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

Survey Methodology

An optimisation algorithm applied to the one-dimensional stratification problem

by José André de Moura Brito, Tomás Moura da Veiga and
Pedro Luis do Nascimento Silva

Release date: June 27, 2019

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 numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

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.

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, June 2019
Vol. 45, No. 2, pp. 295-315
Statistics Canada, Catalogue No. 12-001-X

An optimisation algorithm applied to the one-dimensional stratification problem

José André de Moura Brito, Tomás Moura da Veiga and Pedro Luis do Nascimento Silva[1]

Abstract

This paper presents a new algorithm to solve the one-dimensional optimal stratification problem, which reduces to just determining stratum boundaries. When the number of strata *H* and the total sample size *n* are fixed, the stratum boundaries are obtained by minimizing the variance of the estimator of a total for the stratification variable. This algorithm uses the Biased Random Key Genetic Algorithm (BRKGA) metaheuristic to search for the optimal solution. This metaheuristic has been shown to produce good quality solutions for many optimization problems in modest computing times. The algorithm is implemented in the R package *stratbr* available from CRAN (de Moura Brito, do Nascimento Silva and da Veiga, 2017a). Numerical results are provided for a set of 27 populations, enabling comparison of the new algorithm with some competing approaches available in the literature. The algorithm outperforms simpler approximation-based approaches as well as a couple of other optimization-based approaches. It also matches the performance of the best available optimization-based approach due to Kozak (2004). Its main advantage over Kozak's approach is the coupling of the optimal stratification with the optimal allocation proposed by de Moura Brito, do Nascimento Silva, Silva Semaan and Maculan (2015), thus ensuring that if the stratification bounds obtained achieve the global optimal, then the overall solution will be the global optimum for the stratification bounds and sample allocation.

Key Words: Optimal stratification; Genetic algorithm; Integer programming; Nonlinear optimization; BRKGA Metaheuristic.

1 Introduction

Stratified sampling is a widely used approach to achieve efficiency in sampling designs. The substantial literature on optimal stratification (to be reviewed later in this paper) signals to both the importance of this topic for research and to its wide range of applications. Recently, Hidiroglou and Kozak (2017) compared optimization-based and approximate methods for one-dimensional stratification of skewed populations and concluded that optimization methods are superior and should be used in practice.

In this paper, we propose applying a new optimisation algorithm to determine the stratum boundaries, which we coupled with an approach to obtain the globally optimal sample size allocation to the defined strata. The one-dimensional stratification problem is addressed using a global optimisation technique (a metaheuristic) called Biased Random Key Genetic Algorithm (BRKGA), proposed by Gonçalves and Resende (2011). This technique does not ensure achieving the global optimum for the stratum boundaries, but has been shown to produce good quality solutions for many optimization problems in modest computing times (see Gonçalves and Resende, 2004; Gonçalves, Mendes and Resende, 2005; Festa, 2013 and Oliveira, Chaves and Lorena, 2017).

Our approach for sample allocation given a defined stratification (see de Moura Brito et al., 2015), namely a stratification by a specified variable and given number of strata, is based on an integer programming formulation, and always achieves the global optimum for either minimizing the total sample

1. José André de Moura Brito, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106, Centro, Rio de Janeiro RJ, 20231-050. E-mail: jambrito@gmail.com; Tomás Moura da Veiga, SHIN CA 08 Lote 01 Torre 01, 101 Lago Norte, Brasília DF, 71503-508. E-mail: tmvtomas@gmail.com; Pedro Luis do Nascimento Silva, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106 Centro, Rio de Janeiro, RJ, 20231-050. E-mail: pedronsilva@gmail.com.

size given precision constraints or minimizing variance for a fixed total sample size, while providing an exact integer sample allocation, and allowing the specification of minimum and maximum sample sizes per strata, as is often required in practical applications. The approach is implemented in the *stratbr* R package (see de Moura Brito et al., 2017a), thus providing a practical alternative to existing approximate methods, which it clearly outperforms in terms of efficiency. It also compares favourably with other optimization methods, which are not guaranteed to provide the optimal allocation given the stratification.

We compared this new approach with methods proposed by Dalenius and Hodges (1959), Gunning and Horgan (2004), Kozak (2004, 2006), Keskintürk and Er (2007), and de Moura Brito, Silva Semaan, Fadel and Brito (2017b), using a set comprising 27 real and artificial survey populations. Our empirical study is much larger than that of Hidiroglou and Kozak (2017), who have used only two populations in their comparison. It is also larger than other studies in earlier literature.

We have not considered, as suggested, comparing our approach with classification or regression trees or other machine learning algorithms that synthesise one or more covariates into groupings that can be used for strata. The main reason for this is that such methods do not consider the variance of the target sample estimator or the sample size given precision constraints as the criteria to optimize. Therefore, they cannot be expected to achieve the optimum for the problem we wish to address. In addition, for classification or regression trees the analyst must also specify a "response variable", in addition to the predictors or auxiliary variables. In many typical sampling situations, the analyst will not have access to data on such a "response variable", and must aim to minimize variance of the estimator for the total of the size or stratification variable instead (as is the case in most of the literature on this topic).

Although we have addressed only the "one-dimensional stratification problem", in that a single size measure is used for stratification, one could always use some predictive model or alternative variable reduction technique to summarise auxiliary variables or covariates into a single "x" or size variable to be used in our proposed approach. Nevertheless, our approach can easily be extended to address multivariate stratification coupled with optimum allocation given the nature of the components of the approach.

The paper is divided as follows: Section 2 contains the key concepts of stratified sampling. Section 3 contains a detailed description of the stratification problem. Section 4 presents the Biased Random Key Genetic Algorithm (BRKGA) and its novel implementation to resolve the stratification problem, in combination with the optimal allocation method proposed by de Moura Brito et al. (2015). Section 5 contains the results of the application of the proposed method compared to those of five other methods available in the literature previously mentioned. Section 6 presents the conclusions of the comparative study.

2 Stratified sampling

In stratified sampling, the first step is to partition the elements of the target population into well defined, preferably homogeneous, mutually exclusive and exhaustive subgroups called strata. Each population

element (unit) is the focus of the survey and provider of the information which it aims to obtain. Survey units can be households, people, farms, business establishments or companies, etc.

Stratified sampling is recommended in practice for several reasons:

- It can improve precision of the overall population estimates for a fixed total cost;
- It enables controlling sample sizes and precision of estimates for the strata, if required;
- It may facilitate balancing workload distribution;
- It may help reduce travelling costs between survey elements, if stratification includes geography.

When the strata are formed such that the intra-stratum variability is small for a key set of variables, stratification is considered successful, since this enables achieving better precision for the estimates relative to other stratification schemes.

Figure 2.1 presents the basic notation to be used in this paper. In stratified sampling, the population U is partitioned into $H > 1$ nonempty subpopulations called strata (Lohr, 2010), of sizes N_1, $N_2, \ldots, N_h, \ldots, N_H$. These subpopulations are non-overlapping and such that, when taken jointly, combine to form the full population, such that:

$$N_1 + N_2 + \ldots + N_H = N. \tag{2.1}$$

$U = \{1, 2, \ldots, N\}$ – Set of elements comprising the target population;

N – Number of population elements, or population size;

n – Number of elements in the sample, or sample size;

$H > 1$ – Total number of strata;

h – Index for the strata;

U_h – Set of elements in stratum h, satisfying $U_h \subset U$ and $U_h \neq \varnothing$;

N_h – Number of population elements, or population size, in stratum h;

s_h – Set of elements sampled in the h^{th} stratum $- s_h \subset U_h$;

n_h – Number of sample elements, or sample size, in stratum h;

y_i – Value of survey variable for population element $i\,(i \in U)$;

x_i – Value of stratification variable for population element $i\,(i \in U)$;

$Y_h = \sum_{i \in U_h} y_i$ – Total of survey variable y in stratum h;

$X_h = \sum_{i \in U_h} x_i$ – Total of stratification variable x in stratum h;

$\bar{Y}_h = Y_h / N_h$ – Population mean of survey variable y in stratum h; and

$\bar{X}_h = X_h / N_h$ – Population mean of stratification variable x in stratum h.

Figure 2.1 Notation to be used in the paper.

Cochran (1977) lists the following factors which affect the efficiency of a stratified sampling design: choice of stratification variable(s); number of strata (H); delimitation of strata; total sample size (n); allocation of the total sample to the strata; and selection method for sampling within strata. The strata are defined using one or more variables for which the values are known for each population element. In the sequence, samples are selected independently within each of the H strata. Samples sizes in the strata are such that $n_1 + n_2 + \ldots + n_H = n$.

Stratified Simple Random Sampling (SSRS) corresponds to the case when sampling within each stratum is carried out using simple random sampling without replacement. Under SSRS, the Horvitz-Thompson (HT) estimator of the overall population total $Y = \sum_{h=1}^{H} Y_h$ is given by Cochran (1977) as:

$$\hat{Y}_{\text{SSRS}} = \sum_{h=1}^{H} N_h \bar{y}_h \tag{2.2}$$

where $\bar{y}_h = \sum_{i \in s_h} y_i / n_h$ is the sample average for elements sampled in stratum h.

The sampling Variance (Var) and Coefficient of Variation (CV) of the estimator $\hat{Y}_{\text{SSRS}}$ are given respectively by:

$$\text{Var}\left(\hat{Y}_{\text{SSRS}}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) S_{hy}^2 / n_h \tag{2.3}$$

$$\text{CV}\left(\hat{Y}_{\text{SSRS}}\right) = \sqrt{\text{Var}\left(\hat{Y}_{\text{SSRS}}\right)} \Big/ Y \tag{2.4}$$

where $S_{hy}^2 = \sum_{i \in U_h} \left(y_i - \bar{Y}_h\right)^2 / \left(N_h - 1\right)$ is the population variance of the survey variable y in stratum h.

Analogous quantities are defined for the HT estimator for the total $X = \sum_{h=1}^{H} X_h$ of the stratification variable x, namely:

$$\hat{X}_{\text{SSRS}} = \sum_{h=1}^{H} N_h \bar{x}_h \tag{2.5}$$

$$\text{Var}\left(\hat{X}_{\text{SSRS}}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) S_{hx}^2 / n_h \tag{2.6}$$

$$\text{CV}\left(\hat{X}_{\text{SSRS}}\right) = \sqrt{\text{Var}\left(\hat{X}_{\text{SSRS}}\right)} \Big/ X \tag{2.7}$$

where $\bar{x}_h = \sum_{i \in s_h} x_i / n_h$ is the sample average of x for elements sampled in stratum h, and $S_{hx}^2 = \sum_{i \in U_h} \left(x_i - \bar{X}_h\right)^2 / \left(N_h - 1\right)$ is the population variance of the variable x in stratum h.

3 The one-dimensional stratification problem

Consider the population vector $X_U = \{x_1, x_2, \ldots, x_N\}$ corresponding to the stratification variable x. Without loss of generality, we assume that the population elements in U are ordered by the stratification variable such that $x_1 \leq x_2 \leq \ldots \leq x_N$. The stratum boundaries are used to define the H strata according to the rule:

1) $U_1 = \{i \in U \mid x_i \leq b_1\}$;

2) $U_h = \{i \in U \,|\, b_{h-1} < x_i \leq b_h\}$ for $h = 2, 3, \ldots, H-1$;

3) $U_H = \{i \in U \,|\, b_{H-1} < x_i\}$.

The stratification problem corresponds to determining the cut-off points, i.e., the stratum boundaries $b_1 < b_2 < \ldots < b_h < \ldots < b_{H-1}$ such that the variance (or equivalently the CV) of the estimator of total $\hat{Y}_{\text{SSRS}}$ is minimised. In this section, we consider that the total number of strata H is defined before applying the optimal stratification methods considered.

In practice, the values of the survey variable y are not available and hence the variance in expression (2.3) is not computable. A common approach is to minimise instead the variance (or CV) of the estimator $\hat{X}_{\text{SSRS}}$ for the total of the stratification variable x. Several authors have developed methods that focus on this optimization problem, which from now on we call the one-dimensional "stratification problem". We adopted the same approach here.

Finding the boundary points that minimize the variance (2.6) or the CV (2.7) corresponds to a hard problem both from analytic and computational points of view. This is so because the integer population and sample sizes (N_h and n_h, respectively) depend in a nonlinear way on the stratum boundaries. According to de Moura Brito, Ochi, Montenegro and Maculan (2010a), depending on N, H, and the number of distinct population x values, the number of possible choices for the boundary points can be very large.

In view of this difficulty, over the past decades, various methods were developed to search for the optimum stratum boundaries, aiming to provide at least solutions which correspond to local minima of good quality.

Dalenius (1951) tackled the problem for the case $H = 2$ by approximating the variance in (2.6) by ignoring the finite population correction, which is equivalent to assuming that the sampling within strata would have been simple random sampling with replacement. The approximate variance to be minimised is then given by:

$$\text{Var}\left(\hat{X}_{\text{SSRS}}\right) \cong \sum_{h=1}^{H} N_h^2 S_{hx}^2 / n_h. \tag{3.1}$$

Under the Neyman allocation (Cochran, 1977) using the x variable, and replacing the sample sizes n_h in (3.1) by their theoretical values $n_h = N_h S_{hx} \Big/ \sum_{k=1}^{H} N_k S_{kx}$ leads to the expression used by Dalenius (1951):

$$\text{Var}\left(\hat{X}_{\text{SSRS}}\right) \cong \left(\sum_{h=1}^{H} N_h S_{hx}\right)^2 \Big/ n. \tag{3.2}$$

Dalenius and Hodges (1959) considered the case when $H > 2$, and offered an analytic solution which relied on approximating the distribution of the x variable by its histogram with a moderate number of classes. Still considering the approximate variance and assuming Neyman's allocation, Ekman (1959) provided a solution using a geometric approach to find the stratum boundaries. Hedlin (2000) further extended Ekman's solution while retaining the original variance (2.6) as the function to be minimised, which he labelled the extended Ekman rule.

Hidiroglou (1986) proposed an approach which pre-specifies the required precision (CV) for the estimator of total, and which divides the population into two strata $(H = 2)$ such that the total sample size n is minimised. In this paper, the second stratum corresponds to a "take-all" or "certainty" stratum, where all elements are included in the sample with probability one $(n_2 = N_2)$. Lavallée and Hidiroglou (1988) generalized the approach to $H > 2$ strata, while retaining the idea that the stratum containing the largest population units is to be sampled completely. Their approach relied on adopting a special type of allocation called Power Allocation (Bankier, 1988). More recently, Rivest (2002) further generalised the approach of Lavallée and Hidiroglou (1988) while considering that the target is minimising the variance for the estimator of a total of a model-based prediction of the survey variable y, instead of the stratification variable x.

Gunning and Horgan (2004) proposed the so-called Geometric method for defining the stratum boundaries. This method assumes that the CVs of the stratification variable x are approximately constant, and that the distribution of the stratification variable is approximately uniform within each stratum. Under these assumptions, the optimum stratum boundaries would form a Geometric progression, thus leading to a very simple analytic solution.

Keskintürk and Er (2007) proposed an approach based on a global optimisation technique called Genetic Algorithms. Following a similar idea, de Moura Brito et al. (2017b) applied another global optimisation technique called GRASP to the stratification problem. Here we followed a route like Keskintürk and Er (2007), but have adopted an efficient choice of Genetic Algorithm, namely the Biased Random Key Genetic Algorithm (BRKGA), described in Section 4.

Kozak (2004) proposed a method called random search, followed later by Kozak and Verma (2006), where this approach was compared to the Geometric method of Gunning and Horgan (2004). Khan, Nand and Ahmad (2008) used ideas of dynamic programming to develop an algorithm that determines the stratum boundaries considering that the stratification variable has either a Triangular or Normal distribution, and that sampling within strata is with replacement. De Moura Brito, Maculan, Lila and Montenegro (2010b) proposed an exact algorithm based on graph theory, where proportional allocation to the strata is assumed.

Er (2011) compared the efficiency of several methods available in the literature, taking the Geometric approach of Gunning and Horgan (2004) as the initial solution. Kozak (2014) compared his random search with the Genetic Algorithm proposed by Keskintürk and Er (2007). Rao, Khan and Reddy (2014) developed a method that tackles the stratum boundary determination and stratum allocation problems simultaneously. Their algorithm relies on the assumption that the stratification variable follows a Pareto distribution. Our approach is more general and does not assume that the size variable follows a particular distribution.

4 Biased Random-Key Genetic Algorithm

The Biased Random-Key Genetic Algorithm (BRKGA, from now on), proposed by Gonçalves and Resende (2011), is a metaheuristic approach which has been applied to address several optimization

problems – see for example Festa (2013) and Oliveira et al. (2017). The principle behind this approach mimics the biologic theory of evolution of species.

The algorithm starts with an initial "population" of feasible solutions to the target problem generated by a specified random mechanism. This population then evolves over successive iterations by preserving the best solutions available at each iteration (elite solutions), and by replacing the other (non-elite) solutions with solutions generated through random perturbation operations that mimic crossing and mutation in natural populations. Over the iterations, solutions are selected to be preserved or to evolve based on the value of the function to be optimized.

In BRKGA, the candidate solutions are encoded, i.e., are represented by vectors where the components are numbers in the $(0; 1)$ interval. Given an observed vector, a decoding procedure must be applied. The decoding procedure maps the value of a vector with a corresponding feasible solution of the target optimization problem. The decoding procedure is what connects BRKGA with the specific optimization problem to be addressed. Figure 4.1 displays the pseudo-code for a generic BRKGA algorithm.

The approach is described and illustrated in detail in Section 4.1 using an example that considers the one-dimensional stratification problem and describes all the steps mentioned in Figure 4.1.

1) Generate the initial population composed of p random vectors (keys) $\mathbf{v}$, where each value is a random draw from the Uniform $[0; 1]$ distribution.
2) Apply the decoding procedure to each vector $\mathbf{v}$ in the population, yielding p feasible solutions to the optimization problem.
3) Compute the value of the objective function for each solution in the population.
4) Select the best $p_e\,(1 < p_e < p)$ solutions (designated elite) based on the values of the objective function and add them to the population that will be considered in the next iteration.
5) Generate $p_m\,(1 < p_m < p)$ new random vectors as in step 1), called *mutants*, and add them to the population that will be considered in the next iteration.
6) Generate the remaining $(p - p_e - p_m)$ vectors, designated *crossed*, to complete the population that will be considered in the next iteration by crossing one of the p_e vectors corresponding to an elite solution with one of the $(p - p_e)$ vectors corresponding to one of the non-elite solutions in the current iteration.
7) Iterate from step 2) while the stopping criteria are not satisfied.

Figure 4.1 Pseudo-code for a BRKGA.

4.1 BRKGA for the one-dimensional stratification problem

First consider the population vector $X_U = \{x_1, x_2, \ldots, x_N\}$, and derive the set $C = \{c_1, c_2, \ldots, c_K\}$ containing the K distinct values of x observed in the population. For example, if $X_U = \{1, 3, 3, 5, 6, 7, 7, 7, 8, 9, 10, 10, 11\}$, then $C = \{1, 3, 5, 6, 7, 8, 9, 10, 11\}$. When $K > 100$, compute the ten largest

percentiles of x to obtain the set $Q = \{q_{90}, q_{91}, \ldots, q_{99}, q_{100}\}$. When $K \leq 100$, compute the selected percentiles of x to obtain the set $Q = \{q_5, q_{10}, \ldots, q_{95}, q_{100}\}$. The cut-off point of 100 for K was chosen after some initial experimentation of the approach with some of the populations considered in the numerical experiments to be described in Section 5. The alternative definitions for the set Q help with achieving larger diversity in the set of feasible solutions to be generated by the BRKGA.

To apply BRKGA to the one-dimensional stratification problem, each solution is represented by a vector $\mathbf{v} = \{v_1, \ldots, v_H\}$ with H positions, where the first $H-1$ positions contain values between 0 and 1, and position H contains the value of a percentile of the distribution of the stratification variable x.

Then take $x_{\min}$ as the smallest value in C, and v_H as an element selected at random from Q. For the first iteration, sample the values in the first $H-1$ positions of each vector $\mathbf{v}$ independently from the Uniform $[0; 1]$ distribution.

The decoding procedure to obtain a solution to the one-dimensional stratification problem from each vector $\mathbf{v}$ generated is defined as:

$$b_h = x_{\min} + v_h (v_H - x_{\min}) \quad \text{for} \quad h = 1, \ldots, H-1. \tag{4.1}$$

After obtaining the $H-1$ values for b_h, these must be sorted in increasing order, such that the elements of the resulting vector $\mathbf{b} = \left(b_{(1)}, b_{(2)}, \ldots, b_{(H-1)}\right)$ form the solution boundary points for the corresponding vector $\mathbf{v}$, where $b_{(h)}$ is the h^{th} order statistic of the values $b_1, \ldots, b_{H-1}$ calculated using (4.1).

To illustrate an example of decoding, suppose that $H = 4$, $x_{\min} = 10$, $K = 300$, $Q = \{200, 215, 280.5, 300, 318, 400, 425, 478, 500, 510\}$ Consider also the vector $\mathbf{v} = (0.48, 0.35, 0.20)$ generated as described above. Then it follows that: $b_1 = 10 + 0.48 \times (200 - 10)$; $b_2 = 10 + 0.35 \times (200 - 10)$; and $b_3 = 10 + 0.20 \times (200 - 10)$. Then it follows, after sorting, that $\mathbf{b} = (48, 76.5, 101.2)$.

Given the vector $\mathbf{b}$, the values of N_h and S_{hx}^2 are easily obtained for each of the H strata. The values of the sample sizes n_h for each of the strata are obtained by applying the approach for optimal allocation proposed by de Moura Brito et al. (2015). This approach computes the sample sizes n_h such that a weighted sum of variances (or CVs) of the estimators of totals of m survey variables is minimized, while the total sample size n is kept fixed.

Since here we consider the variance of the estimator for the total of the stratification variable x as the target for minimization, we set $m = 1$ and use formulation (D) as provided in de Moura Brito et al. (2015) to resolve the one-dimensional optimal allocation problem taking equation (2.6) as the variance to be minimized. Note that the approach used provides the global optimum for the allocation problem.

The algorithm then proceeds as indicated in Figure 4.1 by generating an initial set of p vectors $\mathbf{v}$. In step 2, each of these vectors $\mathbf{v}$ is decoded to obtain a feasible solution $\mathbf{b}$ to the optimum stratification problem. In step 3, the optimum allocation corresponding to $\mathbf{b}$ is obtained and the value of the objective

function is calculated. Steps 4 to 6 are then applied to obtain the next population of feasible solutions, and the process is repeated until the stopping criteria are satisfied. Step 4 identifies the p_e elite solutions and add these to the next population. In Step 5, p_m mutant solutions are generated and added to the next population. In Step 6, $(p - p_e - p_m)$ crossed solutions are generated using the "uniform crossover" operator proposed by Spears and DeJong (1991) to produce a new vector $\mathbf{v}$ from one of the p_e elite solutions and one of the current $(p - p_e - p_m)$ non-elite solutions. The process is as follows: once the two vectors (say $\mathbf{v}_e$ and $\mathbf{v}_n$) to cross are selected, an auxiliary random-key vector $(\mathbf{v}_a)$ is generated with independent draws from the Uniform $[0; 1]$ distribution. Let $r_c > 0.5$ be a pre-specified probability that a value is copied from the elite vector $\mathbf{v}_e$. Then the crossed vector $\mathbf{v}_c$ is formed by taking the values from $\mathbf{v}_e$ in the positions where the corresponding value in $\mathbf{v}_a$ is less than r_c (equal to 0.7 in the example of Figure 4.2) and from $\mathbf{v}_n$ in all other positions.

To produce each one of the $(p - p_e - p_m)$ vectors for the next generation, the algorithm selects a vector v_e at random (using the *sample* function from R) from the p_e elite vectors and another vector v_n from the $p - p_e$ non-elite vectors, and crosses these vectors. The selection of vectors from both subsets is done with replacement, implying that individual elite or non-elite vectors may be selected for crossing more than once.

Vectors\positions	1	2	3
$\mathbf{v}_e$	0.31	0.77	0.65
$\mathbf{v}_n$	0.26	0.18	0.36
$\mathbf{v}_a$	0.58	0.89	0.11
$\mathbf{v}_c$	0.31	0.18	0.65

Figure 4.2 Uniform crossing with $r_c = 0.7$.

Now consider the example where $H = 4$, $x_{\min} = 10$, $K = 300$, $p = 8$, $p_e = 3$, $p_m = 3$, $r_c = 0.7$ and $Q = \{200, 215, 280.5, 300, 318, 400, 425, 478, 500, 510\}$. Figure 4.3 illustrates the application of all the steps in BRKGA to the one-dimensional stratification problem, for two consecutive iterations of the algorithm.

The BRKGA approach described here for the one-dimensional optimal stratification problem was implemented in the R package *stratbr* (see de Moura Brito et al., 2017a), which is available from CRAN. This package was used to obtain all the results presented in Section 5.

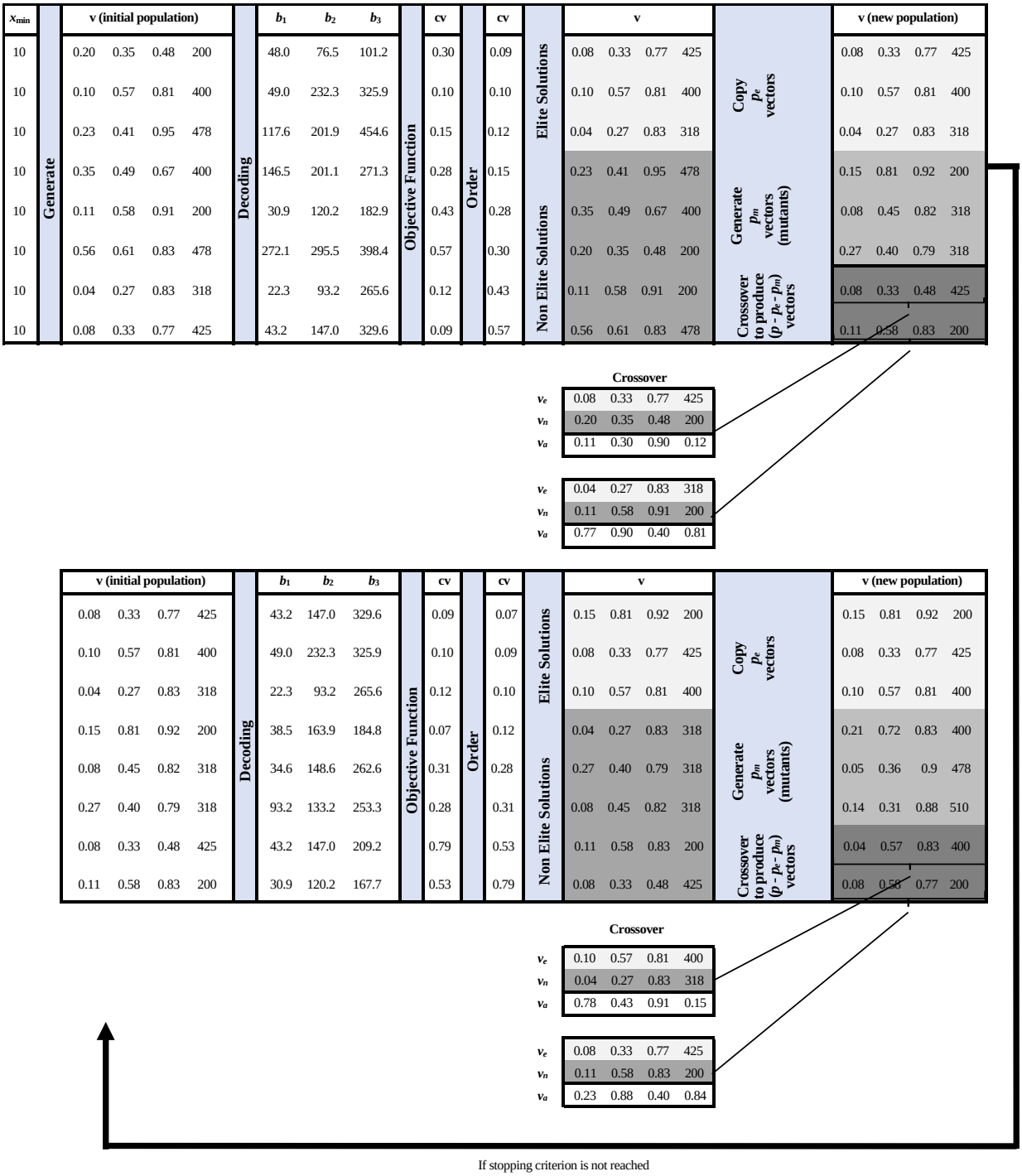

Figure 4.3 Illustration of BRKGA approach for optimal stratification.

5 Computational results

In this section, we present the results of application of six methods to solve the stratification problem, namely: Dalenius and Hodges (DH), Geometric (GH), Kozak (KO), Genetic Algorithm of Keskinturk and Er (KE), GRASP (GR) and the new BRKGA method described in Section 4 (BR). All experiments where carried out using R version 3.3.1. The methods DH, GH and KO are available from the R package *stratification* of Baillargeon and Rivest (2014) (version 2.2-5). With these methods, the Neyman sample allocation method was used. The KE method is available from the R package *GA4Stratification* of Er, Keskintürk and Daly (2010) (version 1.0). With this method, the maximum number of iterations considered was 10,000 and the values of the other parameters required were the same as those reported by Keskintürk and Er (2007), namely using $p = 35$ candidate solutions in each population, a mutation rate of 15% and the sample allocation based also on the Genetic Algorithm. Both the GR and the BR methods were implemented in R by the authors, and the code is provided in package *stratbr* of de Moura Brito et al. (2017a) (version 1.2) available from CRAN.

For the BR method, $p = 100$ candidate solutions were considered in each iteration, with 20% of the solutions being made elite $(p_e = 20)$ and 30% of the solutions being mutant $(p_m = 30)$ in each iteration. The probability of copying a gene from the elite vector was set at $r_c = 0.6$. The total number of iterations was set at 1,500. For the sample allocation, both the BR and the GR methods were coupled with the formulation proposed by de Moura Brito et al. (2015) which is available from the R package *MultAlloc*, also available from CRAN.

To compare the relative efficiency of these methods, they were applied to 27 different populations. Some of these populations are available from the R packages *stratification* and *GA4Stratification*, and were previously used in other comparison studies such as Keskintürk and Er (2007), Er (2011), and de Moura Brito et al. (2017b). Appendix A contains brief descriptions of all these populations, including information on which variable was considered as the "x variable" in each population. Table 5.1 provides some summaries to describe these populations.

The 27 populations considered here form a very diverse set, with total sizes varying from a few hundred (ME84 and P75 with $N = 284$ are the smallest) to several thousand (Coffee with $N = 18{,}570$ is the largest). In the size measure that matters most for efficiency of our optimization algorithm, namely the number K of distinct values of the stratification variable, there's also large variation (from $K = 51$ for Kozak1 to $K = 5{,}453$ to Kozak3). They also display wide variation in the asymmetry of the x variable's distributions, ranging from modestly negative (-0,70 for Beta103 to a substantial 40.04 for CensoCO).

All the calculations for the computational experiment were performed using R in a computer with 24 GB RAM, with 8 processors of 3.40 GHz (I7). Taking advantage of the multicore architecture in modern computers, the *snowfall* R package was used to parallelize the BRKGA algorithm. More specifically, at each iteration, the decoding procedure produces a set of solutions for the boundary points. These boundary points are then supplied to the *MultAlloc* package for optimum allocation, to obtain the sample sizes in each stratum, and then to compute the objective variance function. Since the computation time for this step is

impacted directly by the use of this global optimization formulation, the allocation and calculation of the objective function were parallelized.

Table 5.1
Summaries of the stratification variable for the 27 populations

Populations	N	K	Minimum	Maximum	Skewness
AgrMinas	844	226	5.00	47,800.00	7.32
BeefFarms	430	353	50.00	24,250.00	4.56
Beta103	1,000	1,000	357.98	985.96	-0.70
CensoCO	9,977	79	1.00	911.00	40.04
Chi5	1,000	1,000	0.06	23.43	1.40
Coffee	18,570	538	0.01	13,212.00	19.69
Debtors	3,369	1,129	40.00	28,000.00	6.44
HHinctot	16,025	224	1.00	6,900.00	2.71
Iso2004	487	487	6.36	1,044.66	10.03
Kozak1	4,000	51	72.00	3.00	1.40
Kozak3	2,000	581	2,793.00	6.00	3.55
Kozak4	10,000	5,453	74,400.00	62.00	4.20
ME84	284	264	173.00	47,074.00	8.64
MRTS	2,000	2,000	1.41	4,863.66	8.61
P100e10	1,000	1,000	73.56	127.32	-0.03
P75	284	68	4.00	671.00	8.43
Pop500	500	261	0.01	47,841.42	21.53
Pop800	800	402	0.01	4,735.10	22.13
pop1076	1,076	88	5.00	1,643.00	13.23
pop1616	1,616	165	5.00	2,618.00	11.09
pop2911	2,911	247	5.00	2,497.00	11.50
REV84	284	277	347.00	59,877.00	7.83
SugarCaneFarms	338	101	18.00	280.00	2.26
Swiss	2,896	881	0.00	3,634.00	2.73
USbanks	357	200	70.00	977.00	2.07
UScities	1,038	116	10.00	198.00	2.87
UScolleges	677	576	200.00	9,623.00	2.45

Note: N is the total population size, and K is the number of unique values for the stratification variable.

All six methods considered in the numerical experiment were applied to each of the 27 populations, for numbers of strata H equal to 3, 4, 5 and 6. These values were used since they are often considered in applications, as well as in similar comparative studies available in the literature, such as Er (2011) and Gunning and Horgan (2004). We did not consider larger values for H since the additional gains in efficiency for $H > 6$ are modest. The sample size $n = 100$ (i.e., fixed cost) was used, as in the numerical experiments of Er (2011) and Kozak and Verma (2006).

To assess the efficiency of the methods, the CVs of the estimator for the total of the stratification variable x were calculated for each population and number of strata, leading to $27 \times 4 = 108$ scenarios for each

method. CVs were obtained from equation (2.7) and multiplied by 100, to be presented as percentages. Table 5.2 provides the CVs attained by the six methods. The shaded cells indicate methods providing the best solution (minimum CV) for each of 108 scenarios. The NAs in these tables represent cases where solutions could not be obtained due to problems of the specific stratification method or with the corresponding allocation.

Analyzing the results provided in Table 5.2, and in particular, the shaded cells, it is evident that BR has excellent performance when compared to the five competitors considered. This perception is reinforced by the plots in Figure 5.1, where BR was compared with all competitors. Points above the straight line represent scenarios where the method chosen for comparison is outperformed by BR. It is evident from these plots that the three best performing methods are GR, KO and BR.

Table 5.3 provides the percentage of times that each method produced the best solution over the 108 scenarios. Both BR and KO display performance which is superior to that of the other methods and have tied in the number of times that they have achieved the best solution. DH produced the best solution for only three of the 108 scenarios, and GH has never produced a best solution.

The Geometric method GH, besides leading to high CVs, also often provided infeasible solutions, where the stratum limits lead to allocations where sample sizes were larger than the corresponding population sizes. This method also sometimes partitioned the population such that there were very few population elements in some strata. According to Gunning and Horgan (2004), and as noted by Keskintürk and Er (2007), since the interval widths increase geometrically, the GH method will not perform well when the stratification variable has small values, since this will lead to some narrow strata. This method is also not applicable when the smallest value in the stratification variable is zero.

For most populations, the KE method has produced CVs close to those obtained by the KO, GR and BR methods, which are the most efficient in terms of computing time. Large variation on computing time was observed between different methods. The KE method showed the worst results in this criterion, having displayed computing times much larger than those of the competing methods. The KO method, on the other hand, was the fastest in terms of computing time, while at the same time often achieving the best possible precision (lowest CV). The BR method showed computing time in between those of the KO and KE methods.

The graph in Figure 5.2 shows the percentages of times that each of the methods BR, KO, KE and GR produced the best solution, separated by number of strata. It shows a clear advantage of the BR method when compared to the KE and GR methods. When compared to KO, BR performed better for $H = 3$ and $H = 6,$ while KO was the winner for $H = 4$ and $H = 5.$ GR performed as well as KO for $H = 3$ and $H = 6,$ but was outperformed by both BR and KO for $H = 4$ and $H = 5.$ KE was the clear loser in this analysis, for any number of strata H.

We have also searched for associations between performance and other potential drivers, such as the skewness or the size $(N$ or $K)$ of the populations, but have not found any meaningful association within our limited set of populations.

Table 5.2
CVs for the estimator of total of the stratification variable by scenario

Populations	H	CV_{DH}	CV_{GH}	CV_{KO}	CV_{KE}	CV_{GR}	CV_{BR}
AgrMinas	3	4.158	7.187	4.050	4.089	4.050	4.050
	4	2.714	4.965	2.643	2.811	2.645	2.645
	5	2.325	3.828	1.945	2.262	1.945	1.945
	6	1.821	2.975	1.593	1.932	1.580	1.580
BeefFarms	3	2.758	2.491	1.875	2.086	1.875	1.875
	4	1.853	1.825	1.188	1.557	1.188	1.188
	5	1.455	1.369	0.902	1.280	0.902	0.902
	6	1.148	1.167	0.726	0.990	0.726	0.726
Beta103	3	0.561	0.810	0.560	0.560	0.559	0.559
	4	0.413	0.579	0.410	0.408	0.410	0.410
	5	0.337	0.500	0.329	0.329	0.329	0.329
	6	0.280	0.418	0.276	0.275	0.277	0.276
CensoCO	3	NA	4.839	4.334	4.336	4.334	4.334
	4	NA	4.388	3.078	3.062	3.078	3.078
	5	NA	NA	2.401	2.435	2.401	2.401
	6	NA	NA	1.949	1.956	1.943	1.943
Chi5	3	2.522	4.217	2.502	2.489	2.502	2.502
	4	1.897	3.199	1.889	1.881	1.889	1.889
	5	1.518	2.875	1.515	1.538	1.515	1.515
	6	1.258	NA	1.248	1.251	1.248	1.248
Coffe	3	10.049	12.598	6.906	6.876	6.906	6.906
	4	NA	10.450	4.996	5.027	4.996	4.996
	5	NA	8.124	3.877	3.939	3.877	3.877
	6	NA	6.756	3.176	3.477	3.176	3.176
Debtors	3	5.626	6.150	5.554	5.554	5.554	5.554
	4	4.098	4.387	4.049	4.049	4.049	4.049
	5	3.163	3.595	3.131	3.131	3.131	3.131
	6	2.639	2.897	2.562	2.562	2.562	2.562
HHinctot	3	3.206	5.106	3.184	3.184	3.184	3.184
	4	2.436	4.542	2.429	2.430	2.429	2.429
	5	1.993	4.225	1.973	1.979	1.973	1.973
	6	1.676	3.794	1.629	1.629	1.629	1.629
Iso2004	3	2.716	3.330	1.894	1.894	1.894	1.894
	4	2.059	2.154	1.206	1.206	1.207	1.207
	5	1.616	1.839	0.908	0.908	0.909	0.909
	6	1.380	NA	0.702	0.703	0.704	0.703
Kozak1	3	1.695	2.432	1.695	1.695	1.695	1.695
	4	1.305	2.020	1.301	1.301	1.301	1.301
	5	1.051	1.705	1.050	1.052	1.050	1.050
	6	0.904	1.402	0.890	0.917	0.890	0.890
Kozak3	3	3.673	5.049	3.663	3.659	3.663	3.663
	4	2.733	3.980	2.723	2.724	2.723	2.723
	5	2.208	3.199	2.178	2.231	2.178	2.178
	6	1.823	2.733	1.817	1.827	1.819	1.817
Kozak4	3	4.263	5.811	4.257	4.239	4.257	4.257
	4	3.219	4.696	3.204	3.193	3.205	3.204
	5	2.606	3.873	2.589	2.587	2.591	2.589
	6	2.168	3.236	2.155	2.155	2.157	2.158
ME84	3	1.703	2.527	1.296	1.296	1.296	1.296
	4	1.402	1.642	0.870	0.870	0.870	0.870
	5	1.050	1.549	0.661	0.661	0.661	0.661
	6	0.907	1.213	0.521	0.577	0.521	0.521

Table 5.2(continued)
CVs for the estimator of total of the stratification variable by scenario

Populations	H	CV_{DH}	CV_{GH}	CV_{KO}	CV_{KE}	CV_{GR}	CV_{BR}
MRTS	3	4.363	5.829	4.167	4.167	4.167	4.167
	4	3.406	5.259	2.960	2.960	2.961	2.960
	5	2.498	4.015	2.297	2.485	2.297	2.297
	6	2.167	3.445	1.836	1.836	1.838	1.836
P100e10	3	0.375	0.444	0.373	0.371	0.373	0.373
	4	0.295	0.346	0.294	0.294	0.294	0.294
	5	0.236	0.288	0.236	0.236	0.236	0.236
	6	0.198	0.242	0.196	0.198	0.196	0.196
P75	3	1.635	2.592	1.459	1.459	1.459	1.459
	4	1.415	1.798	0.966	0.966	0.966	0.966
	5	1.047	1.563	0.829	0.835	0.713	0.713
	6	0.896	1.250	0.769	0.553	0.552	0.552
pop1076	3	4.597	3.715	2.437	2.775	2.437	2.437
	4	NA	2.853	1.624	2.164	1.624	1.624
	5	NA	2.168	1.204	1.869	1.203	1.203
	6	NA	1.827	0.953	1.549	0.951	0.951
pop1616	3	4.989	4.318	3.898	3.921	3.898	3.898
	4	3.823	3.267	2.564	2.716	2.564	2.564
	5	3.187	2.508	1.882	2.183	1.882	1.882
	6	NA	2.050	1.527	1.962	1.496	1.496
pop2911	3	5.925	5.935	5.605	5.569	5.605	5.605
	4	4.070	3.992	3.807	3.807	3.807	3.807
	5	3.262	3.183	2.918	2.943	2.918	2.918
	6	2.632	2.649	2.281	2.418	2.281	2.281
Pop500	3	NA	0.678	0.092	0.127	0.092	0.092
	4	NA	0.178	0.059	0.082	0.060	0.060
	5	NA	0.194	0.043	0.059	0.045	0.046
	6	NA	0.117	0.033	0.046	0.036	0.037
Pop800	3	NA	3.133	1.555	2.448	1.555	1.555
	4	NA	2.755	0.996	1.511	0.996	0.996
	5	NA	1.620	0.701	1.261	0.702	0.702
	6	NA	1.436	0.546	0.823	0.550	0.548
REV84	3	1.901	2.777	1.614	1.776	1.614	1.614
	4	1.500	1.975	1.120	1.120	1.120	1.120
	5	1.235	1.700	0.835	0.836	0.835	0.835
	6	0.881	1.315	0.666	0.666	0.667	0.666
SugarCaneFarms	3	1.640	1.929	1.627	1.628	1.627	1.627
	4	1.152	1.440	1.118	1.122	1.118	1.118
	5	0.912	1.186	0.839	0.858	0.839	0.839
	6	0.707	1.041	0.691	0.732	0.682	0.682
Swiss	3	3.726	NA	3.682	3.683	3.690	3.682
	4	2.830	NA	2.781	2.781	2.787	2.781
	5	2.246	NA	2.227	2.549	2.232	2.228
	6	1.905	NA	1.860	1.880	1.864	1.860
USbanks	3	1.861	1.843	1.802	1.802	1.802	1.802
	4	1.364	1.417	1.270	1.270	1.270	1.270
	5	1.118	1.079	0.861	0.861	0.861	0.861
	6	0.794	0.850	0.718	0.710	0.710	0.710
UScities	3	2.738	2.705	2.655	2.687	2.655	2.655
	4	1.972	1.951	1.927	1.934	1.927	1.927
	5	1.483	1.451	1.436	1.437	1.436	1.436
	6	1.260	1.305	1.228	1.214	1.209	1.209
UScolleges	3	2.928	3.169	2.749	2.749	2.749	2.749
	4	2.106	2.185	2.018	2.018	2.018	2.018
	5	1.707	1.838	1.606	1.607	1.607	1.606
	6	1.486	1.488	1.323	1.323	1.323	1.323

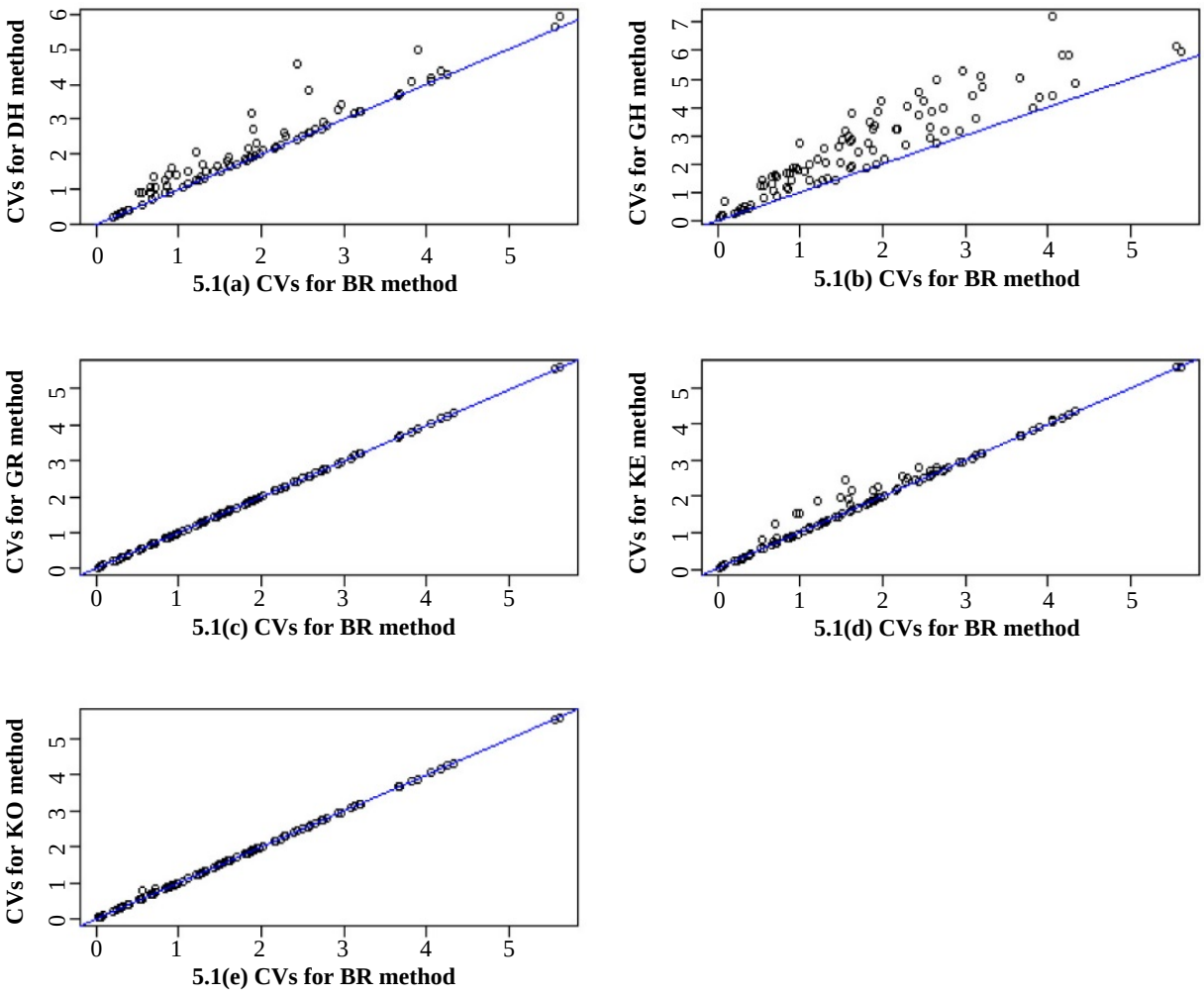

Figure 5.1 Comparing CVs of total estimators under alternative stratification methods, for all populations and numbers of strata (H).

Table 5.3
Percentage of times that method produced the best solution

Method	% Times best
DH	2.8
GH	0.0
KE	42.6
GR	71.3
KO	78.7
BR	78.7

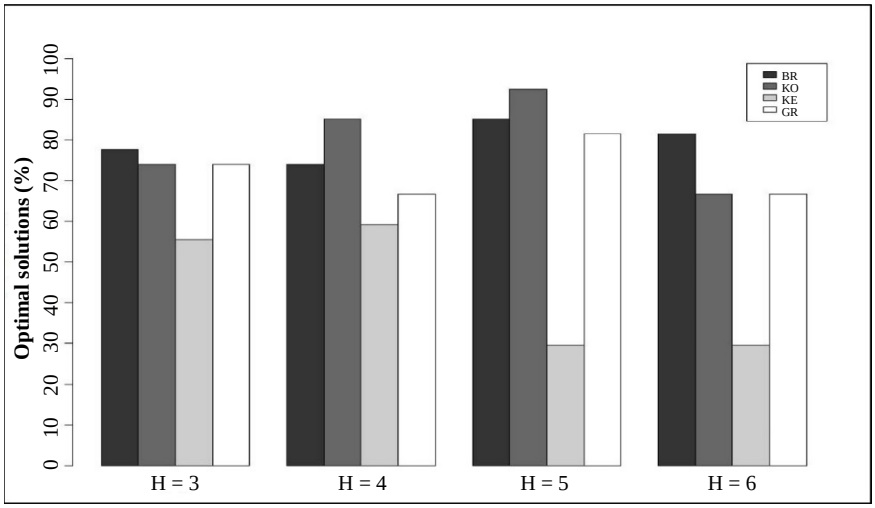

Figure 5.2 Percentage of best solutions yielded by method and number of strata (H).

6 Conclusions

As already mentioned, stratified sampling is a very important idea in survey sampling design, in helping to achieve improved precision of survey estimates for a given sample size or survey budget. This is particularly true for skewed or heterogeneous populations often found in business or establishment surveys. The potential gains due to stratification are strongly dependent on the delimitation of the strata and on the allocation of the sample to the strata, given a specified stratification variable and sample selection method.

The present paper presented a new optimization method for the stratification problem based on the Biased Random-Key Genetic Algorithm (BRKGA). Our approach (named BR) couples BRKGA for the definition of the stratum boundaries with the formulation for optimum sample allocation proposed in de Moura Brito et al. (2015), which is efficient with respect to computing time for large populations (large N).

The results reported here for the comparison of this approach with the five competitors considered suggest that BR offers a good alternative for addressing the stratification and allocation problems in a practical situation.

Our approach can be easily generalised to cases where the stratification variable x is not "measured", but instead represents a summary of several covariates in the form of a predicted y variable. The same is true for generalising to two or more numeric x variables, which can be easily accomplished by changing the decoding function used to retrieve feasible solutions from the BRKGA algorithm with the R package *stratbr* (see de Moura Brito et al., 2017a).

Future work will focus on developing and evaluating alternative decoding procedures which may be used in BR, aiming to produce solutions with superior quality when compared to those produced using the decoding procedure considered here. This research may also focus on solving the dual problem of minimising the total sample size for a specified precision, such as did Lavallée and Hidiroglou (1988).

Finally, additional empirical work may focus on considering varying sample sizes across the various study populations, as did Kozak (2004) and Gunning and Horgan (2004).

Appendix A

Table A1
Description of the 27 populations considered in the numerical experiment

Population	Description
AgrMinas	Agricultural production of municipalities in Minas Gerais State, Brazil, from 2006 Agricultural Census. Stratification variable: planted area.
BeefFarms	Australian beef farms, stratified into seven industrial regions, as considered by (Chambers and Dunstan, 1986). Stratification variable: farm size.
Beta103	Simulated population generated from a Beta distribution with parameters $a = 10$ and $b = 3$ as considered by (Keskintürk and Er, 2007).
Chi5	Simulated population generated from a Chi-square distribution with $df = 5$ as considered by (Keskintürk and Er, 2007).
Coffee	Coffee farms in the state of Paraná, Brazil, in the 1996 Agricultural Census, as considered by (de Moura Brito et al., 2015). Stratification variable: number of coffee trees.
CensoCO	Data from the 2012 school census in Brazil for the mid-west region. Stratification variable: number of classrooms.
Debtors	Population of debtors of an Irish firm as considered by (Er, 2011). Stratification variable: Irish debtors' stated liabilities.
HHinctot	Population of gross family income values (income before tax) from a Family Expenditure Survey 2001 carried out by Statistics Canada, as considered by (Er, 2011).
Iso2004	Data on net sales of 487 Turkish Industrial Enterprises out of the 500 largest enterprises in 2004, obtained by the Istanbul Industrial Chamber, as considered by (Keskintürk and Er, 2007). Stratification variable: net sales.
Kozak1, Kozak3, Kozak4	Populations considered by (Kozak and Verma, 2006). Stratification variable: were generated based on following formula: $X = \exp(Z)$, where Z is a realization of a normal random variable.
ME84	This data is from Särndal, Swensson and Wretman (1992) as considered by (Er, 2011). Stratification variable: number of municipal employees in 1984.
MRTS	Population simulated from the Monthly Survey on Sales in Retail Trade from Statistics Canada, as considered by (Er, 2011). Stratification variable: the size measure used for Canadian retailers in the Monthly Retail Trade Survey (MRTS) carried out by Statistics Canada. This size measure is created using a combination of independent survey data and three administrative variables from the corporation tax return.
P75	Population in thousands of the 284 Swedish municipalities in 1975, as considered by (Er, 2011). Stratification variable: population in thousands.
P100e10	Population simulated from a Normal distribution with $\mu = 100$ and $\sigma = 10$ as considered by (Keskintürk and Er, 2007).
pop1076	Population extracted from the Brazilian Annual Manufacturing Survey as considered by (de Moura Brito et al., 2017b). Stratification variable: number of employees.
pop1616	Population extracted from the Brazilian Annual Manufacturing Survey as considered by (de Moura Brito et al., 2017b). Stratification variable: number of employees.
pop2911	Population extracted from the Brazilian Annual Manufacturing Survey as considered by (de Moura Brito et al., 2017b). Stratification variable: number of employees.
Pop500	Population with $N = 500$ simulated from the Log-Normal Distribution $X = e^z$ where Z is Normal with $\mu = 4$ and $\sigma^2 = 2.7$ as considered by (Hedlin, 2000).
Pop800	Population with $N = 800$ simulated from the Log-Normal Distribution $X = e^z$ where Z is Normal with $\mu = 4$ and $\sigma^2 = 2.7$ as considered by (Hedlin, 2000).
REV84	Value of buildings in million Swedish Crown for the 284 Swedish municipalities in 1984, as considered by (Er, 2011). Stratification variable: the revenues from the 1985 municipal taxation.
SugarCaneFarms	Australian sugar cane farms as considered by (Chambers and Dunstan, 1986). Stratification variable: total cane harvested.
USbanks	Assets in millions of US Dollars for the large north American commercial banks, as considered by (Er, 2011). Stratification variable: the resources in millions of dollars of large commercial US banks.
UScities	Population in thousands for North American cities in 1940, as considered by (Er, 2011). Stratification variable: population in thousands.
UScolleges	Numbers of students in four-year US faculties in 1952-1953 as considered by (Er, 2011). Stratification variable: number of students.
Swiss	Data on Swiss municipalities in 2003, as available from the SamplingStrata package in R. Stratification variable: area under cultivation.

References

Baillargeon, S., and Rivest, L.-P. (2014). Stratification: Univariate stratification of survey populations. R package version 2.2-5. http://CRAN.R-project.org/package=stratification.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Chambers, R., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 3, 597-604.

Cochran, W. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.

Dalenius, T. (1951). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1-2, 133-148.

Dalenius, T., and Hodges, J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 285, 54, 88-101.

De Moura Brito, J.A.M., do Nascimento Silva, P.L. and da Veiga, T.M. (2017a). Stratbr: Optimal Stratification in Stratified Sampling. R package version 1.2. https://CRAN.R-project.org/package=stratbr.

De Moura Brito, J.A.M., do Nascimento Silva, P.L., Silva Semaan, G. and Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41, 2, 427-442. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14249-eng.pdf.

De Moura Brito, J.A.M., Maculan, N., Lila, M. and Montenegro, F. (2010b). An exact algorithm for the stratification problem with proportional allocation. *Optimization Letters*, 4, 185-195.

De Moura Brito, J.A.M., Ochi, L., Montenegro, F. and Maculan, N. (2010a). An iterative local search approach applied to the optimal stratification problem. *International Transactions in Operational Research*, 17, 6, 753-764.

De Moura Brito, J.A.M., Silva Semaan, G., Fadel, A. and Brito, L.R. (2017b). An optimization approach applied to the optimal stratification problem. *Communications in Statistics: Simulation and Computation*, 46, 4419-4451.

Ekman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 1, 219-229.

Er, S. (2011). Comparison of the efficiency of the various algorithms in stratified sampling when the initial solutions are determined with geometric method. *International Journal of Statistics and Applications*, 1, 1, 1-10.

Er, S., Keskintürk, T. and Daly, C. (2010). GA4Stratification: A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. R package version 1.0. http://CRAN.R-project.org/package=stratification.

Festa, P. (2013). A biased random-key genetic algorithm for data clustering. *SI:BIOCOMP, Math. Biosci.*, 245, 1, 76-85.

Gonçalves, J.F., and Resende, M.G.C. (2004). An evolutionary algorithm for manufacturing cell formation. *Comput. Ind. Eng*, 47, 247-273.

Gonçalves, J.F., and Resende, M. (2011). Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics*, 17, 487-525.

Gonçalves, J.F., Mendes, J.J.M. and Resende, M.G.C. (2005). A hybrid genetic algorithm for the job shop scheduling problem. *Eur. J. Oper. Res*, 167, 77-95.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004002/article/7749-eng.pdf.

Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 1, 40, 27-31.

Hidiroglou, M.A., and Kozak, M. (2017). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*, https://doi.org/10.1111/insr.12230.

Keskintürk, T., and Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, 52, 53-67.

Khan, M.G.M., Nand, N. and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34, 2, 205-214. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10761-eng.pdf.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 5, 797-806.

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

Kozak, M. (2014). Comparison of random search method and genetic algorithm for stratification. *Communications in Statistics – Simulation and Computation*, 43, 2, 249-253.

Kozak, M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, 32, 2, 157-163. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9550-eng.pdf.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 1, 33-43. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1988001/article/14602-eng.pdf.

Lohr, S. (2010). *Sampling: Design and Analysis*, 2nd Ed. Washington: Duxbury Press.

Oliveira, R.M., Chaves, A.A. and Lorena, L.A.N. (2017). A comparison of two hybrid methods for constrained clustering problems. *Applied Soft Computing*, 54, 256-266.

Rao, D.K., Khan, M.G.M. and Reddy, K.G. (2014). Optimum stratification of a skewed population. *International Journal of Mathematical, Computational, Physical and Quantum Engineering*, 8, 3, 497-500.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 2, 191-198. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002002/article/6432-eng.pdf.

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

Spears, W., and De Jong, K. (1991). On the virtues of parameterized uniform crossover. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, 230-236.