An ILS Approach Applied to the Optimal Stratification Problem

José Brito[a], Luiz Ochi[b], Flávio Montenegro[c] and Nelson Maculan[d]

[a]*IBGE, Instituto Brasileiro de Geografia e Estatística, DPE/COMEQ, Av.Chile, 500, 10 Andar, Centro, Rio de Janeiro, RJ, Brazil.* , `jose.m.brito@ibge.gov.br`
[b]*UFF, Universidade Federal Fluminense, Instituto de Computação, Rua Passo da Pátria 156, Bloco E, 3 andar, São Domingos, Niterói, RJ, Brazil.* , `satoru@dcc.ic.uff.br`
[b]*IBGE, Instituto Brasileiro de Geografia e Estatística, DPE/COMEQ, Av.Chile, 500, 10 Andar, Centro, Rio de Janeiro, RJ, Brasil.* , `flavio.montenegro@ibge.gov.br`
[d]*COPPE, Universidade Federal do Rio de Janeiro, P.O. Box 68511, 21941-972 Rio de Janeiro, Brazil.* , `maculan@cos.ufrj.br`

Keywords: Stratification, Metaheuristics, ILS, Strata Boundaries.

Abstract.

Stratified sampling is a technique that consists in separating the elements of a population into non-overlapping groups, called strata. This paper describes a new algorithm to solve the one-dimensional case, which reduces the stratification problem to just determining strata boundaries. Assuming that the number L of strata and the total sample size n are predetermined, we obtain the strata boundaries by taking into consideration an objective function associated with the variance. In order to solve this problem, we have implemented an algorithm based on the ILS metaheuristic. Computational results obtained from a real data set are presented and discussed.

1 Introduction

Stratification is a widely used sample survey technique (Cochran, 1977). The sampling frame is divided into strata and independent samples are drawn from each stratum. One reason to use stratification is that the survey designer forms homogenised strata, which are achieved if important study variables vary less within strata than within the unstratified population. According to Scheaffer (1990) the main reasons for using stratified random sampling are:

- Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample. This result is particulary true if measurements within strata are homogeneus.
- The cost per observation in the survey may be reduced by stratification of the population elements into convenient groupings.
- Estimates of population parameters may be desired for subgroups of the population that identify the strata.

The determination of stratum boundaries is one of the main problems of stratified sampling. Some rules and methods to adress this problem have been written by several authors such as Ekman (1959), Dalenius and Hodges (1959), Lavallée and Hidiroglou (1988), Hedlin (1998, 2000), Kozak (2004); Kozak and M.R. (2006), Gunning and Horgan (2004).

Strata boundaries are determined by taking into account a stratification variable X, whose values are known for all population units. The process aims to find homogeneous strata in such a way to minimize the sum of the variances of a study variable Y, correlated with X, or minimize the variance of the variable X itself (Hedlin, 2000), within the strata. The variance calculus depends on the sample allocation scheme taken in consideration, that is, on the number or proportion of sample units that must be allocated within each stratum. In this work, we will consider the stratification problem with the Neyman optimal allocation (Cochran, 1977), that targets to provide the more homogeneous strata.

Some previous works have been proposed to solve this combinatorial problem with Neyman allocation by using non-search based approaches. Also, those proposals focuse on just to find homogeneous strata, expecting to minimize the sum of the variances within the strata as a consequence. As an alternative, we propose in this work an ILS (Iterative Local Search) based metaheuristic algorithm that both performs a guided random exploration of the search space and focuses its search directly on the sum of variances itself, in order to provide good (non-exact) solutions to the problem.

This paper is organized as follows: In section 2, the stratification problem is introduced. In section 3, we give a brief description of the ILS metaheuristic and apply it for solving the stratification problem. Finally, the last section presents and compares some computational results obtained by applying the ILS algorithm and three algorithms from literature, considering a real and an artificial data set.

2 Stratified Random Sampling

In stratified sampling the population of N units is partitioned into strata of $N_1, N_2, ..., N_L$ units, respectively (Cochran, 1977; Sarndal, 1992). These subpopulations are nonoverlapping. Together they comprise the whole population, so that

$$N_1 + N_2 + ... + N_h + ... + N_L = N \tag{1}$$

To obtain the full benefit from stratification, the values of N_h must be known. When the strata have been determined, a sample n is selected by some design, determining a sample of size n_h within each stratum h, so that

$$n_1 + n_2 + ... + n_h + ... + n_L = n \tag{2}$$

Because the selections in different strata are made independently, the variances for the individual strata can be added to obtain variances for the whole population. Since only the within-stratum variances enter into the variances, the principle of stratification is to partition the population in such a way that the units within a stratum are as similar as possible.

Stratification is a common technique that may produce a gain in precision in the estimates of the characteristics of the whole population. It may be possible to divide a heterogeneous population into subpopulations, each one internally homogeneous.

2.1 Stratification Problem

In this problem, a sample of size n is taken from the population $U = \{1, 2, ..., N\}$, considering a study variable Y related to the desired estimates in the survey. The population is partitioned into a given number L of strata, namely $A_1, A_2, ..., A_L$. Strata are built taking into account a size variable X, which is correlated with the study variable Y and whose observations are known for all the population elements.

Let $Y_U = y_1, y_2, ..., y_N$ be a population vector associated to the study of variable Y and $X_U = x_1, x_2, ..., x_N$ be the population vector generated by the corresponding observations of the stratification variable X, where $x_1 \leq x_2 \leq ... \leq x_N$. Strata are determined by the cutting points (boundaries) $b_1 < b_2 < ... < b_h < ... < b_{L-1}$, in such a way that

$A_1 = \{i : x_i \leq b_1\}$

$A_h = \{i : b_{h-1} < x_i \leq b_h\}\ h = 2, 3, ..., L-1$

$A_L = \{i : b_{L-1} < x_i\}$

Once given the boundaries of strata, a simple random sample of size n_h is taken of each stratum. Such boundaries should be defined in order to minimize the variance.

$$v(\hat{Y}) = \sum_{h=1}^{L} N_h^2 \frac{S_{hy}^2}{n_h}(1 - \frac{n_h}{N_h}), \tag{3}$$

where N_h and n_h are respectively the number of frame units and the sample size in stratum h. And S_{hy}^2 is the study variable variance in stratum h:

$$S_{hy}^2 = \sum_{i=1}^{N_h} \frac{\left(y_{hi} - \overline{Y}_h\right)^2}{N_h - 1} \tag{4}$$

$\overline{Y}_h$ and y_{hi} are respectively the variable mean in stratum h and the value of a variable of interest y for the unit i in the stratum h .

When the question is to allocate the sample size among strata, there are several alternative methods such as equal, proportional and Neyman allocation (Cochran, 1977). The equal allocation method is the simplest one, where all the stratum sample sizes are the same. With the proportional allocation method, the sample size in each stratum is proportional to the size of the stratum. These two methods are efficient and suitable if the variances within the strata are similar Cyert (1962). On the other hand, if the stratum variances differ substantially, as in, for example, highly skewed populations, the Neyman allocation method should be used. This method is based on the principle of sampling fewer elements from homogeneous strata and more elements from strata with high internal variability.

In this work, we considered the Neyman's Allocation to allocate the total sample size n among the L strata. Assuming that sample costs are to be equal for all strata, the stratum sampling sizes using Neyman's allocation scheme are given by (Cochran, 1977):

$$n_h = \frac{n.N_h.S_{yh}}{\sum_{k=1}^{L} N_k.S_{yk}} \tag{5}$$

Alternatively, we can also consider the minimization of the coefficient of variation (cv):

$$cv(\hat{Y}) = 100.\frac{\sqrt{v(\hat{Y})}}{Y} \tag{6}$$

where $Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}$

We also work under the assumptions that the values of the stratification variable are known and that, for simplicity, those values are equal to the study variable ones. Many authors draw on this assumption, among others Dalenius and Hodges (1959),Ekman (1959),Lavallée and Hidiroglou (1988), Hedlin (2000) and Mehta (n.d.).

We emphasize that finding a global minimum for (3) or (6) is a hard task, either analytically or by intensive computing methods, because S_{hy}^2 is a nonlinear function of $b_1 < b_2 < ... < b_{L-1}$ and the number of different choices for these values may be very high. In section 3 we discuss this issue using some combinatorial arguments.

Therefore, several methods which yield a local minimum have been suggested. A well-known method of strata definition was proposed in Dalenius and Hodges (1959). It consists of approximating the distribution of the variable of stratification X in the population, by using a histogram with various classes. Its adoption therefore implies the assumption that the variable of stratification has a uniform distribution (Cochran, 1977) in each class. In this case, the problem of stratification has an ordinary solution when the Cumulative Root Frequency (*CumRoot*) Algorithm, or Dalenius-Hodges rule is applied (please, see (Cochran, 1977), chapter 5 and Sarndal et al. (Sarndal, 1992)).

The method implemented by Hedlin (1998, 2000) is associated with an extended Ekman rule (Ekman, 1959). For this reason, the method of Hedlin is also called method of Hedlin altered. According to Hedlin (2000), the strata delimitation is considered such that the variance of the total estimator of a variable of interest, given by (3), has to be a minimum, considering n and L fixed previously and applying Neyman's allocation in each strata.

Kozak (2004) presents the modified random search algorithm as a method of the optimal stratification presented by Rivest (2002). In a more recent work, Kozak and M.R. (2006) performs comparisons between random search method and the geometric and Lavallée and Hidiroglou (1988) approaches.

Gunning and Horgan (2004) and Horgan (2006) developed an algorithm that is easier to implement and that applies the general term of a geometric progression to establish the boundaries of the strata. Stratifying a population by a variable is to subdivide it into intervals with cutting points $b_0 < b_1 < ... < b_L$. The division should be based on the auxiliary variable X , correlated with the study variable Y . In order to define cutting points $(b_0, b_1, ..., b_L)$ Gunning and Horgan (2004) use the following recurrence relation:

$$b_h^2 = b_{h+1}.b_{h-1} \tag{7}$$

By this relation the stratum boundaries are terms of the following geometric progression:

$$b_h = a.r^h \ (h = 1, ..., L-1) \tag{8}$$

Thus, $a = b_0$ is the minimum value of the variable $(b_0 = x_1)$, and $ar^L = b_L$, the maximum value $(b_L = x_N)$ of the variable. It follows that the constant ratio can be calculated as $r = (\frac{b_L}{b_0})^{\frac{1}{L}}$. After stratum boundaries definition, Neyman's allocation (5) is applied.

Gunning and Horgan (2007) also proposed a new approach to generate initial boundaries to Lavallé and Hidiroglou algorithm, in order to improve the rate of convergence to the optimal solution, often resulting in smaller sample sizes.

Khan and N. (2008) use a dynamic programming algorithm for determining the optimum strata boundary points. This algorithm is applied considering two particular cases: the variable X have a normal or a triangular distribution. They also consider the hypothesis that the sampling is with replacement, or that sampling ratios $\frac{n_h}{N_h}$ are small.

3 Proposed Algorithm

We present a new proposal to solve the stratification problem with Neyman allocation. It is a search-based method that intends to work for variables with any distribution. In the subsequent section, we will present some computational results obtained by applying this algorithm to some skewed populations.

3.1 Iterated Local Search (ILS)

Iterated Local Search is a metaheuristic presenting desirable features like simplicity, robustness and high effectiveness, when applied to a wide range of problems. According to Lourenço et al.(Glover and Kochenberger, 2002), its essential idea ”lies in focusing the search not on the full space of solutions but on a smaller subspace defined by the solutions that are locally optimal for a given optimization engine”. The success of this method is directly associated with the choice of the local search procedure, the perturbation procedure and the acceptance criterion.

The pseudo-code below shows the essential ILS steps. Step (1) constructs an initial solution to which a local search is applied in step (2), in order to produce a solution s^*. Aiming at reaching a better solution from s^*, steps (3) and (4) apply procedures of perturbation and local search, respectively, that result in a new solution $s^{''}$ to be confronted with the solution s^*.

In step (5), if the new solution $s^{''}$ satisfies an acceptance criterion based on s^*, then the attribution $s^* = s^{''}$ is performed, otherwise the solution s^* is maintained. Then, steps (3), (4) and (5) proceed iteratively along m iterations, when the final solution s^*, the best one, is outputted.

```
$(1) s_0 = GenerateInitialSolution$
$(2) s^* = LocalSearch(s_0)$
$Repeat$
$(3) s' = Perturbation(s*)$
$(4) s^{''} = LocalSearch(s')$
$(5) s^* = AcceptanceCriterion(s^*, s^{''})$
$Until(Termination\ condition\ met)$
```

3.2 ILS Algorithm for Stratification

In order to implement the new methodology, we had to modify the input data structure. Since the N observations X_U are ordered in ascending order, it is possible to gather them, taking into account only their distinct values. Thus, we have P distinct values of X_U, gathered in a set $B = \{b_1, b_2, ..., b_P\}$, which are the eligible cutting points to stratify the whole population.

Consider, for example, $N = 12$, $L = 3$ and $X_U = (2, 4, 4, 8, 10, 11, 14, 15, 15, 15, 17, 18)$. Then we have $B = (2, 4, 8, 10, 11, 14, 15, 17, 18) = (b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9) \Rightarrow P = 9$, and we may define, for example, $A_1 = \{i | x_i \leq b_3, x_i \in X_U\}$, $A_2 = \{i | b_3 < x_i \leq b_6, x_i \in X_U\}$ and $A_3 = \{i | b_6 < x_i, x_i \in X_U\}$ as a candidate solution for the stratification problem (with cutting points $b_3 = 8$ and $b_6 = 14$).

Analogously, for L strata and P boundaries , we have to find $(L - 1)$ boundaries b_k from B. In short, considering a finite population of size N, that will be divided in L strata, and the ordered values X_U, we establish the set B. The solution for this problem will then consist of boundaries b_k , selected from B, that give the minimum variance according to (3).

The solution of the problem above can be obtained by considering the enumeration of all the possible divisions of the observations associated to the set I, that is, by evaluating the variances of all the solutions and selecting the one with the lowerst variance. However, this procedure may take an excessively high computational time even for moderately high values of the number of observations of I and/or the number L of strata. In fact, determining the number of solutions to be considered corresponds to solve the following combinatorial problem:

Determine the number m of non-negative integer solutions of the equation

$$w_1 + w_2 + ... + w_h + ... + w_q = r \tag{9}$$

Each w_h corresponds to the number of elements of I allocated in each stratum h, q corresponds to the number L of strata and r corresponds to the total of elements in I. In accordance with the stratification problem, it is necessary to introduce a minor modification in equation (9) in order to guarantee that the number of population observations is greater than or equal to 2 in each stratum ($w_h \geq 2$).

Changing the variable $w_h = t_h + 2$, $t_h \geq 0$, and considering $q = L$ and $r = |I|$, the equation (9) can be rewritten as

$$t_1 + t_2 + ... + t_h + ... + t_L = |I| - 2L \tag{10}$$

and the number of solutions of the equation (10) is given by:

$$m = \frac{(L + |I| - 2L - 1)!}{(|I| - 2L)!(L-1)!} \tag{11}$$

The number m of solutions increases very rapidly with L and $|I|$. For example, if $|I| = 100$ and $L = 5$ we will generate $m = 3.049.501$ solutions, and we will reach $m = 40.430.556.376$ solutions if $|I| = 1000$ and $L = 5$. It is important to remark that, in the case of an exhaustive procedure, we will need to generate each one of these solutions and also evaluate the stratum variances (see equation 3) for each set of strata, in order to stablish the stratum sample sizes.

Then, trying to reduce the potentially high number of operations needed to obtain the optimal solution using the exhaustive enumerating process, we will instead solve the stratification problem by applying an algorithm based on ILS metaheuristc and provide a good feasible solution.

Generate Initial Solution: Initially, a set $B = \{b_1, b_2, ..., b_P\}$ with P values, corresponding to possible cutting points, is defined. Then, a given number q of initial vectors are generated, each one containing $(L-1)$ values randomly selected from B and ordered in ascending order. Note: In this work, whenever we refer to a random selection we are considering an uniform distribution. If $L = 5$ and $P = 40$, we may have, for example, the following solution vectors:

b_3	b_9	b_{16}	b_{28}
b_2	b_{11}	b_{25}	b_{39}
b_4	b_{16}	b_{22}	b_{37}
b_6	b_{14}	b_{28}	b_{34}

Among the q solutions, the chosen initial solution s^0 is the one whose cutting points define strata with minimum total variances, according to equation (3). A number $q' < q$ of the remaining solutions with the smaller total variances according to (3), is stored into a set E. Set E is updated at each algorithm iteration by replacing its worst solution by the one resulting from the local search.

Perturbation Procedure: One of the $(L-1)$ cutting points b_i of the solution s^* and one of the $(P-L+1)$ cutting points b_j not belonging to s^* are randomly selected. Then, we replace b_i by b_j in s^*, generating s'. After this replacement, if the condition $b_s < b_i < ... < b_t$ is not satisfied anymore, then we must reorder the elements of s' in ascending order.

Local Search Procedure: It consists in a procedure of replacement followed by a path relinking procedure (Glover and Laguna, 1997). In the replacement procedure, which is applied to all the m algorithm iterations, one of the $(L-1)$ cutting points b_i of the solution s', resulting from the perturbation procedure, is randomly selected. Each cutting point b_i is then replaced by $b_{i+1}, b_{i+2}, ..., b_{k-2}$ and $b_{r+2}, ..., b_{i-2}, b_{i-1}$, where b_k is the next cutting point in the solution and b_r is the previous one. In the example below, considering $i = 16, k = 28$ and $r = 10$, we have

s"	...	b_r	b_i	b_k	...
s"	...	b_{10}	b_{16}	b_{28}	...
s"	...	b_{10}	b_{17}	b_{28}	...
s"	...	b_{10}	b_{18}	b_{28}	...
s"	...	b_{10}	...	b_{28}	...
s"	...	b_{10}	b_{26}	b_{28}	...
s"	...	b_{10}	b_{15}	b_{28}	...
s"	...	b_{10}	b_{14}	b_{28}	...
s"	...	b_{10}	b_{13}	b_{28}	...
s"	...	b_{10}	b_{12}	b_{28}	...

The path relinking procedure is applied to each w $(m \ mod \ w = 0)$ iterations, considering the current best solution and all the solution $s^e \in E$. By performing increments or decrements in each one of the cutting points associated to the solutions s^e, intermediary solutions s^i are obtained. They are stored if $f(s^i) < f(s^*)$.
The following example illustrates the application of this procedure:

s^*	b_3	b_{13}	b_{20}
s^e	b_1	b_{11}	b_{23}
s^i	b_2	b_{11}	b_{23}
s^i	b_3	b_{11}	b_{23}
s^i	b_3	b_{12}	b_{23}
s^i	b_3	b_{13}	b_{23}
s^i	b_3	b_{13}	b_{22}
s^i	b_3	b_{13}	b_{21}
$s^i = s^*$	b_3	b_{13}	b_{20}

Acceptance Criterion: This is a simple criterion consisting of evaluating the relative distance between the current best solution s^* and the solution s" provided by the local search:

$$\boxed{\begin{array}{l} If \ (f(s") < f(s^*)) \ Then \ s^* = s" \\ \ else \, if \ \ |f(s^*) - f(s")|/f(s^*) < \epsilon \ \ Then \ s^* = s" \end{array}}$$

where f is the objective function value considering equation (3) and ϵ is a tolerance factor.

4 Computational Results

In this experiment, we compare the performance of the new algorithm with the Geometric (Gunning and Horgan, 2004), Cumulative Root Frequency and Random Search (Kozak, 2004) algorithms. We used the Delphi language to implement the ILS and R language (http://www.r-project.org/) to implement the Geometric, the Cumulative Root Frequency and the Random Search algorithms . All computational results were obtained on an AMD Core Duo 2.31 Ghz CPU with 2GB RAM running Windows XP.

Since the Geometric algorithm is deterministic (see (Gunning and Horgan, 2004)) and since the Cumulative Root Frequency algorithm Dalenius and Hodges (1959) is a procedure of direct application, possible performance differences due to a different implementation of the Geometric algorithm are not expected to be important in this study.

Thus, we did not perform detailed time comparisons between algorithms (roughly, for the instances in this paper, the Geometric and Cumulative Root Frequency algorithms ran immediately, and the ILS spent on average less than five seconds of CPU time). Also, as many other authors (see section 2.1), we work under the assumption that the values of a study variable Y are equal to those of the stratification variable X.

In order to perform the comparisons, we used just sixteen populations skewed populations (with degrees of skewness varying from 1.4 to 34.8), which were arbitrarily chosen : 1) six populations extracted from *PAM - Produção Agrícola Municipal de 2004* (Municipal Agricultural Production), and associated to the total area of harvest in cities in the states of Ceará (CE), Minas Gerais (MG), Paraná (PR), Rio Grande do Sul (RS), Santa Catarina (SC) and São Paulo (SP); 2) two populations (PopId) extracted from *Pesquisa Industrial Anual de 2004* (Annual Industrial Survey), and associated to the number of persons employed ; 3) one population (PopAgr) extracted from *Censo Agropecuário 1995-1996* (Agricultural Census), associated to the effective production of coffee; and 4) seven populations randomly generated (PopRd1-7) using a macro implemented by Hedlin (Hedlin, 2000). Although we have performed computational tests just to skewed populations, we emphasize that the method we are proposing is intended to apply for populations with any distribution of the study and/or stratification variables.

For application of the ILS algorithm in each instance, we define the following parameters: number of initial solutions q equal to 30, number of solutions of E equal to 10, tolerance factor ϵ equal to 0.05 and the total number of iterations (m) equal to 50. A path relinking procedure was applied every $w = 10$ iterations. And to the Random Search algorithm, according Kozak (2004) we define number of iteractions $R = 1000$, $p = 3$ e and the initial strata boundaries are obtained applying the Geometric Algorithm.

Table 1 gives information about the instances such as: identification, total population, P number of observations in set B and sample size. It also contains the values of the coefficients of variation (equation (6)) obtained by applying the ILS, Geometric, Cumulative Root Frequency and Random Search algorithms for each studied combination (population × number of strata).

As we can observe in Table 1, the ILS algorithm provided, in general, minimum coefficients of variation, meaning a more robust performance when compared with the other three algorithms.

Following Gunning and Horgan (2004) and others, in order to investigate whether the ILS leads to a more efficient estimation than the Cumulative Root Frequency, Geometric and Random Search algorihms, we can evaluate the relative efficiency via equations (12) and (14). Table 2 contains values of the relative efficiencies for each combination studied (population × number of strata).

$$eff_{V_{Geo},V_{ILS}} = \frac{V_{Geo}(\hat{X})}{V_{ILS}(\hat{X})} \tag{12}$$

$$eff_{V_{Cf},V_{ILS}} = \frac{V_{Cf}(\hat{X})}{V_{ILS}(\hat{X})} \tag{13}$$

$$eff_{V_{Cf},V_{RS}} = \frac{V_{RS}(\hat{X})}{V_{ILS}(\hat{X})} \tag{14}$$

where V_{Geo} , V_{Cf} , V_{RS} and V_{ILS} are the variances (3) under the Geometric, Cumulative Root Frequency, Random Search and ILS algorithms, respectively.

From the analysis of Table 2, we can see that the ILS algorithm produced in most cases, better solutions than the Geometric, Cumulative Root Frequency and Random Search ones in roughly all the cases, regardless of the number of strata. Also, the Geometric algorithm had the worst performance.

Pops				ILS			Geometric			Cum.Freq			RS		
	Size	Obs.	Sample	Strata (L)			Strata (L)			Strata (L)			Strata(L)		
Label	N	P	n	3	4	5	3	4	5	3	4	5	3	4	5
PopCE	184	173	80	1.7	1.2	0.9	5.6	3.9	3.5	2.7	1.9	1.5	1.2	1.3	1.1
PopMG	845	238	100	4.1	2.7	2.1	9.4	6.8	4.8	4.6	3.0	2.4	4.4	2.9	2.3
PopPR	397	268	120	2.3	1.6	1.2	4.0	2.9	2.3	2.6	2.0	1.5	2.7	2.0	1.7
PopRS	489	197	60	4.9	3.4	2.6	9.8	7.9	6.6	5.2	3.6	2.9	7.1	4.9	3.1
PopSC	283	176	100	2.3	1.7	1.3	5.6	4.1	3.0	2.8	1.9	1.4	2.5	1.7	1.3
PopSP	586	272	100	3.3	2.5	2.0	7.8	5.2	3.6	3.5	2.7	2.2	3.4	2.2	1.8
PopId1	2911	247	140	4.4	3.0	2.4	4.6	3.2	2.5	4.6	3.5	2.5	4.5	3.0	2.5
PopId2	1076	88	40	4.7	3.6	2.8	6.6	5.2	3.8	7.3	4.4	3.0	6.6	5.2	3.8
PopAgr	20472	784	100	6.7	4.8	3.9	6.8	5.0	4.0	6.9	5.2	4.0	7.4	5.5	4.4
PopRd1	1000	1000	100	0.6	0.4	0.3	0.8	0.6	0.5	0.6	0.5	0.4	0.7	0.5	0.4
PopRd2	1000	1000	100	4.1	3.0	2.3	10.1	7.9	6.4	4.3	3.7	2.5	4.0	3.2	2.5
PopRd3	1000	1000	100	2.5	1.9	1.5	4.2	3.2	2.4	2.5	2.0	1.7	2.5	2.1	1.7
PopRd4	1000	1000	100	1.4	1.1	0.9	1.8	1.4	1.2	1.7	1.1	1.0	1.8	1.4	1.1
PopRd5	1000	1000	50	0.5	0.4	0.3	0.6	0.5	0.4	0.5	0.5	0.4	0.5	0.4	0.4
PopRd6	1907	707	200	2.4	1.8	1.2	2.4	2.2	1.6	3.5	2.1	1.4	2.4	2.3	1.6
PopRd7	3838	1110	195	3.8	2.8	2.2	4.1	3.3	2.6	3.9	2.9	2.3	4.1	3.3	2.7
		Mean	$\overline{cv}$	3.1	2.2	1.7	5.3	4.0	3.1	3.6	2.6	1.9	3.5	2.6	2.0

Table 1: Results for the ILS, Geometric, Cumulative Root Frequency and Random Search Algorithms

Pops	$eff_{Geo,ILS}$			$eff_{Cf,ILS}$			$eff_{RS,ILS}$		
	Strata (L)			Strata (L)			Strata (L)		
Label	3	4	5	3	4	5	3	4	5
PopCE	10.9	10.6	15.1	2.5	2.5	2.8	0.5	1.2	1.5
PopMG	5.3	6.3	5.2	1.3	1.2	1.3	1.2	1.2	1.2
PopPR	3.0	3.3	3.7	1.3	1.6	1.6	1.4	1.6	2.0
PopRS	4.0	5.4	6.4	1.1	1.1	1.2	2.1	2.1	1.4
PopSC	5.9	5.8	5.3	1.5	1.2	1.2	1.2	1.0	1.0
PopSP	5.6	4.3	3.2	1.1	1.2	1.2	1.1	0.8	0.8
PopId1	1.1	1.1	1.1	1.1	1.4	1.1	1.0	1.0	1.1
PopId2	2.0	2.1	1.8	2.4	1.5	1.1	2.0	2.1	1.8
PopAgr	1.0	1.1	1.1	1.1	1.2	1.1	1.2	1.3	1.3
PopRd1	1.8	2.3	2.8	1.0	1.5	1.4	1.4	1.6	1.8
PopRd2	6.1	6.9	7.1	1.1	1.5	1.1	1.0	1.1	1.2
PopRd3	2.8	2.8	2.6	1.0	1.1	1.3	1.0	1.2	1.3
PopRd4	1.7	1.6	1.8	1.4	1.1	1.2	1.7	1.6	1.5
PopRd5	1.4	1.6	1.8	1.0	1.1	1.1	1.0	1.0	1.8
PopRd6	1.0	1.5	1.8	2.1	1.4	1.4	1.0	1.6	1.8
PopRd7	1.2	1.4	1.4	1.1	1.1	1.1	1.2	1.4	1.5
Mean $\overline{eff}$	3.4	3.6	3.9	1.4	1.4	1.3	1.2	1.4	1.4

Table 2: Efficiences of the ILS, Geometric, Cumulative Root Frequency and Random Search Algorithms

In a future work, we intend to implement and incorporate to the ILS algorithm more sophisticated procedures and acceptance criterion, expecting to get better solutions than those presented in this study. For example, trying to improve the local search we may adopt a VNS aproach, following the basic concepts of Ribeiro and Urrutia (2008) and Hansen and Mladenovic (2001). The generation of initial solutions, that corrently is carried out almost purely at random, may also be modified by applying some greedy construction based on previous evaluation of insertions of cut points. Finally, we intend to consider the application of a Markovian acceptance criterion, as proposed in Martin and E.W. (1991).

Acknowledgments

We would like to thank the referees for their very valuable comments and suggestions.

References

Cochran, W.: 1977, *Sampling Techniques*, John Wiley & Sons, New York.

Cyert, R.M., D. H.: 1962, *Statistical Sampling for Accounting Information*, Prentice-Hall, Englewood Cliffs, NJ.

Dalenius, J. and Hodges, J.: 1959, Minimum variance stratification, *Skandinavisk Aktuarietidskrift* **54**, 88–101.

Ekman, G.: 1959, An aproximation useful in univariate stratification, *The Annals of Mathematical Statistics* **30**, 219–229.

Glover, F. and Kochenberger, G.: 2002, *Handbook of Metaheuristics*, Kluwer Academic Publishers.

Glover, F. and Laguna, M.: 1997, *Tabu Search*, Kluwer Academic Publishers, Norwell, MA.

Gunning, P. and Horgan, J.: 2004, A new algorithm for the construction of stratum boundaries in skewed populations, *Survey Methodology* **30**, 159–166.

Gunning, P. and Horgan, J.: 2007, Improving the lavallé and hidiroglou algorithm for stratification os skewed populations, *Journal of Statistical Computation and Simulation* **77**, 277–291.

Hansen, P. and Mladenovic, N.: 2001, Variable neighborhood search: Principles and applications, *European Journal of Operational Research* **130**, 449–467.

Hedlin, D.: 1998, On the stratification of highly skewed populations, RD Report. Statistics Sweden, Sweden.

Hedlin, D.: 2000, A procedure for stratification by an extended ekman rule, *Journal of Official Statistics* **16**, 15–29.

Horgan, J.: 2006, Stratificatin of skewed populations:a review, *International Statistical Review* **74**, 67–76.

Khan, M.G.M., N. N. and N., A.: 2008, Determining the optimum strata boundary points using dynamic programming, *Survey Methodology* **34**, 205–214.

Kozak, M.: 2004, Optimal stratification using random search method in agricultural surveys, *Statistics in Transition* **6**, 797–806.

Kozak, M. and M.R., V.: 2006, Geometric versus optimization approach to stratification: A comparison of efficiency, *Survey Methodology* **32**, 157–163.

Lavallée, P. and Hidiroglou, M.: 1988, On the stratification of skewed populations, *Survey Methodology (Statistics Canada)* **14**, 33–43.

Martin, O., O. S. and E.W., F.: 1991, Large-step markov chains for the traveling salesman problem, *Complex Systems* **5**, 299–326.

Mehta, S.K., S. R. K. L.: n.d., On optimum stratification for allocation proportional to strata totals.

Ribeiro, C.C., A. D. N. T. F. R. C. and Urrutia, S.: 2008, An efficient implementation of a vns/ils heuristic for a real-life car sequencing problem, *European Journal of Operational Research* **191**, 596–611.

Rivest, L. P.: 2002, A generalization of the lavallée and hidiroglou algorithm for stratification in business surveys, *Survey Methodology (Statistics Canada)* **28**, 191–198.

Sarndal, C.E, S. B. W. J. H.: 1992, *Model Assisted Survey Sampling*, New York, Spring Verlag.

Scheaffer, R.L., M. W. O. L.: 1990, *Elementary Survey Sampling*, PWS-KENT Publishing Company.