RAIRO-Oper. Res. 55 (2021) 979–996
https://doi.org/10.1051/ro/2021051

RAIRO Operations Research
www.rairo-ro.org

HEURISTIC APPROACH APPLIED TO THE OPTIMUM STRATIFICATION PROBLEM

José André Brito[1], Leonardo de Lima[2,*], Pedro Henrique González[3], Breno Oliveira[3] and Nelson Maculan[4]

Abstract. The problem of finding an optimal sample stratification has been extensively studied in the literature. In this paper, we propose a heuristic optimization method for solving the univariate optimum stratification problem to minimize the sample size for a given precision level. The method is based on the variable neighborhood search metaheuristic, which was combined with an exact method. Numerical experiments were performed over a dataset of 24 instances, and the results of the proposed algorithm were compared with two very well-known methods from the literature. Our results outperformed 94% of the considered cases. Besides, we developed an enumeration algorithm to find the optimal global solution in some populations and scenarios, which enabled us to validate our metaheuristic method. Furthermore, we find that our algorithm obtained the optimal global solutions for the vast majority of the cases.

Mathematics Subject Classification. 90C59, 62D05.

Received August 5, 2020. Accepted March 30, 2021.

1. Introduction

The optimum stratification problem, related to the field of probability sampling [8], can be formulated according to two possible goals: (A) minimizing the variance of an estimator given a fixed sample size or (B) minimizing the sample size for a fixed level of precision. In the literature, most methods were developed aiming at the first goal [2, 3, 5, 6, 10, 15, 18, 22, 25, 26, 30, 37–39], while the second goal has been less studied [23, 24, 28, 33, 35].

This article's optimization problem consists of minimizing the total sample size, simultaneously satisfying the constraints of precision and minimum sample size of each stratum. In this paper, we proposed an optimization approach aiming to give good solutions to the problem associated with the goal (B). It is worth mentioning that the stratification methods proposed in the literature do not take into account the constraint of a minimum sample size per stratum and do not solve problems with negative entries. Our method fills this gap since this

Keywords. Sampling, stratification, VNDS, exact methods.

[1] National School of Statistical Sciences, Rio de Janeiro, Brazil.
[2] Federal University of Paraná, Curitiba, Brazil.
[3] Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil.
[4] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
*Corresponding author: leonardo.delima@ufpr.br

is very important for the real-life situations of some sampling research in official statistics agencies, such as the Brazilian Institute of Geography and Statistics (IBGE).

Our approach is developed into two steps: (i) definition of each stratum by obtaining the cutoff points using the Variable Neighborhood Search (VNS), by Hansen *et al.* [21]; (ii) sample allocation is obtained optimally by solving an Integer Programming formulation given by Brito *et al.* [4]. The proposed method was implemented in *R* language, and it is available in the CRAN at `https://cran.r-project.org/web/packages/stratvns/`.

We applied the proposed algorithm to a dataset of 24 instances (populations) with size (N) ranging from 284 to over 16 057. Most instances are very well-known and are available in *R* on the packages *stratification, GA4Stratification* and *sampling*. To evaluate the proposed algorithm's performance against the classical algorithms of Hidiroglou and Kozak, computational experiments were carried out with the 24 instances, considering eight associated scenarios: the number of strata ($L = 3, 4, 5, 6$) *versus* two levels of precision, $cv = 5\%$ and $cv = 10\%$. Considering these scenarios, the proposed metaheuristic outperformed the classical methods in all cases but one, in which the result was the same as theirs. Hence, our method produced the smallest sample sizes respecting the level of precision constraint. Also, we implemented an exhaustive method called *StratEnum* which was applied to some of the 24 instances and compared the metaheuristic results with those of the *StratEnum*. The results show that the proposed algorithm obtained the global optimum in 98.64% of the cases.

The remainder of this paper is organized as follows. In Section 2, we present the basic concepts of sampling, considering, in particular, the concepts of stratified sampling. In Section 3, we present a detailed description of the optimal stratification problem. In Section 4, the proposed metaheuristic method and an enumeration method are presented. In Section 5, we apply our methodology to a dataset from the literature and compare our results to those obtained by the classical methods proposed by Kozak [28] and Lavallée and Hidiroglou [33]. Finally, in Section 6, this study's conclusions and possible extensions are presented.

2. Sampling concepts and stratified sampling

In the light of society and government demands, the institutes and bureaus that produce official statistics have been carrying out various surveys that aim to collect information about the characteristics of interest of different types of population (people, households, companies, schools, farms, among others), for which there is a need to produce a set of statistics. From these statistics, for example, governments can plan and implement their economic and social policies. Three examples are the demographic census carried out, in general, every ten years, the survey of the domestic product (GDP), and the surveys associated with price indices.

Given geographic, logistical, or cost issues, these surveys are mostly carried out by sampling, that is, instead of performing out a census of the entire population, a survey is conducted based on a subset of selected units of the population called sample. According to [8], some of the advantages of using sampling instead of the complete enumeration of the population have cost reduction, time reduction, a more comprehensive data collection, and higher accuracy in the collection of information. Most surveys carried out by statistical institutes use probability sampling. More specifically, in this type of sampling, the population of interest elements have a greater than zero probability of being selected (from a register) to compose the sample, considering adopting a sampling plan.

The most common examples of sampling plans [36] are simple random sampling, stratified sampling, systematic sampling, and cluster sampling. Still, in this sense, these researches are based on adopting a complex sampling plan, that is, one that can combine two or more sampling schemes, particularly stratified sampling, which is intrinsically associated with the problem studied in this article. Therefore, to facilitate the understanding of the stratification problem addressed in this research, we present below (based on [8, 36]) the notations, definitions, and expressions associated with stratified sampling. The reasons to use stratification are: improving the accuracy of estimates, the possibility of representing different groups within a population, ensuring the spread of the sample, and administrative issues.

The use of stratified sampling implies partitioning a population U of N units into L subpopulations consisting of $N_1, N_2, \ldots, N_L$ units. Such subpopulations are called population strata and denoted by $E_1, E_2, \ldots, E_L$. The following constraints are respected for the construction of the strata:

$$U = E_1 \cup E_2 \cup \ldots \cup E_L, \tag{2.1}$$

$$E_j \cap E_k = \emptyset, \qquad \forall j,k \in \{1,\ldots,L\}, j \neq k, \tag{2.2}$$

$$N = \sum_{h=1}^{L} N_h.$$

Once these strata are determined, a simple random sample denoted by s_h is selected in each stratum E_h,($h = 1,\ldots,L$), with selections made independently in the different strata, so that the total sample size is given by the sum of the samples allocated to the strata. Each sample s_h has an associated size denoted by n_h, with the total sample size (n) being defined by the sum of the sample sizes of each stratum:

$$n = \sum_{h=1}^{L} n_h. \tag{2.3}$$

To produce the statistic associated with a variable Y of interest, information (such as age and income, among others) is collected for all investigated sample elements, which is selected from the L strata. For example, the expression of the total estimator $\hat{Y}_{\mathrm{AE}}$, is defined according to the following equation:

$$\hat{Y}_{\mathrm{AE}} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i \in s_h} y_{hi},$$

where y_{hi} value of ith unit in stratum h.

Additionally, in order to calculate the value of the variance of the total estimator associated with $\hat{Y}_{\mathrm{AE}}$ in equation (2.6), are defined, respectively, in equations (2.4) and (2.5) define the mean and the population variance for each stratum, respectively:

$$\overline{Y}_h = \frac{1}{N_h} \sum_{i \in E_h} y_i, \tag{2.4}$$

$$S_{hy}^2 = \frac{1}{N_h - 1} \sum_{i \in E_h} (y_i - \overline{Y}_h)^2. \tag{2.5}$$

Finally, we have that the variance of the total estimator ($\hat{Y}_{\mathrm{AE}}$) and its associated coefficient of variation are respectively given by:

$$V(\hat{Y}_{\mathrm{AE}}) = \sum_{h=1}^{L} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{hy}^2}{n_h} \tag{2.6}$$

and

$$cv(\hat{Y}_{\mathrm{AE}}) = \frac{\sqrt{V(\hat{Y}_{\mathrm{AE}})}}{T_Y}, \tag{2.7}$$

where T_Y (population total) is defined by

$$T_Y = \sum_{i=1}^{N} y_i.$$

The expression presented in equation (2.6) or (2.7) allows us to evaluate how accurate is the result obtained from an estimate derived from one of the variables considered in the research. The lower the value of $V(\hat{Y}_{\mathrm{AE}})$ or $cv(\hat{Y}_{\mathrm{AE}})$, the better the stratification.

3. Optimal stratification problem

Suppose that information must be collected for a population U composed of N elements distributed in a set $P = \{1, 2, ..., N\}$. A sample is drawn from this population to gather information for a set of variables of interest. One of these variables is defined by $Y = (y_1, y_2, \ldots, y_N)$. The goal is to estimate variable Y by using $\hat{Y}_{\text{AE}}$. Moreover, a variable of size X is considered and used for the stratification of P. The values of X are known for each population unit, *i.e.*, $X = (x_1, x_2, \ldots, x_N)$.

In order to stratify P, the observations of X are distributed in a non-decreasing order in each stratum $E_h, h = 1, \ldots, L$, which are constructed as a function of $(L-1)$ strata boundaries, denoted by $b_1, b_2, \ldots, b_{L-1}$ such that $b_1 < b_2 < \cdots < b_{L-1}$, as follows:

$$E_1 = \{x_j \in X | x_j \leq b_1\}, \tag{3.1}$$

$$E_h = \{x_j \in X | b_{h-1} < x_j \leq b_h\}, \text{ for each } h = 2, \ldots, L-1, \tag{3.2}$$

$$E_L = \{x_j \in X | b_{L-1} < x_j\}. \tag{3.3}$$

After constructing each stratum, in the case of simple stratified sampling, a random sample of size $n_h (h = 1, \ldots, L)$ is drawn from each stratum in such a way that $n_1 + \cdots + n_L = n$, that is, equation (2.3) is satisfied. Based on this information, the stratification problem will be solved by determining the strata boundaries $b_1 < b_2 < \cdots < b_{L-1}$ in such a way that the variance of the estimator of the total of the variable Y is minimum.

Since, in general, the values of Y are not known for the entire population, the variance presented in equation (2.6) cannot be calculated. A typical procedure to solve this problem is to replace Y with X in the variance equation, considering the correlation between both variables. As a result, strata boundaries and the variance equation are given as a function of X. Many authors start from this assumption, such as: Dalenius and Hodges [10], Lavallée and Hidiroglou [33], and Hedlin [22]. In this phase, the importance of selecting a proper auxiliary variable or stratification variable becomes evident. Its relation to the variable of interest may be analyzed through previous studies or surveys or by carrying out a pilot survey. Once the replacement is done, the following variance equation must be minimized:

$$V(\hat{X}_{\text{AE}}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{hx}^2}{n_h}, \tag{3.4}$$

$$cv(\hat{X}_{\text{AE}}) = \frac{\sqrt{V(\hat{X}_{\text{AE}})}}{T_X}, \tag{3.5}$$

where T_X (population total) is defined by

$$T_X = \sum_{i=1}^{N} x_i.$$

Table A.1 in the appendix section presents a summary of the notation defined in this paper.

Notice that in order to compute $V(\hat{X}_{\text{AE}})$ or $cv(\hat{X}_{\text{AE}})$, a two-level problem needs to be solved: (1) to determine the population strata, which consequently makes us obtain the values of N_h and S_{hx}^2; (2) to define the sample sizes n_h that will be allocated to these strata.

The resolution of this problem, in its two levels, defines the optimal stratification problem. Concerning the first level, the solution is associated with delimiting strata, that is, defining cutoff points that allow segmenting the population into L strata. At the second level, already considering the defined strata, we have the problem of optimal allocation, which can be addressed according to one of the following objectives:

(i) Determine the sample sizes n_h so that the sum $\sum_{h=1}^{L} n_h$ is minimal, subject to the "level of precision constraint", defined *a priori* as $cv(\hat{X}_{\text{AE}}) \leq cv_t$, where cv_t is a target coefficient of variation (fixed precision);

(ii) Minimize $V(\hat{X}_{\text{AE}})$, subject to $\sum_{h=1}^{L} n_h = n$, where n is defined *a priori*.

In the first objective, the goal is the cost reduction associated with the sample size, and in the second objective, the focus is on obtaining maximum precision.

In general, the steps that are considered in order to solve the optimal stratification problem can be summarized as follows:

(1) *Define the objective function to be optimized;*
(2) *Define the allocation method;*
(3) *Choose the variable to be stratified and define the number of strata* (L)*;*
(4) *Define the number* $L-1$ *of cutoff points;*
(5) Compute the sample size of each stratum, which we denote by $n_1, n_2, \ldots, n_L$, according to the allocation method defined in item 2;
(6) Select the objects in each stratum according to the selection method defined in item 2, and with the size n_h of each stratum $h = 1, \ldots, L$.

As described in this section, Step 1 consists of defining of defining the objective function by either: (i) minimizing the total estimator variance, that is, maximizing the precision considering a fixed sample size; (ii) minimizing the sample size, that is, minimizing cost given a fixed precision. Most of the proposed methods in the literature are related to the objective function described in (i), as can be seen in [2, 3, 5, 6, 10–16,18,22,25,26,37–40]. The objective function described in (ii) was only studied by Hidiroglou [23], Lavallée and Hidiroglou [33], Kozak [28], Hidiroglou and Kozak [24] and Lisic *et al.* [35]. Note that both objective functions are correlated since minimizing variance requires the sample size as an input, and minimizing the sample size requires precision (that is, variance or variation coefficient) as an input to the problem.

In Step 2, the sample allocation method is chosen among the Uniform, Proportional, Neyman, and Power methods (see [5]). Notice that none of those allocation methods provides an integer sample size. Step 4 consists of choosing $L-1$ cutoff points denoted by $b_1, \ldots, b_{L-1}$, considering the variable to be stratified. This step divides the population into L strata. Once all cutoff points boundary is defined, it is possible to obtain the subpopulation size N_h and the corresponding variance S^2_{hx} of each stratum h. According to [5], finding a global minimum for this very important problem is a difficult analytically and computationally, since S^2_{hx} is a nonlinear function of the values $b_1, b_2, \ldots, b_{L-1}$. Step 5 is the second level of the stratification problem, where the sample size n_h of each stratum is defined in such a way that n, the total sample size which optimizes the objective function, is obtained. As the second level of the problem has already been solved optimally by Brito *et al.* [4], the method proposed in this work uses the allocation presented in that research and focuses only on the first level of the stratification problem in order to solve the second objective (to minimize the sample size).

To illustrate the stratification problem, we present an example based on a fictitious population of size $N = 18$, whose observations associated with the stratification variable X are defined by

$$X = \{1, 1, 2, 2, 3, 3, 4, 4, 5, 7, 7.8, 8, 10, 10, 15, 31\}.$$

We consider the number of strata as $L = 2$, and the target coefficient of variation as $cv_t = 0.02$. The first step towards stratification of a population concerns to the choice of the cutoff points. Here, since $L = 2$, it implies the determination of only one cutoff point denoted by b_1. Once b_1 is obtained, stratum E_1 and E_2 are defined, which allows the determination of population sizes N_1 and N_2 in the strata and their respective population variances S^2_{1x} and S^2_{2x}. An allocation method is applied, and the sample sizes n_1 and n_2 associated with the strata are obtained. Consequently, the value of n (total sample) that corresponds to the objective function of the problem is determined. In Table 1 we consider two choices to b_1, and their impacts to the minimization process. First, we take $b_1 = 4$ which is associated to $n = 7$. Then, by choosing $b_1 = 8$, we get $n = 6$. It is observed that the choice of a cutoff point is determinant for the sample size and the coefficient of variation. In this case, when $n = 7$ and $n = 6$, the following values for $cv(\hat{X}_{\mathrm{AE}})$, 18.34% and 17.95% were observed, respectively, both satisfying the level of precision constraint.

It is worth mentioning that we did not find any article with a method that guarantees the global optimum achievement for this minimization problem at the first level, except for the exhaustive enumeration method,

Table 1. Example of optimizing the stratification problem.

b_1	E_1	E_2	N_1	N_2	S_{1x}^2	S_{2x}^2	n_1	n_2	**n**
4	1 1 1 2 2 3 3 4 4	5 7 7 8 8 10 10 15 31	9	9	1,5	62,9	2	5	7
8	1 1 1 2 2 3 3 4 4 5 7 7 8 8	10 10 15 31	14	4	6,8	99,0	3	3	6

which is described in Section 4.2. However, applying enumeration is infeasible for medium or large-sized populations (depending on N and L). Therefore, metaheuristic methods are proper for larger instances, which justifies the proposal of this work.

In this paper, we aim to minimize the sample size by using a metaheuristic procedure. Our approach uses the Variable Neighborhood Search (VNS) metaheuristic to define good cutoff points for the cutoff sampling problem. We use the exact method proposed by Brito *et al.* [4] to allocate samples in each stratum with those points. Therefore, the first level of the problem is solved using a metaheuristic method, and the second level is by an exact method.

4. Developed methods

We developed a metaheuristic method for the first level of the optimal stratification problem. In this sense, we implemented the Variable Neighborhood Search (VNS), proposed by Hansen and Mladenović [20]. Each solution, generated by the optimization process of the VNS, is given as an input to the mathematical formulation proposed by Brito *et al.* [4], which optimally solves the second level of the problem. The integration of both levels' solution is done in Algorithm 1, proposed in this article, and named *StratVNS*. This algorithm was implemented in *R* and is available in the *stratvns*, `https://cran.r-project.org/web/packages/stratvns/`.

In this context, a solution is feasible whenever it attains all of the following conditions for each stratum: the sampling size is at least $n_{\min}$, that is, $n_h \geq n_{\min}$; the number of population elements is at least $N_{\min}$, that is, $N_h \geq N_{\min}$; the coefficient of variation is is less than or equal to a target coefficient of variation cv_t, that is, $cv \leq cv_t$, see equation (3.5). Feasibility is checked for every intermediate solution of the algorithm (values of N_h and S_{hx}^2).

In line 2 of Algorithm 1, the duplications of population X are removed, producing the set Q whose values will be used as possible cutoff points by the algorithm to determine the values of N_h and S_{hx}^2. In line 3, the procedure *InitialSolution()* generates 20 random solutions and returns the best feasible solution associated with the smallest sample size. From lines 5 to 11, we execute a multi-start of the VNS metaheuristic, and the algorithm stops when either the maximum CPU time or the maximum number of iterations is achieved. In line 7, the evaluation of objective function is done for the solution b' by calling the integer formulation of [4], which optimally solves the second level of the stratification problem returning the sample size. Notice that the same happens inside the VNS procedure, which is presented in the next subsection.

In Table 2, we give a brief description of the functions of Algorithm 1.

4.1. Variable Neighborhood Search procedure

In Algorithm 2, the proposed Variable Neighborhood Search (VNS) is presented. Each solution is represented by a vector of cutting points $b = (b_1, b_2, \ldots, b_{L-1})$. A set of neighborhood structures for a given integers k and s, denoted by $\mathrm{NG}_{k,s}(x)$, is defined as follows: given a solution x, a solution $x' \in \mathrm{NG}_{k,s}(x)$ has k (such that $1 \leq k \leq L-2$) of its elements randomly chosen to be modified, and the remaining $(L-1-k)$ elements are fixed. Assume the k chosen elements are $x_1, \ldots, x_k$ and let t_j be the position of x_j in set Q for each $j = 1, \ldots, k$. Then, each new element x'_j replacing x_j will be chosen from the interval $q_{t_j-s} \leq x_j \leq q_{t_j+s}$. Note that the neighborhood structures $\mathrm{NG}_{k,s}(x)$ are nested.

Algorithm 1: StratVNS Algorithm.

Input: $X, L, cv_t, n_{\min}, N_{\min}, k_{\max}, i_{\max}, CPUtimeMax$
$Q = \text{RemoveDuplications}\,(X)$;
$b = \text{InitialSolution}\,(Q, L, n_{\min}, N_{\min}, cv_t)$;
$i = 1$;
while $(i \leq i_{\max})$ *and* $(CPUtime \leq CPUtimeMax)$ **do**
  $b' = \text{VNS}(Q, X, b, k_{\max}, n_{\min}, N_{\min}, cv_t)$;
  if *EvalObjFunc*$(b', cv_t) \leq$ *EvalObjFunc*(b, cv_t) **then**
    $b = b'$;
  end
  $i = i + 1$;
end
Output: b, n_h, N_h, S^2_{hx}.

Table 2. Description of the functions of Algorithm 1.

Routines	Description
RemoveDuplications(X)	Remove all data duplications from the X vector
EvalObjFunc(b, cv_t)	Returns the optimal sample size considering the cutoff points of b by calling the integer programming formulation of [4]
InitialSolution(Q, L, $n_{\min}$, $N_{\min}$, cv_t)	Twenty random solutions are generated. The best feasible solution (associated with the smallest sample size) is chosen

With the *Shaking*(b, k, s) procedure, solution b' is obtained by perturbing solution b, considering the neighborhood $\mathbf{NG}_{k,s}(b)$. The shaking procedure corresponds to randomly choosing a neighbor from the set $\mathbf{NG}_{k,s}(b)$, which will be used as an input to the local search. This new solution b' differs from b by k elements after randomly choosing a neighbor in $\mathbf{NG}_{k,s}(b)$ and is given as an input to the local search.

Due to the high computational cost of obtaining all solutions in a given neighborhood, we use the reduced VNS (RVNS) procedure. The RVNS method consists of obtaining $t_{\max}$ random solutions selected from $\mathbf{NG}_{k,s'}$ where no descent is required. The random solutions are compared with the current solution, and an update takes place in case of improvement. The application of an exhaustive local search procedure or, even of first improvement, is not feasible in this case, since each evaluation of the objective function implies solving an integer programming problem. Then, the RVNS method generates $t_{\max}$ random solutions from $\mathbf{NG}_{k,s'}(b')$ and returns the best solution (smallest sample size), say b'', among all $t_{\max}$ solutions. There is no guarantee that b'' is a local optimum. The neighborhood structures of the *Shaking* and RVNS procedures are the same, except for the range of the modification. Algorithm 2 stops when the maximum number of iterations with no improvements (reduction of the sample size) is achieved.

For the sake of clarity, consider the following example. Suppose that the stratification variable of the population of interest has the following values $X = \{1, 1, 1, 2, 2, 3, 3, 4, 4, 5, 7, 7, 8, 8, 10, 10, 15, 31\}$, and the number of strata is equal to $L = 3$, and $cv_t = 10\%$. In Algorithm 1, after applying RemoveDuplications(X), we obtain the set $Q = \{1, 2, 3, 4, 5, 7, 8, 10, 15, 31\}$. In the sequel, the *InitialSolution()* provides the initial solution $b = (5, 8)$ corresponding to the cutoff points, which is given that is given to the VNS procedure as an input. In Algorithm 3, let $k = 1$ and $s = s' = 2$. Assume that the element $b_2 = 8$ was randomly chosen to be modified in the *Shaking()* procedure. In this case, the new cutoff point b'_2 is randomly chosen from the set $\{5, 7, 10, 15\}$. Suppose that $b'_2 = 10$, which implies that $b' = (5, 10)$. Let $t_{\max} = 4$, and suppose that b'_1 is chosen from the set $\{3, 4, 7, 8\}$ inside the RVNS() procedure. In this case, the set of possible solutions is $\mathcal{S} = \{(3, 10), (4, 10), (7, 10), (8, 10)\}$.

The feasibility is checked and the objective function is evaluated for every solution in $\mathcal{S}$. After that, we obtain $b'' = (4, 10)$ as the feasible solution with minimum sample size. In Line 9, the EvalObjFunc() is computed for $b = (5, 8)$ and $b'' = (4, 10)$, which means that the second level is optimally solved for both solutions, and the one with minimum sample size is kept as the best solution. In this case, *EvalObjFunc((5,8), 0.01)* $= 8$ and *EValObjFunc((4,10), 0.01)* $= 6$, the current solution is updated, k is incremented to 2, and the algorithm continues.

Algorithm 2: VNS general framework.

Input: Currently solution b. Define integers s, $s', nIterWithNoImpMax$ and $t_{\max}$. Define the set of neighborhood structures $\mathbf{NG}_{k,s}$ and $\mathbf{NG}_{k,s'}$ for $k = 1, \ldots, k_{\max}$

2 nIterWithNoImp = 0
3 **while** *(nIterWithNoImp < nIterWithNoImpMax)* **do**
4     $k = 1$
5   nIterWithNoImp = nIterWithNoImp + 1
6   **while** $k \leq k_{\max}$ **do**
7         $b' = Shaking(b, k, s)$;
8         $b'' = \text{RVNS}(b', k, t_{\max}, s')$;
9   **if** $EvalObjFunc(b'', cv_t) < EvalObjFunc(b, cv_t)$ **then**
10             $b = b''$;
11             $k = 1$;
12   nIterWithNoImp = 0
13   **else**
14             $k = k + 1$;
15   **end**
16   **end**
17 **end**

Output: b

4.2. Enumeration method

Alternatively, to apply the VNS algorithm, one can consider using the exhaustive enumeration algorithm described in Algorithm 3. This algorithm guarantees the global optimum for the stratification problem in its two levels: the cutoff points and the allocation of the sample to the strata. However, its use is only feasible under certain conditions that we will explain further on. This algorithm was developed based on the discretization considered in Section 4 for applying the VNS method, that is, the elements of the Q set. The algorithm generates all possible stratifications of a population. It is based on the determination of all integer and non-negative solutions of the following linear equation:

$$w_1 + \cdots + w_h + \cdots + w_L = |Q|, \tag{4.1}$$

$$w_h \geq 2, h = 1, \ldots, L, \tag{4.2}$$

where each w_h corresponds to the number of observations of Q that are in the hth strata. Notice that equation (4.2) implies $N_h \geq 2, h = 1, \ldots, L$.

For each $(w_1, \ldots, w_L)$ solution that satisfies equations (4.1) and (4.2), we have corresponding values of N_h and S^2_{hx}. These values, together with the coefficient of variation fixed *a priori* (cv_t), are used as input for the formulation proposed by Brito *et al.* [4]. To meet the constraint associated with equation (4.2), the following substitution can be made in equation (4.1): $w_h = z_h + 2, (h = 1, \ldots, L)$, producing:

$$z_1 + \cdots + z_h + \cdots + z_L = |Q| - 2L. \tag{4.3}$$

The total T of entire solutions of equation (4.3) (see [42]), corresponding to the total number of feasible solutions S for the stratification problem (possible stratum sizes) at the first level, can be defined by:

$$T = \frac{(|Q| - L - 1)!}{(|Q| - 2L)!(L - 1)!}. \tag{4.4}$$

For example, assuming $|Q| = 200$ and $L = 3$, we have $T = \frac{196!}{194!2!} = 19\,110$ feasible solutions that must be evaluated by applying the formulation of [4]. Keeping $|Q|$ fixed and increasing L by one, $T = 1\,216\,185$, indicating a substantial number of solutions that must be listed for equation (4.4). Tests performed with this algorithm showed that its application that its application is computationally feasible when $T \leq 10^7$. Algorithm 3 is the enumeration algorithm, denoted by StratENUM.

Algorithm 3: StratENUM Algorithm.

Input: Q, L, cv_t
1 Determine all solutions of equation (4.3) and keep them in matrix Z of dimension $T \times L$;
2 Compute N_h and S^2_{hx} for each solution obtained in line 1 and keep them in matrix V of dimension $T \times 2L$ (values of N_h and S^2_{hx}) ;
3 Considering the parameter cv_t, apply the optimal allocation according to [4] to each row of matrix V in order to obtain the sample size associated with each of them;
4 Select the row of the matrix V whose N_h and S^2_{hx} are associated with the smallest sample size obtained in line 3;
Output: $N_h, S^2_{hx}, n_h, n, cv$

5. Computational experiments

In this section, computational experiments for the StratVNS and StratENUM are presented. The best two known algorithms in the literature, LH88 [33] and Ko04 [28], were used as a baseline to evaluate the efficiency of the proposed algorithm. The functions associated with the LH88 and Ko04 algorithms are implemented in the stratification package (strata.LH function with default arguments) of the *R* Language, and functions associated with StratVNS and StratENUM are implemented in the stratvns package. All routines used in STRATVNS are available at `http://github.com/pehgonzalez/stratification`. In the Appendix A, we show an example using all of these functions.

All numerical experiments were performed on a computer with an AMD-FX6300 six-core processor, with a 3.5 GHz CPU and 16 GB of RAM. To evaluate the proposed methods, the StratVNS algorithm was applied to 24 benchmark public datasets from the literature. The StratENUM algorithm was applied to all instances, such that $T \leq 10^7$, for different values of L. We have implemented all algorithms described in the previous sections in *R* language.

The remainder of this section is organized as follows: Section 5.1 presents the characteristics of each benchmark instance, and in Section 5.2, the results obtained in our computational experiments are presented.

5.1. Instances

Twenty four instances were used to test the proposed algorithms. The instances either come from statistical packages or are generated from statistical distributions. For example, consider the populations used in the following works: [5, 18, 22, 25]. We have used instances available in the statistical software *R* in the following packages (`http://cran.r-project.org/web/packages/available_packages_by_name.html`): *stratification*, *GA4Stratification* ([25]) and *sampling*. Also, some instances from different authors were used as presented in Table 3, which also shows the identification code, name, source, and description of each instance. The descriptive statistics of these 24 instances are summarized in Table 4. The first column presents the identification code of

TABLE 3. Description of the 24 populations from literature.

ID	Name	Reference	Description
U01	BeeFarms	[7]	Australian cattle farms stratified by industrial regions
U02	beta103	GA4Stratification	Population generated from Beta distribution with parameters a = 10 and b = 3
U03	chi1	GA4Stratification	Population generated by the Chi-Square distribution with 1 degree of freedom
U04	chi5	GA4Stratification	Population generated by the Chi-Square distribution with 5 degrees of freedom
U05	debtors	Stratification	Debtor population in an Irish firm
U06	HHINCTOT	Stratification	Canada 2001 family income before taxes
U07	iso2004	GA4Stratification	Net sales of Turkish industrial companies in 2004. Population divided by 1000
U08, U09, U10, U11	Kozak1, ..., Kozak4	[31]	Populations given in the article by Kozak and Verma
U12	me84	Sampling	Number of municipal employees in 284 municipalities in Sweden in 1984
U13	mrts	Stratification	Simulated population of the Monthly Wholesale Trade Survey from Statistics Canada
U14	p100e10	GA4Stratification	Population generated by the Normal distribution with $\mu = 100$ and $\sigma = 10$
U15	p75	GA4Stratification	Population in thousands of 284 municipalities in Sweden in 1975
U16	pop800	[22]	Generated from the Log-Normal distribution ($X = e^Z$), where Z follows an N ($\mu = 4; \sigma^2 = 2.7$)
U17	rev84	Sampling	Property values in millions of Swedish kronor from 284 municipalities in 1984
U18	SugarCaneFarms	[7]	Australia's sugar cane farm population
U19	Swiss	Sampling	Information on Swiss municipalities (2003)
U20	TaxableIncome	Sampling	Income of municipalities in Belgium in 2001 (in euros, divided by 1000)
U21	Usbanks	Stratification	Million dollar funds from major US commercial banks
U22	Uscities	Stratification	Population in thousands of American cities in 1940
U23	Uscolleges	Stratification	Number of students at four-year US colleges in 1952–1953
U24	Rchisq2-30	[1]	Population generated by the Chi-Square distribution with 30 degrees of freedom

the population. The second column presents the population size (N). The third column shows the number of distinct values: the cardinality of set Q, denoted by $|Q|$. The fourth and fifth columns present the minimum and maximum values of the stratification variable X, corresponding to the population. The last column shows the coefficient of skewness. It is worth highlighting three points: There are only two populations of a size larger than or equal to $N = 10\,000$, the population U14 has zero skewness, and the population U16 has the largest positive skewness 22.2.

TABLE 4. Basic information of the 24 populations from literature.

ID	N	$\|Q\|$	Minimum	Maximum	Skewness
U01	430	353	50	24 250	4.6
U02	1000	1000	358	986	−0.7
U03	1000	1000	0	13	2.7
U04	1000	1000	0.1	23.4	1.4
U05	3369	1129	40	28 000	6.4
U06	16 057	225	0	690 000	2.7
U07	487	487	63 583	10 446 592	10.1
U08	4000	51	3	72	1.4
U09	4000	2837	243	28 578	2.7
U10	2000	581	6	2793	3.5
U11	10 000	5453	62	74 398	4.2
U12	284	264	173	47 074	8.7
U13	2000	2000	141	486 367	8.6
U14	1000	1000	74	127.3	0.0
U15	284	68	4	671	8.5
U16	800	402	1	473 510	22.2
U17	284	277	347	59 877	7.9
U18	338	101	18	280	2.3
U19	2896	881	0	3634	2.7
U20	589	589	1097	5 416 419	9.3
U21	357	200	70	977	2.1
U22	1038	116	10	198	2.9
U23	677	576	200	9623	2.5
U24	1000	1000	13	61	0.6

TABLE 5. Possible values for the parameters of the StratVNS algorithm.

Parameter	Values
$k_{\max}$	2, 3
$t_{\max}$	5, 10, 15, 20
s	10, 20, 30, 40
s'	30, 40, 50, 60
$i_{\max}$	2, 3, 4
maxstart	2, 3, 4
nIterWithnoImpMax	5, 10

5.2. Parameters tuning

In addition to the stratification parameters, the VNS parameters were tuned. The values associated with these parameters, except for the maximum CPU time, were defined from experiments previously carried out with populations U01, U16, and U23. More specifically, the following sets of values were initially defined for each of these parameters according to Table 5.

Then, considering each 2304 combinations of these parameters, the number of strata $L = 3$ and $L = 4$, and cv_t equal to 5% and 10%, the StratVNS algorithm was applied 10 times in populations U01, U16, and U23, with 92 160 executions in total. For each combination, the median was calculated with the sample sizes produced in the ten runs. Finally, from the analysis of the five smallest median values, associated with the combinations of parameters, in each of the three populations evaluated considering the number of strata and the target coefficient

TABLE 6. Configuration of parameters.

Parameter	Values
$k_{\max}$	3
$t_{\max}$	15
s	30
s'	50
$i_{\max}$	3
maxstart	3
nIterWithnoImpMax	5
cpuTime	3600 s

TABLE 7. Total Sample Size (n) produced by each algorithm and number of strata (L) of the 24 populations, for $cv_t = 10\%$.

ID	$L=3$			$L=4$			$L=5$			$L=6$		
	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS
U01	26	23	22	14	13	13	14	11	10	13	13	12
U02	6	6	6	8	8	8	10	10	10	12	12	12
U03	22	21	20	14	12	12	10	10	10	12	12	12
U04	9	8	8	8	8	8	10	10	10	12	12	12
U05	35	34	34	21	19	19	14	12	12	12	12	12
U06	13	11	11	9	8	8	10	10	10	12	12	12
U07	23	21	21	16	14	13	14	10	10	14	12	12
U08	6	6	6	8	8	8	10	10	10	12	12	12
U09	11	10	10	8	8	8	10	10	10	12	12	12
U10	17	15	15	10	9	9	10	10	10	12	12	12
U11	19	19	20	12	11	11	10	10	10	12	12	12
U12	18	17	17	10	9	8	10	10	10	12	12	12
U13	22	21	21	13	11	13	10	10	10	12	12	12
U14	6	6	6	8	8	8	10	10	10	12	12	12
U15	20	18	17	11	10	9	11	10	10	12	12	12
U16	22	22	22	*	17	15	*	12	11	*	13	13
U17	17	17	16	10	9	9	10	10	10	12	12	12
U18	7	6	6	8	8	8	10	10	10	12	12	12
U19	15	15	15	11	9	9	10	10	10	12	12	12
U20	22	21	21	15	14	14	15	10	10	12	12	12
U21	8	8	7	8	8	8	10	10	10	12	12	12
U22	11	10	9	8	8	8	10	10	10	12	12	12
U23	12	10	10	9	8	8	10	10	10	12	12	12
U24	6	6	6	8	8	8	10	10	10	12	12	12

Notes. *Means that the algorithm did not converge.

of variation cv_t, the standard configuration with the greatest repetition among the five values was sought. The best configuration obtained is presented in Table 6, and it was used in all the experiments carried out in this work.

5.3. Computational results

This section compares the computational results obtained from the classical methods LH88 and Ko04 with those from the proposed algorithms, StratVNS and StratEnum. We considered eight scenarios corresponding to the parameters $cv_t = 10\%, 5\%$ and $L = 3, 4, 5, 6$ for each instance. We evaluate the results produced by the

TABLE 8. Total Sample Size (n) produced by each algorithm and number of strata (L) of the 24 populations, for $cv_t = 5\%$.

ID	$L = 3$			$L = 4$			$L = 5$			$L = 6$		
	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS
U01	54	54	54	38	37	36	30	26	25	22	20	19
U02	6	6	6	8	8	8	10	10	10	12	12	12
U03	73	73	72	44	42	42	30	27	27	21	20	20
U04	30	28	28	19	17	17	12	11	11	13	12	12
U05	121	120	120	72	70	70	47	44	44	33	32	32
U06	43	42	42	28	25	25	20	17	17	18	13	12
U07	43	42	42	33	31	31	24	22	23	24	18	18
U08	14	13	12	11	8	8	11	10	10	12	12	12
U09	38	37	37	23	22	22	16	15	15	12	12	13
U10	58	57	57	34	33	33	24	22	22	19	17	16
U11	75	73	74	44	42	43	29	28	29	23	20	21
U12	*	40	40	*	23	23	*	16	16	20	13	14
U13	74	73	73	42	41	41	29	27	28	22	20	22
U14	6	6	6	8	8	8	10	10	10	12	12	12
U15	39	38	38	*	25	25	*	17	17	*	14	13
U16	42	42	42	*	28	28	*	20	20	*	19	16
U17	41	41	41	26	25	25	18	17	17	17	14	13
U18	20	20	20	13	12	12	11	10	10	13	12	12
U19	57	56	56	35	33	33	24	22	22	19	16	16
U20	58	57	57	39	37	37	31	28	25	20	19	18
U21	25	25	25	20	14	14	13	11	10	15	12	12
U22	37	33	32	20	18	18	19	11	10	20	12	12
U23	38	37	37	25	23	23	21	16	15	17	12	12
U24	6	6	6	8	8	8	10	10	10	12	12	12

*Means that the algorithm did not converge.

TABLE 9. Number of best solutions by strata and cv.

	$cv_t = 10\%$			$cv_t = 5\%$		
L	LH88	Ko04	StratVNS	LH88	Ko04	StratVNS
3	7	18	23	8	21	23
4	9	20	23	3	23	23
5	18	22	24	3	19	21
6	21	23	24	5	17	20

StratVNS algorithm using LH88 and Ko04 as a baseline. Notice that the Ko04 and LH88 algorithms can produce infeasible solutions since neither of them was designed to obey constraints associated with the minimum sample size per stratum $n_{\min}$, especially when $n_{\min} \geq 3$. However, this did not occur for the populations used in these experiments.

In Tables 7 and 8, we present the results produced by each method (LH88, Ko04, StratVNS) for the number of strata ranging from 3 to 6 for all populations, with mininum sample size $n_h \geq 2$, $cv_t = 10\%$, and $cv_t = 5\%$, respectively. Table 9 and Figures 1 and 2 show a summary of the results presented in Tables 7 and 8. More specifically, the total of best solutions produced by each of the algorithms is presented for all eight scenarios.

Table 9 and Figures 1 and 2 show that the StratVNS algorithm had a performance far superior to the others, as it produced the most significant amount of best solutions for all strata numbers.

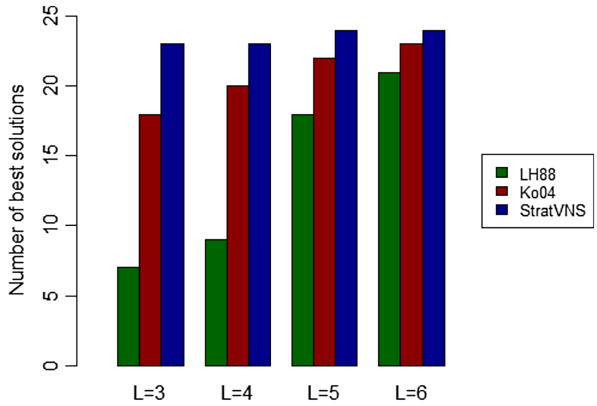

FIGURE 1. Number of best solutions per method with $cv_t = 10\%$.

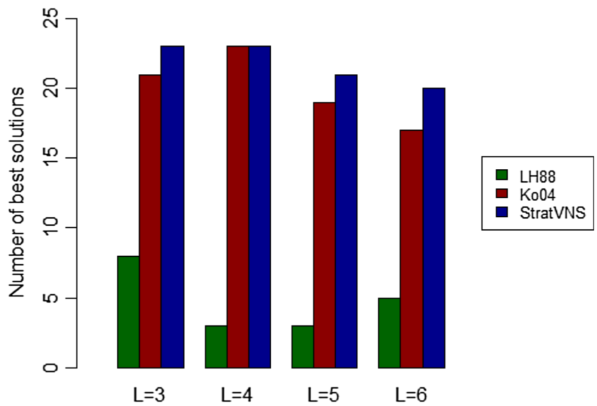

FIGURE 2. Number of best solutions per method with $cv_t = 5\%$.

In addition to the experiment with the StratVNS algorithm, a second experiment was carried out considering applying the StratENUM algorithm described in Section 4.2. Using the targets $cv_t = 5\%$ and $cv_t = 10\%$, StratEnum was applied to all populations where $T \leq 10^7$. In these cases, 74 global optimal solutions (37 for each cv_t) were obtained corresponding to the sample sizes (n), in a total of 192 possible solutions ($cv_t \times$ 24populations $\times$ number of strata). The optimal solutions produced were compared to the corresponding solutions produced by the StratVNS algorithm. Table 9, in the appendix, shows the disaggregated results (by population, cv_t and strata) to the global optimum produced by the StratENUM algorithm and the solutions produced by the StratVNS algorithm. Table A.2 also shows the obtained sample sizes and the total number of feasible solutions (tfeasible) evaluated by the StratENUM algorithm to get the global optimum. It is observed that the StratVNS algorithm achieved the global optimum in 37 out of 37 cases for $cv_t = 10\%$ (corresponding to 100%), and in 36 out of 37 cases (corresponding to 97%) for $cv_t = 5\%$.

For $cv_t = 10\%$, the average execution time of the StratVNS and StratENUM for 37 cases was, respectively, 17 and 5448 s. For $cv_t = 5\%$, the average execution time of the StratVNS and StratENUM was 25 and 9127 s, respectively. The StratENUM maximum execution time was obtained for population $U6$, corresponding to 26 h for $cv_t = 10\%$, and 55 h for $cv_t = 5\%$.

It is worth mentioning that the strata boundaries provided by StratVNS algorithm were not the same from the other two methods of the literature. To illustrate this fact, we present the strata boundaries, sizes, and variances of the population U12 for $L = 4$ and $cv_t = 10\%$ in Table A.3 of the appendix.

From the obtained results, the StratVNS algorithm can be considered an alternative to the classical methods of the literature to solve the stratification problem when we consider minimizing the sample size. Besides, the proposed algorithm proved to be the most appropriate one when there is a constraint on the minimum sample size per stratum.

6. Conclusion

The stratification problem has been studied since [9]. No algorithm in the literature always obtains the optimal global solution for this problem according to the stratified population. So far, the methods that produced the most favorable results were [28,33]. However, they do not allow us to include the constraint of minimum sample size per stratum. They also do not allow negative values in the observations of the stratification variable. In this article, it was possible to go one step further towards obtaining better quality solutions. The StratVNS algorithm produced better solutions than the two algorithms known in the literature in 94% of the studied instances. Moreover, it was possible to produce a global minimum for the considered stratification problem for the first time by applying the StratENUM algorithm, which is new in the literature. Furthermore, our method applies to all types of population, even those with negative values.

Possible extensions of this work include a generalization of the method for the multivariate stratification problem and changing the method to minimize the variance of an estimator, given sample size. There is also the possibility of testing other metaheuristics.

Appendix A.

Table A.1. Notation.

Parameter/Variable	Description
$U = \{1, 2, \ldots, N\}$	Population of size N
N	Number of population elements
L	Number of cutoff points
h	Stratum index
N_h	Number of population elements in stratum h
E_h	Set of population elements in stratum h
s_h	Set of population elements sampled in stratum h
n_h	Sample size of stratum h
n	Total sample size given by $\sum_{h=1}^{L} n_h$
Y	Variable of interest
$\overline{Y}_h$	Population mean of stratum h
S^2_{hy}	Population variance of stratum h
$\hat{Y}_{\text{AE}}$	Total estimator
$V(\hat{Y}_{\text{AE}})$	Variance of the total estimator $\hat{Y}_{\text{AE}}$
$cv(\hat{Y}_{\text{AE}})$	Coefficient of variation of the total estimator $\hat{Y}_{\text{AE}}$
cv_t	Target coefficient of variation or precision level
T_Y	Population total of the variable of interest Y
b_i	ith cutoff point

TABLE A.2. Sample sizes corresponding to the global optimum.

		cv_t=10%			cv_t = 5%		
POP	L	nStratENUM	tfeasible	nStratVNS	nStratENUM	tfeasible	nStratVNS
U01	3	22	60 726	22	54	60 726	54
U02	3	6	495 510	6	6	495 510	6
U03	3	20	495 510	20	72	495 510	72
U04	3	8	495 510	8	28	495 510	28
U05	3	34	632 250	34	120	632 250	120
U06	3	11	24 310	11	42	24 310	42
U07	3	21	116 403	21	42	116 403	42
U08	3	6	1081	6	12	1081	12
U09	3	10	4 011 528	10	37	4 011 528	37
U10	3	15	166 176	15	57	166 176	57
U12	3	17	33 670	17	40	33 670	40
U13	3	21	1 991 010	21	73	1 991 010	73
U14	3	6	495 510	6	6	495 510	6
U15	3	17	2016	17	38	2016	38
U16	3	22	79 003	22	42	79 003	42
U17	3	16	37 128	16	41	37 128	41
U18	3	6	4656	6	20	4656	20
U19	3	15	384 126	15	56	384 126	56
U20	3	21	170 820	21	57	170 820	57
U21	3	7	19 110	7	24	19 110	25*
U22	3	9	6216	9	32	6216	32
U23	3	10	163 306	10	37	163 306	37
U24	3	6	495 510	6	6	495 510	6
U01	4	13	6 963 596	13	36	6 963 596	36
U06	4	8	1 750 540	8	25	1 750 540	25
U08	4	8	15 180	8	8	15 180	8
U12	4	8	2 862 209	8	23	2 862 209	23
U15	4	9	39 711	9	25	39 711	25
U17	4	9	3 317 040	9	25	3 317 040	25
U18	4	8	142 880	8	12	142 880	12
U21	4	8	1 216 865	8	14	1 216 865	14
U22	4	8	221 815	8	18	221 815	18
U08	5	10	148 995	10	10	148 995	10
U15	5	10	557 845	10	17	557 845	17
U18	5	10	3 183 545	10	10	3 183 545	10
U22	5	10	5 773 185	10	10	5 773 185	10
U08	6	12	1 086 008	12	12	1 086 008	12

Notes. *The only case in which StratVNS did not produce the global optimum.

TABLE A.3. Detailed Results for Population U12 ($L = 4$ and $cv_t = 10\%$).

	b_h			N_h				S^2_{hx}			
Ko04	1.018, 1	2.929, 9	12.724, 0	163	85	33	3	41.304, 9	231.853, 7	1.949.479, 6	103.280.422, 2
LH88	1.010, 0	2.830, 5	16.302, 0	163	84	34	3	41.304, 9	216.532, 1	1.999.348, 9	103.280.422, 2
StratVNS	1.061, 0	3.070, 0	7.910, 0	170	80	31	3	49.779, 9	253.597, 9	1.854.485, 0	103.280.422, 2

Let X be the vector with 284 values associated with population U12. The results of Table A.3 were obtained using the following commands in R:

strata.LH(X,CV=0.1,Ls=4,alloc=c(0.5,0,0.5),takeall=0, algo = "Kozak", model="none")

strata.LH(X,CV=0.1,Ls=4,alloc=c(0.5,0,0.5),takeall=0, algo = "Sethi", model="none")

STRATVNS(X,L=4)

Acknowledgements. The research of Leonardo de Lima was partly funded by the CNPq grant number 314298/2018-5.

References

[1] S. Baillargeon and L. Rivest, The construction of stratified designs in R with the package stratification. *Surv. Methodol.* **37** (2011) 53–65.

[2] M. Ballin and G. Barcaroli, Joint determination of optimal stratification and sample allocation using genetic algorithm. *Surv. Methodol.* **39** (2013) 369–393.

[3] J.A.M. Brito, L. Ochi, F.M.T. Montenegro and N. Maculan, An iterative local search approach applied to the optimal stratification problem. *Int. Trans. Oper. Res.* **17** (2010) 753–764.

[4] J.A.M. Brito, P.L.N. Silva, G.S. Semaan and N. Maculan, Integer programming formulations applied to optimal allocation in stratified sampling. *Surv. Methodol.* **41** (2015) 427–442.

[5] J. Brito, G. Semaan, A. Fadel and L. Brito, An optimization approach applied to the optimal stratification problem. *Commun. Stat. Simul. Comput.* **46** (2017) 4491–4451.

[6] J. Brito, T. Veiga and P. Silva, An optimisation algorithm applied to the one-dimensional stratification problem. *Surv. Methodol.* **45** (2019) 295–315.

[7] R. Chambers and R. Dunstan, Estimating distribution functions from survey data. *Biometrika* **73** (1986) 597–604.

[8] W.G. Cochran, Samppling Techniques, 3rd edition. *Wiley Series in Probability and Statistics* (2007).

[9] T. Dalenius, The problem of optimum stratification. *Skandinavisk Aktuarietidskrift* **1950** (1950) 203–213.

[10] T. Dalenius and J. Hodges, Minimum variance stratification. *J. Am. Stat. Assoc.* **285** (1959) 88–101.

[11] F. Danish, A mathematical programming approach for obtaining optimum strata boundaries using two auxiliary variables under proportional allocation. *Stat. Trans. New Ser.* **19** (2018) 507–526.

[12] F. Danish and S. Rizvi, Optimum stratification in bivariate auxiliary variables under neyman allocation. *J. Mod. Appl. Stat. Methods* **17** (2018) 2–24.

[13] F. Danish, S. Rizvi, M. Jeelani and J. Reshi, Obtaining strata boundaries under proportional allocation with varying cost of every unit. *Pak. J. Stat. Oper. Res.* **13** (2017) 567.

[14] F. Danish, S. Rizvi, M.K. Sharma, M.I. Jeelani, B. Kumar and Q.F. Dar, Optimum stratification for two stratifying variables. *Rev. Invest. Oper.* **40** (2019) 562–573.

[15] G. Ekman, An approximation useful in univariate stratification. *Ann. Math. Stat.* **30** (1959) 219–229.

[16] G. Glasser, On the complete coverage of large units in a statistical study. *Rev. Int. Stat. Inst.* **30** (1962) 28–32.

[17] F. Glover and G.A. Kochenberger, Handbook of Metaheuristics. Springer (2003).

[18] P. Gunning and J.M. Horgan, A new algorithm for the construction of stratum boundaries in skewed populations. *Surv. Methodol.* **30** (2004) 159–166.

[19] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2011).

[20] P. Hansen and N. Mladenović, Variable neighborhood search: principles and applications. *Eur. J. Oper. Res.* **130** (2001) 449–467.

[21] P. Hansen, N. Mladenović and D. Perez-Brito, Variable neighborhood decomposition search. *J. Heuristics* **7** (2001) 335–350.

[22] D. Hedlin, A procedure for stratification by an extended Ekman rule. *J. Official Stat.* **16** (2000) 15–29.

[23] M.A. Hidiroglou, The construction of a self-representing stratum of large units in survey design. *Am. Stat.* **40** (1986) 27–31.

[24] M. Hidiroglou and M. Kozak, Stratification of skewed populations: a comparison of optimisation-based versus approximate methods. *Int. Stat. Rev.* **86** (2018) 87–105.

[25] T. Keskintürk and S. Er, A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Comput. Stat. Data Anal.* **52** (2007) 53–67.

[26] M. Khan, N. Nand and N. Ahmad, Determining the optimum strata boundary points using dynamic programming. *Surv. Methodol.* **34** (2008) 205–214.

[27] L. Kish, Survey Sampling. Wiley New York, Chichester (1965).

[28] M. Kozak, Optimal stratification using random search method in agricultural surveys. *Stat. Trans.* **6** (2004) 797–806.

[29] M. Kozak, Multivariate sample allocation: application of a random search method. *Stat. Trans.* **7** (2006) 889–900.

[30] M. Kozak, Comparison of random search method and genetic algorithm for stratification. *Commun. Stat. Simul. Comput.* **43** (2014) 249–253.

[31] M. Kozak and M.R. Verma, Geometric versus optimization approach to stratification: a comparison of efficiency. *Surv. Methodol.* **32** (2006) 157–163.

[32] M. Kozak, M.R. Verma and A. Zieliński, Modern approach to optimum stratification: review and perspectives. *Stat. Trans.* **8** (2007) 223–250.

[33] P. Lavallée and M.A. Hidiroglou, On the stratification of skewed populations. *Surv. Methodol.* **14** (1988) 33–43.

[34] B. Lednicki and R. Wieczorkowski, Optimal stratification and sample allocation between subpopulations and strata. *Stat. Trans.* **6** (2003) 287–305.

[35] J. Lisic, H. Sang, Z. Zhu and S. Zimmer, Optimal stratification and allocation for the june agricultural survey. *J. Official Stat.* **34** (2018) 121–148.

[36] S.L. Lohr, Sampling: Design and Analysis, 2nd edition. Chapman and Hall/CRC (2019).

[37] D. Rao, M. Khan and K. Reddy, Optimum stratification of a skewed population. *Int. J. Math. Comput. Sci.* **8** (2014) 492–495.

[38] K. Reddy and M. Khan, Optimal stratification in stratified designs using weibull-distributed auxiliary information. *Commun. Stat. Theory Methods* **48** (2019) 3136–3152.

[39] K. Reddy and M. Khan, stratifyR: an R package for optimal stratification and sample allocation for univariate populations. *Aust. New Zealand J. Stat.* **62** (2020) 383–405.

[40] K. Reddy, M. Khan and S. Khan, Optimum strata boundaries and sample sizes in health surveys using auxiliary variables. *PLoS ONE* **13** (2018) e0194787.

[41] L. Rivest, A generalization of the Lavallé and Hidiroglou algorithm for stratification in business surveys. *Surv. Methodol.* **28** (2002) 191–198.

[42] S. Ross, A First Course in Probability, 10th edition. Pearson (2018).

[43] V. Sethi, A note on the optimum stratification of populations for estimating the population means. *Aust. J. Stat.* **5** (1963) 20–33.