Enhanced Cube Implementation For Highly Stratified Population

Raphaël Jauslin[a], Esther Eustache[a] and Yves Tillé[a]

Abstract

A balanced sampling design should always be the adopted strategy if auxiliary information is available. In addition, integrating a stratified structure of the population in the sampling process can considerably reduce the variance of the estimators. We propose here a new method to handle the selection of a balanced sample in a highly stratified population. The method improves substantially the commonly used sampling designs and reduces the time-consuming problem that could arise if inclusion probabilities within strata do not sum to an integer.

Key words: balanced sampling, clustered sampling, auxiliary information, unequal probability sampling

[a]Institute of statistics, University of Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Switzerland (E-mail: raphael.jauslin@unine.ch)

1 Introduction

In survey statistics, balanced sampling is a particularly efficient method when values of auxiliary variables are available for all units in the population. The idea is to select the sample so that the totals of the Horvitz-Thompson estimators of some auxiliary variables equal the population totals. There are different methods for selecting a balanced sample. Deville and Tillé (2004) have proposed the cube method which successively transforms the vector of inclusion probabilities into a sample. The method has been improved by Chauvet and Tillé (2006) by reducing the computation time.

In many areas, it is very useful to use stratified sampling designs. As already indicated by Neyman (1934), the variance of the Horvitz-Thompson estimator can be reduced by constructing strata such that the variables are homogeneous within the strata. Besides, Chauvet (2009) proposed a specific algorithm to obtain balanced samples in the strata of a population. However, this method becomes cumbersome when the number of strata is large.

A highly stratified population is very common in survey sampling. For example, it may be necessary to select individuals from a population while requiring that at most only one individual from each household in a population is taken. Each household is then a stratum. In spatial statistics, we can also construct small strata of neighboring units to obtain well-spread samples. Highly stratified sampling is also necessary for some donor imputation methods: the objective is to select a respondent for each non-respondent to impute its values. Each non-respondent then defines a stratum in which a respondent must be selected (Hasler and Tillé, 2016).

The balanced and stratified sampling method of Chauvet (2009) has been improved by Hasler and Tillé (2014) to partially resolve the disadvantage of the time required to process a highly stratified population. When the sum of the inclusion probabilities in the strata is not an integer, the computation time can become problematic. This problem arises, for example, when the objective is to select less than one individual per household. Neither of the two methods already proposed solves the computational time problem in these situations.

In this paper, we propose a new method to obtain a stratified balanced sample. This new method is particularly interesting when the population is highly stratified and the inclusion probabilities do not sum to an integer within the strata. We refer readers to Tillé (2020) and Hankin et al. (2019) to have more information on the general settings on stratified balanced sampling design.

The document is organized as follows. The section 2 gives the basic notations and settings. Section 3 presents the problem of selecting a balanced sample. In the section 4, we review the cube method and how it is used to select a balanced sample. In section 5, we discuss the issue of the highly stratified population and

review the methods used to select a sample in this case. In the section 6, we present the new method and the section 7 is devoted to variance estimation. In the section 8, we give the simulation results of the different algorithms on an artificial dataset while the section 9 gives a conclusion on the new method.

2 Basic sampling notations

Consider a finite population U of size N whose units can be defined by labels $k \in \{1, 2, \ldots, N\}$. Let define a variable of interest y. Suppose that we are trying to estimate the following unknown total:

$$Y = \sum_{k \in U} y_k. \tag{1}$$

A sampling design is defined by the probability $p(s)$ of selecting each possible subset $s \subset U$ such that $\sum_{s \subset U} p(s) = 1$. Consider a vector $\mathbf{a} = (a_1, \ldots, a_N)^\top$ that maps elements of a subset s to an N vector of 0s and 1s such that:

$$a_k = \begin{cases} 1 & \text{if } k \in s, \\ 0 & \text{otherwise,} \end{cases}$$

for $k \in U$. For each unit of the population, the inclusion probability π_k, with $0 \leq \pi_k \leq 1$, is defined as the probability of selecting k into a sample s:

$$\pi_k = \mathrm{P}(k \in s) = \mathrm{E}(a_k) = \sum_{s \subset U | k \in s} p(s).$$

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ be the vector of all the inclusion probabilities. Let also $\pi_{k\ell}$ be the probability of selecting units k and ℓ together in the sample, with $\pi_{kk} = \pi_k$. Assuming that $\pi_k > 0$ for all $k \in U$, the total (1) can be estimated using the classical unbiased Horvitz-Thompson estimator defined by

$$\widehat{Y} = \sum_{k \in U} \frac{y_k a_k}{\pi_k}. \tag{2}$$

3 Stratified balanced sampling

Usually, some auxiliary information are available for each unit $k \in U$ in a vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kq})^\top \in \mathbb{R}^q$, with $q \in \mathbb{N}$. A sampling design is said to be balanced on the q auxiliary variables if and only if it satisfies the following balancing equation:

$$\widehat{\mathbf{X}} = \sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \frac{\mathbf{x}_k a_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k = \mathbf{X}.$$

Sometimes, selecting a sample that satisfies exactly the constraints is not possible due to the rounding problem.

In many applications, inclusion probabilities are such that the selected sample has a fixed size. In order to obtain a sampling design with fixed sample size, a linear combination of the auxiliary variables must be proportional or equal to the vector of inclusion probabilities, i.e. there exists $\boldsymbol{\psi} \in \mathbb{R}^q$ such that $\boldsymbol{\psi}^\top \mathbf{x}_k = \pi_k$, for all $k \in U$. Indeed, this gives

$$\sum_{k \in s} \frac{\boldsymbol{\psi}^\top \mathbf{x}_k}{\pi_k} = \sum_{k \in s} \frac{\pi_k}{\pi_k} = n.$$

The size of the sample will be fixed only if n is an integer. If it is not the case, the sample size will be equal to the higher or lower integer to n.

More generally, the problem of selecting a balanced sample is written as the following linear system :

$$\begin{cases} \mathbf{A}^\top \mathbf{a} = \mathbf{A}^\top \boldsymbol{\pi}, \\ \mathbf{a} \in \{0,1\}^N, \end{cases} \tag{3}$$

where $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_N/\pi_N)^\top$. The aim consists then of obtaining a sample $\mathbf{a}$ that satisfies (or approximately satisfies) the constraints.

Suppose that the population U is divided into H strata $U_1, \ldots, U_H$, with respective sizes of $N_1, \ldots, N_H$. The strata form a partition and respect the following properties:

$$U = \bigcup_{h=1}^{H} U_h, \ N_h > 0, \ U_h \cap U_\ell = \emptyset, \ \text{for all } h, \ell \in \{1, \ldots, H\}.$$

Then, this implies that $\sum_{h=1}^{H} N_h = N$. The inclusion probabilities sum to a value n_h in each stratum h, i.e. $n_h = \sum_{k \in U_h} \pi_k$. Let $\mathbf{h} = (h_1, \ldots, h_N)^\top$ be a categorical vector that specifies the stratum to which each unit belongs. For example, $h_k = \ell$ means that unit k belongs to strata U_ℓ, with $k \in U$ and $\ell \in \{1, \ldots, H\}$. Another way for expressing the stratum of each unit is to use the disjunctive form. Let $\mathbf{H}$ be the disjunctive matrix of the corresponding vector $\mathbf{h}$ of size $N \times H$, such that:

$$\mathbf{H} = \big(\mathbf{1}(U_1), \ldots, \mathbf{1}(U_H)\big),$$

where $\mathbf{1}(U_h) \in \mathbb{R}^N$ is an column vector such that its kth element is equal to 1 if the unit k belongs to the stratum U_h and 0 otherwise.

Obtaining a balanced sample in a stratified population is equivalent to adding stratification constraints to the previous linear system (3). These constraints are contained in the matrix $\mathbf{H}$, so the modification of the linear problem gives:

$$\begin{cases} (\mathbf{H} \ \mathbf{A})^\top \mathbf{a} = (\mathbf{H} \ \mathbf{A})^\top \boldsymbol{\pi}, \\ \mathbf{a} \in \{0,1\}^N. \end{cases} \tag{4}$$

The number of constraints in the linear problem is then $(q + H)$. In the next section, a method to select a balanced sample is presented.

4 Cube Method

Deville and Tillé (2004) developed the cube method that selects a balanced sample respecting the inclusion probabilities. The method can deal with equal or unequal inclusion probabilities. The algorithm is separated into two phases.

- The first phase is called the flight phase. It modifies recursively and randomly the vector of inclusion probabilities $\boldsymbol{\pi}$ into a sample by respecting exactly the balancing constraints of the problem. The subspace induced by the linear system (3) could be rewritten using the following notation:

$$\mathcal{A} = \left\{\mathbf{a} \in \mathbb{R}^N | \mathbf{A}^\top \mathbf{a} = \mathbf{A}^\top \boldsymbol{\pi}\right\} = \boldsymbol{\pi} + \text{Null}(\mathbf{A}^\top),$$

  where $\text{Null}(\mathbf{A}^\top) = \left\{\mathbf{u} \in \mathbb{R}^N | \mathbf{A}^\top \mathbf{u} = 0\right\}$. The idea is then to use a vector $\mathbf{u}$ of the null space of $\mathbf{A}^\top$ in order to update randomly the vector $\boldsymbol{\pi}$. The whole procedure of the update can be found in Deville and Tillé (2004). At each step, at least one component is set to 0 or 1. Matrix $\mathbf{A}$ is updated with the new inclusion probabilities. This step is repeated until the null space of $\mathbf{A}^\top$ is empty. At the end of the flight phase, the final updated vector of $\boldsymbol{\pi}$ contains at most q elements that are still not equal to 0 or 1.

- The second phase is called the landing phase. This phase allows to obtain the sample $\mathbf{a}$ that respects as much as possible the balancing constraints. There are two different ways to achieve it, by relaxing the q constraints one by one, or by linear programming.

In the flight phase, the major computational cost comes from the research of a vector in the null space of $\mathbf{A}^\top$. Chauvet and Tillé (2006) have improved this time-consuming inconvenience using a submatrix of $\mathbf{A}$ rather than the entire matrix. The idea is to consider a submatrix that has one more row than the number of columns to ensure to have at least one vector in its null space. This submatrix, denoted by $\mathbf{B}$, has then a size of $(q+1)\times q$, with respect to $q < N$ and $\text{Rank}(\mathbf{B}) \leq q$.

The interest of using this submatrix comes from the following result: a vector $\mathbf{u}$ of $\text{Null}(\mathbf{B}^\top)$ completed by $(N - (q + 1))$ zeros is a vector of $\text{Null}(\mathbf{A}^\top)$. With this idea, all the computations can be done using only a submatrix $\mathbf{B}$. Usually, N is much larger than q, the size of $\mathbf{B}$ is then much smaller than $\mathbf{A}$. This implies obviously an important gain of computational time. The method proposed in this paper uses the same idea. In the next section, the particular case of highly stratified sampling is considered.

5 Highly stratified population

It is always preferable to consider a stratified population in order to estimate the total (1). Indeed, the variance of the Horvitz-Thompson estimator (2) can be considerably reduced compared to the non-stratified estimator (1). However, when the population is highly stratified (i.e. H is very large), the selection of a balanced sample with classical methods becomes difficult due to the too large number of constraints in $\mathbf{H}$. In order to decrease the time-consuming problem, different approaches have already been proposed.

Chauvet (2009) has developed an algorithm to select a balanced sample in a highly stratified population. Firstly, a flight phase is applied inside each stratum. This allows modifying the inclusion probabilities such that these are as balanced as possible in each stratum. Next, a flight phase is applied to the whole population. Finally, a landing phase is carried out on units that are not still selected or rejected. This procedure has the advantage to be simple to implement. Its major deficiency is when the number of strata H becomes too large. The procedure remains then very slow and often cannot even be used.

Hasler and Tillé (2014) have proposed another method to deal with highly stratified population. As the previous method, it begins by applying the flight phase of the cube method to each stratum of the population. Next, it carries out a flight phase on an union of strata by adding another stratum at each step. By doing this, strata are managed one after the other and the inclusion probabilities of certain strata are set to 0 or 1 during this step. The idea behind this procedure is to reduce the matrix $\mathbf{H}$ considered because some strata are removed from the matrix when all its units are selected or rejected. At the end, a landing phase is applied. However, if n_h is not equal to an integer for a stratum U_h, this method also remains very time-consuming. Indeed, some strata are never completely removed during the procedure and then the submatrix of $\mathbf{H}$ considered becomes too large.

The properties of the cube method imply that the inclusion probabilities are satisfied and that the sample is balanced on the auxiliary variables in these two methods. However, they still have difficulty to deal with all the situations of highly stratified sampling. In the next following section, a new method is presented in order to completely resolve these drawbacks.

6 Proposed method

In the fast implementation of the cube method proposed by Chauvet and Tillé (2006), the main modification was to use a matrix smaller than $\mathbf{A}$ to update $\boldsymbol{\pi}$. This allows to considerably decrease the computational cost. The idea of our method is similar but adapted to a stratified population. It considers a matrix

of constraints $\mathbf{B}$ smaller than $(\mathbf{H}\ \mathbf{A})$ during the use of the cube method. The submatrix $\mathbf{B}$ must be found at each step of the flight phase of the cube method. As explained in Section 3, the number of balancing constraints depends on the number of strata H when the population is stratified. Indeed, since all the strata must be taken into account, the number of balancing equations is equal to $(q+H)$. So, in the classical flight phase presented in Section 4, the considered matrix $\mathbf{B}$ is of size $(q+H+1)\times(q+H)$. The columns of $\mathbf{B}$ corresponding to strata which do not contain any unit in the rows of $\mathbf{B}$ are only composed of 0. To update the vector $\boldsymbol{\pi}$, we find a vector in the nullspace of $\mathbf{B}$. Since columns that are only equal to 0 are inefficient for this, they only increase the size of the matrix $\mathbf{B}$ and are irrelevant for finding a null vector.

The idea of the proposed method is then to compute a matrix $\mathbf{B}$, with still one more row than the number of columns, but that considers a smaller number of strata. By considering a matrix $\mathbf{B}$ with fewer rows, the corresponding vector of strata $\mathbf{h}$ will be reduced. This subvector of $\mathbf{h}$ will contain fewer categories and then the corresponding matrix $\mathbf{H}$ will have fewer columns. This is why obtaining the matrix $\mathbf{B}$ with exactly one row more than its number of columns is not as easy as with an unstratified population. Algorithm 1 shows how to find the number of rows to consider in order to obtain the smaller matrix $\mathbf{B}$ such that $\mathbf{B}$ has exactly one row more than its number of columns.

Algorithm 1 Find the submatrix $\mathbf{B}$ of $(\mathbf{H}\ \mathbf{A})$

Let q be the number of auxiliary variables of $\mathbf{A}$. Initialize q^1 by q. For $t = 1, 2, 3, \ldots$ repeat the following steps:

1. Extract the first q^t rows of the vector $\mathbf{h}$ and denote it $\mathbf{h}^t$.
2. Denote H^t the number of different strata in $\mathbf{h}^t$.
3. Update $q^{t+1} = q + H^t + 1$.

while $q^{t+1} > q^t$.
Finally, $\mathbf{B}$ is defined as the q^t first rows of the concatenated matrix $(\mathbf{H}^t\ \mathbf{A}^t)$, where $\mathbf{A}^t$ and $\mathbf{H}^t$ are the submatrix containing only its q^t first rows.

Example 6.1 *Suppose that* $q = 2$ *and that the categorical vector is equal to* $\mathbf{h} = (1,$

1, 2, 2, 3, 3, 3, 4, 4, ...)$^\top$. *We obtain*

$$
\begin{array}{llll}
t=1: & q^1=2, & \mathbf{h}^1=(1,1)^\top, & H^1=1 \rightarrow q^2=2+1+1=4,\\
t=2: & q^2=4, & \mathbf{h}^2=(1,1,2,2)^\top, & H^2=2 \rightarrow q^3=2+2+1=5,\\
t=3: & q^3=5, & \mathbf{h}^3=(1,1,2,2,3)^\top, & H^3=3 \rightarrow q^4=2+3+1=6,\\
t=4: & q^4=6, & \mathbf{h}^4=(1,1,2,2,3,3)^\top, & H^4=3 \rightarrow q^5=2+3+1=6,\\
t=5: & q^5=q^4 & &
\end{array}
$$

$\mathbf{B}$ *contains then* $q^4=6$ *rows and* $5=2+3$ *columns. So it is a matrix with only one more row than the number of columns as desired.*

The matrix $\mathbf{B}$ is found after having computed its number of rows q^t using Algorithm 1. The first q^t elements of $\mathbf{h}$ composed the strata membership vector $\mathbf{h}^t$. The disjunctive matrix $\mathbf{H}^t$ can then be found using $\mathbf{h}^t$. The matrix $\mathbf{B}$ is equal to $(\mathbf{H}^t\ \mathbf{A}^t)$, with $\mathbf{A}^t$ the submatrix of $\mathbf{A}$ containing only its q^t first rows. The units must be ordered in such a way that the strata are clustered. If it is not the case, Algorithm 1 ends up with a matrix that could be large. Indeed, the larger the value $\mathbf{H}^t$ is, the larger the size of the matrix $\mathbf{B}$ will be. The same procedure proposed by Chauvet and Tillé (2006) can then be applied. If the population is highly stratified and the number of auxiliary variables is acceptable, our procedure can be very efficient. Moreover, it handles inclusion probabilities that do not sum to an integer inside strata. Algorithm 2 presents the whole method and is implemented in an R package (Jauslin et al., 2021).

The Algorithm can be decomposed into two flight phases and two landing phases. The first step is to apply a flight phase within each stratum. This step modifies the inclusion probabilities so that they are balanced in each stratum. In the second step, we first check that there are no strata containing only one unit. In this case, the balancing equations of these strata cannot be balanced, so we remove these units from the treatment in order to manage them afterwards. Next, we apply the flight phase on the other units by using the matrix $\mathbf{B}$ found by the Algorithm 1. This second flight phase is repeated until the matrix $\mathbf{B}$ is no longer found or the null space is empty. Note that the dimension of the computed matrix $\mathbf{B}$ can change depending on the stratum vector $\mathbf{h}$ and the inclusion probabilities set to 0 or 1. At the end of this flight phase, we can have H units that are alone in their stratum and also have maximum q units for which the balancing equations could not be satisfied. It is then no longer possible to find a matrix $\mathbf{B}$ such that its kernel is not empty. Note that at the end of this step, the auxiliary variables are still completely satisfied, i.e. the equation (4) with the modified inclusion probabilities is perfectly satisfied.

The objective is to respect the strata sizes as much as possible. Therefore, the constraints on the strata are prioritized. We then perform a landing phase by suppression of variables on the units that are in the strata containing at least two

units. At the end of this first landing phase, we have at most H units which are alone in their strata and thus the corresponding disjunctive matrix is a diagonal matrix. We can drop the variables corresponding to the strata and apply a landing phase only on the auxiliary variable to finally find the sample.

If the vector $\mathbf{h}$ is reorder in a different way before the procedure is started, the matrix $\mathbf{B}$ could change during the execution of the Algorithm 2. This has no impact on the computational cost if the $\mathbf{h}$ vector is still clustered. Figure 1 shows an example of what the matrix $\mathbf{B}$ looks like in the Chauvet and Hasler methods and the one computed by the Algorithm 1.

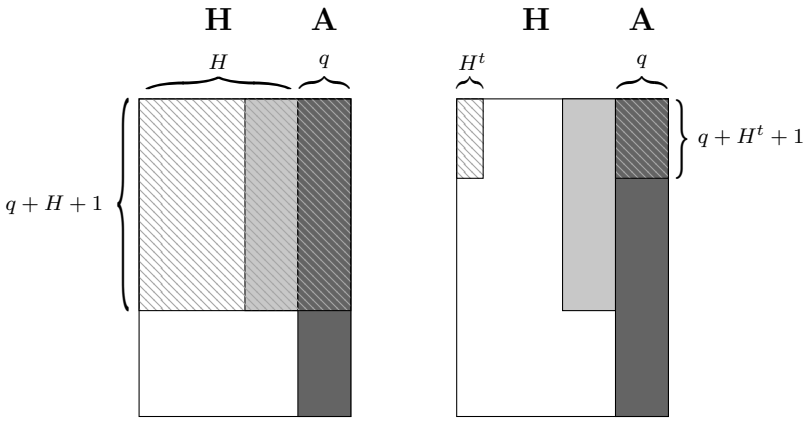

Figure 1: Illustration of how the $\mathbf{B}$ matrix is calculated within the flightphase in the discussed methods. The dark grey area represents the auxiliary variable matrix $\mathbf{A}$ while the white area is the $\mathbf{H}$ matrix. The $\mathbf{B}$ matrix is illustrated by the hatched area. The light grey area is the one inside of the matrix $\mathbf{B}$ where all values are equal to zero. On the left, we see how the matrix $\mathbf{B}$ is computed in Chauvet and Hasler methods, while on the right, it is the matrix proposed by Algorithm 1. We see that the matrix $\mathbf{B}$ on the left contains some columns that are only equal to zero.

7 Variance estimation

The variance can be approximated by using the method proposed by Deville and Tillé (2005). Let the vector

$$\mathbf{z}_k = (\mathbf{H} \quad \mathbf{A})_k,$$

where $(\mathbf{H} \quad \mathbf{A})_k$ denote the kth row of the matrix $(\mathbf{H} \quad \mathbf{A})$. The variance of the Horvitz-Thompson estimator of the total $\widehat{Y}$ can be approximated by

$$\text{var}_{app}(\widehat{Y}) = \sum_{k \in U} c_k \left(\frac{y_k}{\pi_k} - \boldsymbol{\alpha}^\top \mathbf{z}_k \right)^2, \tag{5}$$

where

$$c_k = \pi_k(1-\pi_k)\frac{N}{N-(H+q)} \text{ and } \boldsymbol{\alpha} = \left(\sum_{\ell \in U} c_\ell \mathbf{z}_\ell \mathbf{z}_\ell^\top \right)^{-1} \sum_{\ell \in U} c_\ell \mathbf{z}_\ell \frac{y_\ell^\top}{\pi_\ell}.$$

There exists many different ways to express the quantity c_k and then this leads to various approximations of the variance. Value c_k can in particular be approximated by

$$\widetilde{c}_k = (1-\pi_k)\frac{n}{n-(H+q)}.$$

We define the estimator of the approximated variance as the following equation:

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{k \in s} \widetilde{c}_k \left(\frac{y_k}{\pi_k} - \widetilde{\boldsymbol{\alpha}}^\top \mathbf{z}_k \right)^2. \tag{6}$$

Note that the sum in the equation (6) is on the sample s using $\widetilde{c}_k$ and $\widetilde{\boldsymbol{\alpha}}$ instead of c_k and $\boldsymbol{\alpha}$.

8 Simulations

In this section, the performance of the method is evaluated on real data produced by the Swiss Federal Statistical Office (2020). The dataset contains information on Swiss establishments. We restrict the study to the Swiss region called Espace Mittelland (a region of the second degree of the Nomenclature of Territorial Units for Statistics (NUTS) of Switzerland). This region contains 5 cantons (a region of the third degree of the NUTS) and 675 municipalities. For confidentiality reasons, the units considered are the hectares of land in which at least one establishment is located. In order to be able to estimate the variance, only 3 hectares of land per municipalities are included in the study. This implies that the dataset contains information from 2025 hectares including at least one establishment.

We stratify the units in two different ways: by cantons and by municipalities. The number of strata is then respectively equal to $H_c = 5$ and $H_m = 675$. Figure 2 shows the dataset with the two proposed stratifications. The idea behind this procedure is to compare the execution time for a stratified population with a

low number of strata versus a high one. To compare the method, we will use balancing variables $\mathbf{x}_j$ containing the number of women employed in a sector j, with $j = 1, 2, 3$. Each sector represents a type of activities: sector $j = 1$ involves the natural environment and agriculture, sector $j = 2$ is manufacture and sector $j = 3$ is related to services. Table 1 shows the mean time execution of three methods for highly stratified sampling: the methods of Hasler and Tillé (2014) and Chauvet (2009), detailed in Section 5, and the new one presented in this article.

Inside each stratum, the inclusion probabilities are equal. For each stratification, we carry out two different sampling: one with inclusion probabilities that sum to an integer number within each stratum and one with a non-integer sum. Then, we consider $n_h = 80$ and $n_h = 80.4$, $h \in \{1, \ldots, H_c\}$, for the first stratification. For the second one, we take $n_h = 2$ and $n_h = 1.4$, $h \in \{1, \ldots, H_m\}$. We choose to deal also with non-integers n_h with the aim to compare the impact of this situation on the mean sampling time.

Chauvet's method cannot be compared because its execution time is too long and should be avoided for highly stratified population. However, it remains very efficient if the number of strata is acceptable. If inclusion probabilities in stratum sum to an integer, the Hasler's method performs very well. However, the execution time increases strongly when n_h is not an integer. The proposed method has a very well behaved and the time is considerably reduced for a highly stratified population.

In order to compare the variance of the method with the others, we estimate the variance using some variables of interest y_j that contains the total number of employees of the sector j, $j = 1, 2, 3$. In Table 2, we compare the approximated variance (5), the estimated variance and the simulated variance computing using the equation:

$$v_{sim} = \frac{1}{m} \sum_{s} \{\widehat{Y}(s) - Y\}^2, \tag{7}$$

where m is the number of simulations.

For each method, we vary the number of selected units within each stratum by taking n_h equals to 2 for the stratification with H_c strata and 80 for the stratification with H_m strata. This implies sample of size respectively equal to 400 and 1350. The variance estimator seems to be unbiased for the approximated variance. However, we see that the approximated variance and estimator are slightly biased to the v_i. This comes from the landing phase of each method. We can conclude that the proposed method is comparable in terms of variance to other methods.

Table 1: Results of 1000 simulations on the Swiss establishments dataset. The population size is equal to 2025. We compute the mean time execution in seconds of each sampling procedure. We vary the number of strata H and the number of units selected within each stratum n_h.

	Algorithm		
	Proposed method	Hasler's method	Chauvet's method
Cantons ($H = 5$)			
$n_h = 80$	0.24	0.25	0.24
$n_h = 80.4$	0.24	0.24	0.24
Municipalities ($H = 675$)			
$n_h = 2$	0.42	0.4	418.07
$n_h = 1.4$	0.53	400.55	701.77

Table 2: Results of 1000 simulations on a population of size 2025. The number of strata is equal to 5 for Cantons and 675 for Municipalities. For each variable of interest y_j, $j = 1, 2, 3$ and for each sampling methods, we compute the ratio between the different estimators (i.e. approximated variance (5) as well as the variance estimator (6)) and the variance approximated by the simulations (7).

	Algorithm								
	Proposed method			Hasler's method			Chauvet's method		
	v_{sim}	$\widehat{\text{var}}(\widehat{Y})$	$\text{var}_{app}(\widehat{Y})$	v_{sim}	$\widehat{\text{var}}(\widehat{Y})$	$\text{var}_{app}(\widehat{Y})$	v_{sim}	$\widehat{\text{var}}(\widehat{Y})$	$\text{var}_{app}(\widehat{Y})$
Cantons ($H = 5$)									
y_1	1	0.945	0.959	1	0.994	1.019	1	1.064	1.08
y_2	1	0.905	0.969	1	0.932	1.012	1	0.927	0.983
y_3	1	0.891	0.927	1	0.864	0.911	1	0.84	0.879
Municipalities ($H = 675$)									
y_1	1	1.042	1.049	1	0.952	0.957	1	0.978	0.981
y_2	1	1.024	1.028	1	1.035	1.036	1	0.951	0.954
y_3	1	0.995	0.994	1	0.933	0.937	1	0.849	0.853

9 Conclusion

The stratified sampling procedure is a well-known and appropriate procedure to reduce the variance of the Horvitz-Thompson estimator. In this paper, we propose a new method and implementation that provide an excellent executing time and a flexibility that the existing methods did not allow. In many surveys where the

population is stratified, the sum of inclusion probabilities within each stratum is not an integer. Other methods are not directly applicable in this case. We have shown by mean of simulations that the variance of the estimator is not impacted by our method. All of these results indicate that our proposed algorithm is very efficient to select a sample in a stratified and highly stratified population.

References

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35:115–119.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21:53–62.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:569–591.

Hankin, D., Mohr, M., and Newman, K. (2019). *Sampling Theory: For the Ecological and Natural Resource Sciences*. Oxford University Press, New York.

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81–94.

Hasler, C. and Tillé, Y. (2016). Balanced k-nearest neighbor imputation. *Statistics*, 105:11–23.

Jauslin, R., Eustache, E., and Tillé, Y. (2021). *Different methods for stratified sampling*. R package version 0.1.0.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Swiss Federal Statistical Office (2020). *Statistique structurelle des entreprises (STATENT) Description des données GEOSTAT*. Neuchâtel, Switzerland.

Tillé, Y. (2020). *Sampling and estimation from finite populations*. Wiley, New York.

Algorithm 2

Consider $\boldsymbol{\pi}$ the N vector of inclusion probabilities such that $0 < \pi_k < 1$, for $k \in \{1, \ldots, N\}$.

I. Perform a flight phase on each stratum according to the inclusion probabilities $\boldsymbol{\pi}$ and the balancing constraints in $\mathbf{A}^\top$. The vector $\boldsymbol{\pi}$ is updated by $\boldsymbol{\pi}^1$ such that some of its elements are set to 0 or 1. Compute the set of indices $\mathbf{i}^1 \subset \{1, \ldots, N\}$ containing the unit indices that have an inclusion probability still not equal to 0 or 1 and that are not alone in their strata. Define the set of units that are alone in their strata and denote this set $\widetilde{\mathbf{i}}^t$.

II. Initialize t by 1. Repeat step 1. to 7. until it is no more possible to find the matrix $\mathbf{B}$ or until the vector $\mathbf{u}$ is null.

1. Update the set $\widetilde{\mathbf{i}}^t$ by adding the indices of the units that are contained in only one strata and remove it from $\mathbf{i}^t$.
2. In $\mathbf{A}$, $\mathbf{h}$ and $\boldsymbol{\pi}$, consider only units with indices in $\mathbf{i}^t$.
3. Apply the Algorithm 1 to find the submatrix $\mathbf{B}$ of $(\mathbf{H}\ \mathbf{A})$.
4. Compute $\mathbf{u}$, a vector of the null space of $\mathbf{B}$ completed by 0s to obtain a vector with the same size as $\mathbf{i}^t$.
5. Compute $\lambda_1 > 0$ and $\lambda_2 > 0$, the two larger values such that

$$\begin{array}{rcl} 0 & \leqslant & \pi_k^t + \lambda_1 u_k \leqslant 1 \\ 0 & \leqslant & \pi_k^t - \lambda_2 u_k \leqslant 1 \end{array} \text{, for all } k \in \mathbf{i}^t.$$

6. Update $\boldsymbol{\pi}^t$ by:

$$\boldsymbol{\pi}^{t+1} = \begin{cases} \boldsymbol{\pi}^t + \lambda_1 \mathbf{u} & \text{with probability} \quad \lambda_2/(\lambda_1 + \lambda_2), \\ \boldsymbol{\pi}^t - \lambda_2 \mathbf{u} & \text{with probability} \quad \lambda_1/(\lambda_1 + \lambda_2). \end{cases}$$

7. Update t by $t+1$ and update $\mathbf{i}^t$ the set of indices containing the unit that have an inclusion probability still not equal to 0 or 1.

III. Pool the set $\widetilde{\mathbf{i}}^t$ and the remaining units $\mathbf{i}^t$ and do a landing phase until we have only one unit alone in their strata. This step is done by suppression of variables on the balancing matrix $(\mathbf{H}\ \mathbf{A})$, but only dropping the variables that are in $\mathbf{A}$.

IV. Perform a landing phase by suppression of variables on the balancing variables $\mathbf{A}$ on the remaining units.

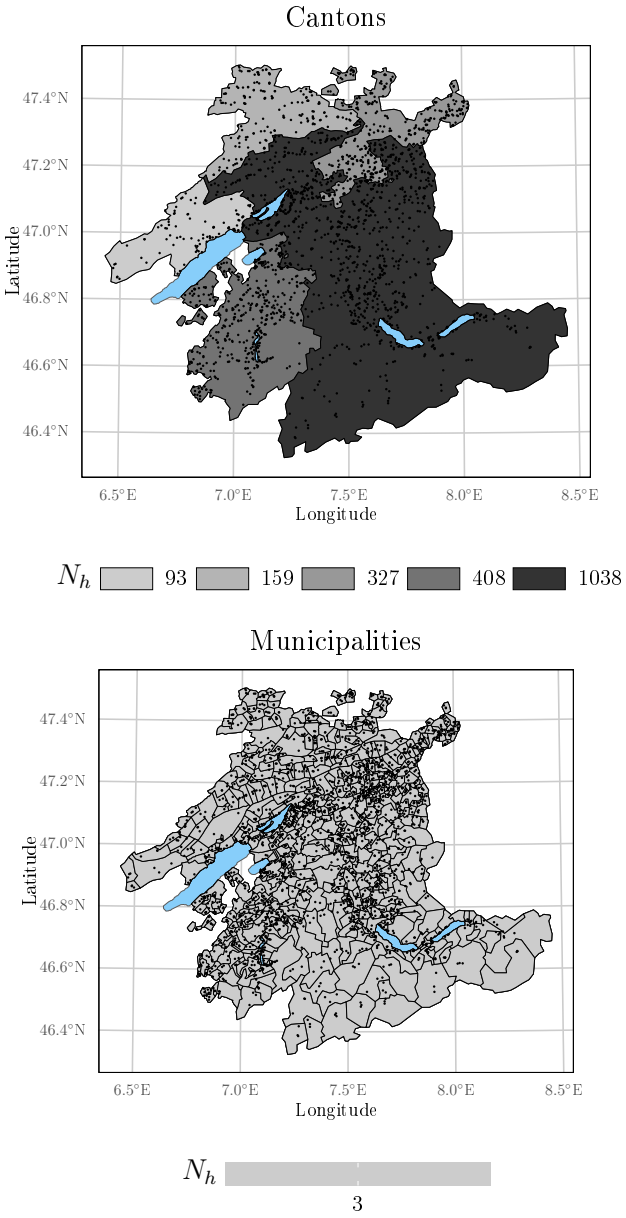

Figure 2: Data extracted from the Swiss establishments data base of the Swiss Federal Statistical Office (2020). The data are restricted to the NUTS region 2. The left plot is showing the separation by Cantons $H_c = 5$, the right one the separation by Municipalities $H_m = 675$. The grey gradient scale gives the number of units considered in each Canton. The data are selected such that each municipality contains 3 units.