimization.

1. Introduction

Most of the statistics produced by official statistics agencies in many countries derives 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 makes possible to produce estimates for the target population parameters with controlled precision and to have data from typically small samples of 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 significant estimates within the survey budget available. 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 used very frequently in a wide range of sample surveys.
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).

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 optimal allocation 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) to minimize the total sample size, subject to having Coefficients of Variation (CVs) for the estimates of totals of the \( m \) survey variables below specified thresholds or (ii) to minimize a weighted sum of relative variances of the estimates of totals for the \( m \) survey variables.

This paper presents a new approach based on developing and applying a non-deterministic optimization approach to achieve the first goal. This 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 and gives some final remarks and concludes the paper.

2. Stratified Sampling and the Optimal Allocation Problem

In stratified sampling (Cochran, 1977) 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 (1) and together form the entire population (2) such that:

\[
U_h \cap U_k = \emptyset \quad h \neq k \quad (1)
\]

\[
\bigcup_{h=1}^{H} U_h = U \quad (2)
\]

\[
N_1 + N_2 + \ldots + N_H = \sum_{h=1}^{H} N_h = N \quad (3)
\]

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 \( 2 \leq n_h \leq N_h \quad \forall h \) and \( n_1 + n_2 + \ldots + n_H = \sum_{h=1}^{H} n_h = n \).

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:

\[
S_{hj}^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_{ij} - \bar{y}_{hj})^2 \quad (4)
\]

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

\[
\bar{y}_{hj} = \frac{1}{N_h} \sum_{i \in U_h} y_{ij} = \bar{y}_{hj} / N_h \quad (5)
\]
for \( h = 1, \ldots, H \) and \( j = 1, \ldots, m \). The population total \( Y_j \) for the \( j \)-th survey variable is

\[
Y_j = \sum_{h=1}^{H} \sum_{j \in J_h} Y_{hj} = \sum_{h=1}^{H} Y_{hj}.
\]

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:

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

(6)

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 (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 some contributions which consider the first goal of the allocation problem, as described in Section 1, see, for example: Kokan (1963), Huddleston, Claypool and Hocking (1970), Bethel (1985, 1989), Day (2010) and Barcaroli (2014). All of the above apply methods based on nonlinear programming theory, convex programming, and heuristics to try and solve the multivariate optimal allocation problem. Here we propose a non-deterministic method to tackle the problem.

Minimize \( n = \sum_{h=1}^{H} n_h \)  

(7)

s.t. \( 2 \leq n_h \leq N_h, h = 1, \ldots, H \)  

(8)

\[ \sqrt{V(t_j) / Y_j} \leq CV_{j} \quad j = 1, \ldots, m \]  

(9)

\( n_h \in \mathbb{Z}_+ \quad h = 1, \ldots, H \)  

(10)

In this formulation, the objective function to be minimized (7) corresponds to the sum of the sample sizes allocated to the strata. Constraint (8) ensures that at least two units are allocated to each stratum, and that the sample size will not exceed the population size for the stratum. Constraint (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 (10) ensures that all the allocated sample sizes are integers.

3. Method

In this paper the multivariate optimal allocation problem will be solved using a non-deterministic optimization method known as biased random-key genetic algorithms (BRKGA).

According to Gonçalves and Resende (2011), the BRKGA was developed from the study of the random-key genetic algorithm – (RKGGA) which was introduced by Bean (1994). In RKGGA, chromosomes (solutions) are represented by a vector which values (genes) are obtained from real values randomly selected according to a uniform distribution [0,1]. After generation, a procedure called decoder is applied in each of those vectors (chromosomes). The decoder aims to establish a mapping between the values of the chromosomes and the possible viable solutions to the optimization problem for which the objective function should be computed. Thus, the decoder procedure is specific to each optimization problem in which a RKGGA is used. A RKGGA evolves from an initial population of random-key vectors to new populations, generating new chromosomes (feasible solutions) for a given number of generations, applying a Darwinian principle of selection. The initial population consists of \( p \) random-key vectors generated as described above. After applying the decoder and computing the objective function, the population is partitioned into two sets: a small one formed by \( p_e \) elite solutions, corresponding to the best solutions with respect to the objective function value, and another one formed by the \( p-p_e \) remaining solutions, where \( p_e < p\). To update the population, a new
A mutant is a random-key vector generated in the same way as the initial population vectors. In every generation a small number \( p_e \) of mutant solutions is introduced into the population and, as well as other solutions, can be decoded into viable solutions to the problem. In RKGA, Bean (1994) considers that each of the \( p-p_e-p_m \) solutions is produced for the next generation through crossover process for two (parent) solutions, randomly selected from the entire population of the current generation. The (children) solutions are produced by applying uniform crossover. The biased random-key genetic algorithm (BRKGA) differs from RKGA in the crossover process. More specifically, the \( p-p_e-p_m \) (children) solutions are produced through crossover for an elite set solution and a non-elite set solution. Figure 1 illustrates the evolution of a population from generation \( g \) to \( g+1 \). Since \( p_e < p_e \), the probability of selection of an elite set chromosome is \( 1/p_e \) (greater than \( 1/(p-p_s) \)), enabling elite set elements to pass its allele for the future generations.

Figure 1 – Solutions produced by BRKGA between two consecutive generations

Figure 2 illustrates the crossover process for two random-key vectors with ten genes each. Chromosome 1 (C1) refers to the elite individual and Chromosome 2 (C2) to the non-elite one. In this example the value of \( p_e = 0.7 \), i.e., the offspring inherits the allele of the elite parent with a probability of 0.7 and of the other parent with a probability of 0.3. A randomly generated real in the interval \([0,1]\) simulates the toss of a biased coin \( v_e \) vector. If the outcome is less than or equal to 0.7, then the child inherits the allele of the elite parent. Otherwise, it inherits the allele of the other parent.

As regards the optimal allocation problem, the crossover process provides the exchange of sample sizes between two solutions, generating children vectors (chromosomes) that correspond to the new allocation solutions.

Once the population of the next generation is completed, that is, when there are \( p \) chromosomes, considering elite, mutants and children chromosomes, each new mutant and child chromosome is evaluated using the objective function. Then, the population is once again partitioned in two sets (elite and non-elite) and a new generation \( (g+2) \) can be started. Figure 3 illustrates the steps of the BRKGA algorithm.

<table>
<thead>
<tr>
<th>C1</th>
<th>0.31</th>
<th>0.77</th>
<th>0.81</th>
<th>0.49</th>
<th>0.32</th>
<th>0.97</th>
<th>0.72</th>
<th>0.15</th>
<th>0.56</th>
<th>0.92</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>0.26</td>
<td>0.15</td>
<td>0.36</td>
<td>0.41</td>
<td>0.93</td>
<td>0.11</td>
<td>0.28</td>
<td>0.56</td>
<td>0.34</td>
<td>0.87</td>
</tr>
<tr>
<td>( v_e )</td>
<td>0.58</td>
<td>0.89</td>
<td>0.11</td>
<td>0.41</td>
<td>0.75</td>
<td>0.99</td>
<td>0.43</td>
<td>0.71</td>
<td>0.88</td>
<td>0.58</td>
</tr>
<tr>
<td>( v_f )</td>
<td>0.31</td>
<td>0.13</td>
<td>0.81</td>
<td>0.49</td>
<td>0.93</td>
<td>0.11</td>
<td>0.72</td>
<td>0.56</td>
<td>0.34</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Figure 2 – Uniform Crossover \( (p_e=0.7) \)

Figure 3 – BRKGA pseudo-code

```
g=0; Generate p vectors of random keys
While (g<Number_generations)
   decode each vector of random keys and sort solutions by their fitness (value of objective function)
   classify solutions as elite or non-elite
   copy elite solutions to next population and generate mutants in next population
   Combine elite and non-elite solutions and add offspring to next population
   g=g+1
End_while
return(best_solution) (chromosome with the lowest value of the objective function)
```
3.1. BRKGA for Allocation Problem – Representation, Decoder and Objective Function (fitness)

As regards the Allocation Problem, each chromosome was defined as a vector \( u \) with \( H \) positions (number of strata) filled with real values uniformly distributed in the interval [0,1]. This vector was then submitted to the decoder procedure, which produced a vector \( a=(n_1,n_2,...,n_h,...,n_H) \) containing the sample sizes for the \( H \) strata. Each \( n_h \) obtained from the decoding equation:

\[
n_h=n_{\text{min}} + \text{round}(u_h^*(N_h-n_{\text{min}})) \quad (h=1, ..., H)
\]

where \( n_{\text{min}} \) is the minimum sample size allocated to each stratum, \( u_h \) is the value of the \( h \)-th component of \( u \) and \( N_h \) is the population size in the \( h \)-th stratum. This decoding process ensures immediate compliance with constraints (8) and (10). This way, considering, for example, \( H=3 \), \( n_{\text{min}}=2 \), \( N_j=100 \), \( N_j=250 \) and \( N_j=80 \) and \( u=(0.75, 0.19, 0.45) \), we have \( a=(76,49,37) \).

The objective function \( f \) used in BRKGA generations has a penalty term \( P \) associated with the constraint (9): 

\[
f = \sum_{h=1}^{H} n_h + P .
\]

This term is derived from the calculation of the term

\[
M = \text{max}_{j=1,...,m}(CVa_j/CV_j).
\]

If \( M \leq 1 \), that is, all the coefficients of variation (\( CV_a \)) associated with the solution \( a=(n_1,n_2,...,n_H) \) are less than or equal to the coefficient of variation (\( CV_j \)) fixed \( \text{a priori} \), we have \( P=0 \). Otherwise, \( P = T^M \) (\( T \in \mathbb{R} \)). This penalty term ensures the generation and perpetuation of feasible solutions, that is, solutions that satisfy constraints (9) during successive generations produced by the BRKGA.

4. Numerical Results

This section provides results for the application of the selected multivariate optimum allocation approaches to a set of datasets. The approaches considered include: BRKGA and an improved version of Bethel’s algorithm (Bethel, 1985 and 1989) presented by Barcarolli (2014). The BRKGA algorithm was developed in R language and Bethel’s algorithm is available in SamplingStrata package. Eleven datasets were used for the numerical experiments, but for space considerations, here we report only the results for three of these datasets described in table 1.

**Table 1** – Number of strata, number of survey variables and total size for the survey populations considered

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( H )</th>
<th>( m )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SugarCaneFarms</td>
<td>4</td>
<td>3</td>
<td>338</td>
</tr>
<tr>
<td>MunicSw</td>
<td>7</td>
<td>4</td>
<td>2,896</td>
</tr>
<tr>
<td>SchoolsNorth</td>
<td>12</td>
<td>2</td>
<td>9,977</td>
</tr>
</tbody>
</table>

The results of the numerical experiments reported here were obtained using a Windows 8 desktop computer with 16GB of RAM and with four i7 processors of 2.90GHz. Furthermore, in all BRKGA experiments, the following values were adopted, respectively, for the parameters \( p \) (population size), \( p_e \) (elite size), \( p_m \) (mutant size), \( \rho_c \) (crossover probability), \( N_g \) (number of generations) and \( T \) (penalty factor): 10,000; 1,500; 3,000; 0.65; 100 and 900. These values were defined through computational experiments carried out \( \text{a priori} \).

For BRKGA algorithm processing time ranged from seconds (for the relatively small SugarCaneFarms dataset) to less than two minutes (for the larger dataset, SchoolsNorth). This demonstrates that the proposed algorithm provide a feasible and efficient alternative to multivariate optimum allocation problems of small and medium size, for populations of sizes (\( N \)) in thousands.

Tables 2 through 4 provide the target coefficients of variation (\( CV_j \)) for each of the survey variables, the sample sizes obtained using the BRKGA (\( n_{\text{BRKGA}} \)) algorithm and Bethel’s algorithm (\( n_{\text{Bethel}} \)), and the coefficients of variation achieved for the estimators of totals of the survey variables considered in each dataset under the two algorithms being compared. In all cases the sample sizes obtained by BRKGA were smaller than (bold) to those obtained using Bethel’s algorithm. The proposed algorithm managed to improve upon the current best method in 9 scenarios considered (three datasets times three levels for the target CVs). BRKGA also outperformed the Bethel’s algorithm in the remaining eight datasets not presented in this paper, producing smaller sample sizes.
Table 2 – Results for the SugarCaneFarms Dataset

<table>
<thead>
<tr>
<th>CV_j(%)</th>
<th>BRKGA Algorithm</th>
<th>Bethel’s Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n_BRKGA</td>
<td>CV(t1) (%)</td>
</tr>
<tr>
<td>2</td>
<td>248</td>
<td>1.85</td>
</tr>
<tr>
<td>5</td>
<td>110</td>
<td>4.59</td>
</tr>
</tbody>
</table>

Table 3 – Results for the MunicSw Dataset

<table>
<thead>
<tr>
<th>CV_j(%)</th>
<th>BRKGA Algorithm</th>
<th>Bethel’s Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n_BRKGA</td>
<td>CV(t1) (%)</td>
</tr>
<tr>
<td>5</td>
<td>1,527</td>
<td>2.01</td>
</tr>
<tr>
<td>10</td>
<td>761</td>
<td>3.57</td>
</tr>
<tr>
<td>15</td>
<td>439</td>
<td>4.97</td>
</tr>
</tbody>
</table>

Table 4 – Results for the SchoolsNorth Dataset

<table>
<thead>
<tr>
<th>CV_j(%)</th>
<th>BRKGA Algorithm</th>
<th>Bethel’s Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n_BRKGA</td>
<td>CV(t1) (%)</td>
</tr>
<tr>
<td>2</td>
<td>1,117</td>
<td>1.88</td>
</tr>
<tr>
<td>5</td>
<td>195</td>
<td>4.69</td>
</tr>
<tr>
<td>10</td>
<td>55</td>
<td>0.0927</td>
</tr>
</tbody>
</table>

The presented results produced by BRKGA indicate this new approach as a good alternative to solve allocation problem. This new approach improve upon the existing methods by tackling the allocation problem directly, and dealing with the non-linearity of either the objective function or the conditions, 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 using approaches which are not guaranteed to reach the global optimum, or that produce real-valued allocations that must be rounded to integer-values. As for future work, new versions of BRKGA will be developed adopting new solution representations and encodings for this problem.

References


