Published in The Case Studies in Business, Industry and Government Statistics (CSBIGS) journal 1, issue 2, 173-182, 2007
which should be used for any reference to this work

Application of Fast SAS Macros for Balancing Samples to the Selection of Addresses

Guillaume Chauvet

Yves Tillé

This article introduces the method of balanced sampling and explains how to apply it to the selection of addresses in the context of the French census. SAS macros which perform the analysis are presented. The exposition is accessible to readers with at least an intermediate statistics level. Prior exposure to sampling ideas is useful but not strictly necessary.

1. Introduction

Balanced sampling consists in drawing random samples which provide exact estimations for auxiliary variables. When samples are drawn with equal probabilities, balanced sampling ensures that proportions known over the whole population are also respected on the sample. The method has been used for several statistical problems. In the French renovated census (see Dumais and Isnard, 2000; Bertrand et al., 2004), the rotation groups are drawn by balanced sampling. Examples related to these samples will be given in the paper, with an illustration of the gain in accuracy induced by balanced sampling. The method has also been used for the selection of the Master sample for household surveys (see Caron and Christine, 2005).

The Cube method (see Deville and Tillé, 2004, 2005) enables us to select balanced samples, with unequal probabilities and a non-restricted number of balancing variables. In this paper, we contribute SAS macros corresponding to the implementation of the fast algorithm proposed in Chauvet and Tillé (2006), with an application to sampling designs close to the ones used by the French census.

The paper is organised as follows. In Section 2, a reminder of the principles of the Cube Method is given. In Section 3, the SAS macro EXE_CUBE using the fast algorithm for balanced sampling is presented with an application to the selection of addresses. In Section 4, we introduce a particular technique for balanced sampling, called stratified balanced sampling. The SAS macro ECHANT_STRAT, which performs stratified balanced sampling, is presented with examples in Section 5. The SAS programs used in the different examples are given in the Appendix.

2. Balanced sampling : the Cube method

Consider a finite population U of size N whose units can be identified by labels $k \in \{1, \cdots, N\}$. The goal is to estimate the total $Y = \sum_{k \in U} y_k$ of a variable of interest y

that takes values $y_k, k \in U$ for the units of the population. Suppose also that the vectors of values $\mathbf{x_k} = \left(x_{k1} \cdots x_{kj} \cdots x_{kp}\right)'$ taken by p auxiliary variables are known for all the units of the population. The p vectors $\left(x_{1j} \cdots x_{kj} \cdots x_{Nj}\right)', j = 1, \ldots, p$ are assumed without loss of generality to be linearly independent.

A sample is denoted by a vector $s = \left(s_1 \cdots s_k \cdots s_N\right)'$ of R^N where s_k takes the value 1 if k is in the sample and is 0 otherwise. A sampling design $p(.)$ is a probability distribution on the set $\Re = \{0,1\}^N$ of all the possible samples. The random sample $S = \left(S_1 \cdots S_k \cdots S_N\right)'$ is a random vector of R^N that takes the value s with probability $\Pr(S = s) = p(s)$. The inclusion probability of unit k is the probability $\pi_k = \Pr(S_k = 1)$ that unit k is in the sample, and the vector of inclusion probabilities is $\pi = \left(\pi_1 \cdots \pi_k \cdots \pi_N\right)'$. Note that

$$\pi = E(S) = \sum_{s \in \Re} p(s) s \in \mathrm{R}^{\mathrm{N}}.$$

Note also that the joint inclusion probability $\pi_{kl} = \Pr(S_k = 1 \text{ and } S_l = 1)$ is the probability that two distinct units are jointly in the sample. The Horvitz-Thompson estimator given by $\hat{Y} = \sum_{k \in U} S_k \, y_k / \pi_k$ is an unbiased estimator of Y. The Horvitz-Thompson estimator of the j^{th} auxiliary total $X_j = \sum_{k \in U} x_{kj}$ is $\hat{X}_j = \sum_{k \in U} S_k \, \mathbf{x_k} / \pi_k$. The Horvitz-Thompson estimator vector,

$$\hat{\mathbf{X}} = \sum_{k \in U} S_k \, \mathbf{x_k} / \pi_k$$

estimates without bias the totals of the auxiliary variables, $\mathbf{X} = \sum_{k \in U} \mathbf{x_k}$.

A sampling design is said to be balanced on the variables $\mathbf{x} = \left(x_1 \cdots x_j \cdots x_p\right)'$ if it satisfies the balancing equations given by

$$\hat{\mathbf{X}} = \mathbf{X}.$$

The cube method proposed by Deville and Tillé (2004) provides a general solution to the problem of selecting balanced samples, with any vector of inclusion probabilities and a non-restricted number of balancing variables. A sampling design balanced on the variable $x_k = \pi_k$ is of fixed size, as

$$\begin{aligned} \sum_{k \in S} \frac{x_k}{\pi_k} &= \sum_{k \in S} 1 = n(S) \\ = \sum_{k \in U} x_k &= \sum_{k \in U} \pi_k = n. \end{aligned}$$

Sampling of fixed size with unequal probabilities is thus a particular case of balanced sampling. A sampling design balanced on the variable $x_k = 1$ ensures that the population size N is perfectly estimated, as

$$\begin{aligned} \sum_{k \in S} \frac{x_k}{\pi_k} &= \sum_{k \in S} \frac{1}{\pi_k} = \hat{N} \\ = \sum_{k \in U} x_k &= \sum_{k \in U} 1 = N, \end{aligned}$$

while $\hat{N}$ is generally random. Note that these two variables are always available in the population. By comparison with an unequal probability sampling design of fixed size, a sampling design balanced on the variables $\mathbf{x}_k = \left(\pi_k \quad 1\right)'$ induces a large gain in accuracy for a variable of interest y_k which is well explained by π_k and 1. More generally, a sampling design causes a large gain in accuracy for a variable of interest y_k which is well explained by the balancing variables $\mathbf{x}_k$.

Nevertheless, in most cases, an exact balanced sampling design does not exist. Deville and Tillé (2004) provide a solution to select approximately balanced samples, with respect to exact inclusion probabilities. Deville and Tillé (2005) give variance approximations for balanced samples, and Chauvet and Tillé (2006) present a very fast implementation of the Cube method. The SAS macros presented in the following sections are based on the algorithm of Chauvet and Tillé (2006). For a detailed presentation of the Cube method, see Deville and Tillé (2004), Rousseau and Tardieu (2004) or Tillé (2006).

We give below a small example of exact balanced sampling in order to give an overview of the progress of the Cube method.

Example: Exact balanced sampling

Consider a population of 5 units. We want to select a random sample, with respect to inclusion probabilities $\pi = \left(0.2 \quad 0.3 \quad 0.6 \quad 0.4 \quad 0.5\right)'$, balanced on the variables given by $x_1 = \pi = \left(0.2 \quad 0.3 \quad 0.6 \quad 0.4 \quad 0.5\right)'$ and $x_2 = \left(0.2 \quad 0.3 \quad 0 \quad 0 \quad 0.5\right)'$. Note that balancing on the variable π ensures that the sampling design is of fixed size, and that a sample of size $\sum_{k \in U} \pi_k = 2$ is selected.

The matrix of constraints is

$$A = \begin{pmatrix} x_1' \\ \pi' \\ x_2' \\ \pi' \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

We initialize with $\pi^0 = \pi = (0.2 \quad 0.3 \quad 0.6 \quad 0.4 \quad 0.5)'$.

First Step:

Let u be a vector in $Ker(A)$ (that is, such that $Au=0$), say $u = (1 \quad -1 \quad 0 \quad 0 \quad 0)'$. Then

$$\pi^1 = \begin{cases} \pi^0 + 0.3\,u = (0.5 \quad 0 \quad 0.6 \quad 0.4 \quad 0.5) & \text{with probability } \frac{0.2}{0.2+0.3} = 0.4 \\ \pi^0 - 0.2\,u = (0 \quad 0.5 \quad 0.6 \quad 0.4 \quad 0.5) & \text{with probability } \frac{0.3}{0.2+0.3} = 0.6 \end{cases}$$

The coefficients 0.3 and -0.2 are computed so that at least one component of the vector is put to 0 or 1. Note that adding a vector in $Ker(A)$ to π^0 ensures that the balancing equations are satisfied, and the probabilities used to determine π^1 ensure that the inclusion probabilities π are satisfied.

We assume that the randomization gives $\pi^1 = (0.5 \quad 0 \quad 0.6 \quad 0.4 \quad 0.5)'$. Since the second component equals 0, the second unit of the population is not included in the sample.

Second Step:

As the second unit of the population is definitely rejected, let v be a vector in $Ker(A)$ with a second component equal to 0, say $v = (0 \quad 0 \quad 1 \quad -1 \quad 0)'$. Then

$$\pi^2 = \begin{cases} \pi^1 + 0.4\,v = (0.5 \quad 0 \quad 1 \quad 0 \quad 0.5) & \text{with probability } \frac{0.6}{0.4+0.6} = 0.6 \\ \pi^1 - 0.6\,v = (0.5 \quad 0 \quad 0 \quad 1 \quad 0.5) & \text{with probability } \frac{0.4}{0.4+0.6} = 0.4 \end{cases}$$

Once again, the coefficients 0.4 and -0.6 are chosen so that at least one component of the vector is put to 0 or 1.

We assume that the randomization gives $\pi^2 = (0.5 \quad 0 \quad 1 \quad 0 \quad 0.5)'$. As the third component (respectively the fourth component) equals 1 (respectively 0), the third unit of the population is included in the sample (respectively the fourth unit is rejected from the sample).

Third Step:

As we already decided to include or not units 2, 3 and 4 in the sample, let w be a vector in $Ker(A)$ with its second, third and fourth components equal to 0, say $w = (1 \quad 0 \quad 0 \quad 0 \quad -1)'$. Then

$$\pi^3 = \begin{cases} \pi^2 + 0.5\,w = (1 \quad 0 \quad 1 \quad 0 \quad 0) & \text{with probability } \frac{0.5}{0.5+0.5} = 0.5 \\ \pi^2 - 0.5\,w = (0 \quad 0 \quad 1 \quad 0 \quad 1) & \text{with probability } \frac{0.5}{0.5+0.5} = 0.5 \end{cases}$$

We assume that the randomization gives $\pi^3 = (0 \quad 0 \quad 1 \quad 0 \quad 1)'$. The sample is composed of units 3 and 5. We can easily verify that the balancing equations are exactly respected.

3. A fast macro for balanced sampling

In this section, a fast macro for drawing balanced samples is proposed and examples related to the selection of addresses by the French census are given. In the French big municipalities (10 000 inhabitants or more), a sample of 1/5 of the addresses is selected each year by balanced sampling with equal probabilities, see Example 1. The variance may be strongly reduced by selecting a sample with probabilities proportional to a measure of size, such as the number of households at each address, see Example 2. In the French renovated census, the addresses of each big municipality are split into five non overlapping rotation groups by means of coordinated balanced sampling (Tillé and Favre, 2004). One rotation group is surveyed each year, so that each address is taken into account in a span of five years. An illustration of coordinated balanced sampling is given in Example 3.

3.1. Description

The EXE_CUBE macro enables us to select a balanced sample and returns a data table containing the result of the sampling. This macro may be found on the website (http://www2.unine.ch/statistics/page10891.html) of the Neuchatel University, and is provided with this article.

3.2. The Input Data

The data relative to the population from which we want to select a balanced sample must be put into a SAS table, containing all the units of the population, and at least:

- an identifying variable,
- the variable with inclusion probabilities,
- the balancing variables.

This table may not contain missing values for these variables. The variable with inclusion probabilities, as well as the balancing variables, must be of numerical type.

3.3. Syntax of the macro

3.3.1. Parameters relative to the Database

All these parameters are compulsory.

- BASE = name of SAS library
  Name of the SAS library containing the SAS table of Input data.

- DATA = name of SAS table
  Name of the SAS table containing the Input data.

- ID = variable
  Name of the variable that identifies the units of the population

- PI = variable
  Name of the variable with the inclusion probabilities

- CONTR = variable(s)
  Names of the variables on which the sample will be balanced. The names must be spaced with blanks.

3.3.2. Parameters relative to the sampling and Output

- ATTER = option
  States the option selected for the landing phase (see Tillé, 2006 p. 159 for a definition of flight and landing phases). Possible values are:

  - ATTER = 1
    The balancing variables are progressively abandoned. The last variable in the CONTR parameter is removed first, then the previous variable and so on.

  - ATTER = 2
    The landing phase is performed by considering all the possible samples among the remaining units, and selecting preferably those providing a low difference to the balancing.

  - ATTER = 3
    The landing is performed as with ATTER=2, but only considering the samples whose size equals the sum of inclusion probabilities. We obtain a fixed sample size. If this option is used, the variable of inclusion probabilities must be put in the CONTR parameter.

  This parameter is optional. The default value is: ATTER=1. This is the fastest landing option. To ensure a reasonable execution time, the option ATTER=2 should not be used with more than 14 balancing variables, and the option ATTER=3 should not be used with more than 18 balancing variables.

- COMPEQ = option

  Equals 1 if the complement of the sample has to be balanced on the same variables as well, and 0 otherwise. This parameter is optional. The default value is: COMPEQ=0.

  Here we use a result of Tillé and Favre (2004). The proof can be found in Chauvet (2006). This option allows for the selection of several non-overlapping samples, balanced on the same variables, with fixed inclusion probabilities. Suppose we want to select two non-overlapping samples, balanced on the variable x, with inclusion probabilities π_k. We select the first balanced sample S_1 as usual, with option COMPEQ=1. Then we select a sample S_2 in the complement of S_1, with inclusion probabilities $\frac{\pi_k}{1-\pi_k}$, balanced on the variable $(z_k)=\left(\frac{x_k}{1-\pi_k}\right)$. This method can be applied to any number of balancing variables. We can select up to $Min_{k\in U}\left[\frac{1}{\pi_k}\right]$ balanced samples with this method, where $[x]$ is the largest integer smaller than x.

  For example, this technique enables us to randomly split a population into non-overlapping samples; in the French renovated census, the municipalities of each French region are randomly split into five samples, called rotation groups, one of them being fully investigated each year. This option multiplies by 2 the number of balancing variables, thus by about 4 the execution time. If all inclusion probabilities are equal, the complement of the sample is automatically balanced on the same variables, so the option becomes useless (see Chauvet, 2006, for details).

- SORT = name of SAS table

  Name of the SAS table containing the Output data. This table belongs to the library quoted in BASE. It

contains all the units of the population, and the variable ECH equal to 1 if the unit has been selected in the sample, and 0 otherwise. This parameter is compulsory.

3.3.3. Some numerical examples

We use an artificial population of 26,471 units, corresponding to a municipality; the dataset is provided with this article. The samples are selected by means of a personal computer (Penthium 4, 1.8 Gh).

Example 1. We first select a sample with equal probabilities 1/5, balanced on the socio-demographic variables quoted in Table 1 (18 variables) and on a constant for the condition of fixed sample size. We use the first landing option. A sample of 5,294 units is drawn in a few seconds. Results are presented in Table 2.

Example 2. We want to select a sample of 1,500 addresses, with probabilities proportional to the size of the address (the size is given by the number of households at the address), balanced on the socio-demographic variables quoted in Table 1 (18 variables). We also balance on the variable of inclusion probabilities and use the third landing option, for the sample to be of exact fixed size. The sample is drawn in less than one minute. The condition of fixed size is satisfied perfectly. Results are presented in Table 3.

Table 1. List of socio-demographic variables

NLOG	Number of households
NLOGCO	Number of households in collective addresses
H0019	Number of men, age less than 20
H2039	Number of men, ages 20 to 39
H4059	Number of men, ages 40 to 59
H6074	Number of men, ages 60 to 74
H7599	Number of men, age more than 75
F0019	Number of women, age less than 20
F2039	Number of women, ages 20 to 39
F4059	Number of women, ages 40 to 59
F6074	Number of women, ages 60 to 74
F7599	Number of women, age more than 75
ACTIFS	Number of working persons
INACTIFS	Number of non-workers
NATFN	Number of people with French nationality by birth
NATFA	Number of people with French nationality by acquisition
NATHE	Number of foreigners outside European Union
NATUE	Number of foreigners from European Union

Example 3. Now, suppose we want to select several samples, balanced on the former variables. We still use probabilities proportional to the size of the addresses; we intend to select 3 samples of 500 addresses. In the population, all reverse inclusion probabilities are strictly higher than 3, thus coordinated sampling is possible. We also balance on the variable of inclusion probabilities. We use the option COMPEQ=1 and the first landing option. Indeed, as the number of balancing variables is very high (38, corresponding to the 19 basic balancing variables, and 19 other variables generated by option COMPEQ=1), the sampling cannot be performed in a reasonable time with options ATTER=2 or 3. Results are presented in Table 4. The fixed size is perfectly obtained for each of the samples.

Table 2. Relative difference between the real total and the Horvitz-Thompson estimator of the total for the balancing variables

Variable	Horvitz-Thompson estimator of the total	Real total	Relative difference (%)
NLOG	251,200	251,199	0.00%
NLOGCO	251,200	251,198	0.00%
H0019	46,765	46,759	0.01%
H2039	74,720	74,729	-0.01%
H4059	45,745	45,763	-0.04%
H6074	20,865	20,870	-0.02%
H7599	10,315	10,316	-0.01%
F0019	45,855	45,852	0.01%
F2039	84,015	84,022	-0.01%
F4059	51,450	51,455	-0.01%
F6074	28,730	28,739	-0.03%
F7599	21,475	21,484	-0.04%
ACTIFS	359,940	359,984	-0.01%
INACTIFS	69,995	70,005	-0.01%
NATFN	323,190	323,219	-0.01%
NATFA	17,440	17,450	-0.06%
NATHE	8,985	8,990	-0.06%
NATUE	80,320	80,330	-0.01%

4. Global balancing and stratified balancing

4.1. Notation

We maintain the same notation as in part 1. We assume here that U is divided into H non-overlapping strata $U_1,\cdots,U_H$. We recall that the sampling design is said to be balanced on the variable x if

$$\sum_{k\in U}\frac{S_k x_k}{\pi_k}=\sum_{k\in U}x_k$$

We say that the sampling design is balanced by strata on the variable x if

$$\sum_{k\in U_h}\frac{S_k x_k}{\pi_k}=\sum_{k\in U_h}x_k,\ \text{for all}\ h=1,\cdots,H$$

Note that if a sampling design is balanced by strata, it is globally balanced on the whole population. This technique has been used in the French renovated census for the construction of rotation groups of small municipalities; in each French region, these rotation groups are made up by selecting samples which are balanced globally on socio-demographic variables, and

Table 3. Relative difference between the real total and the Horvitz-Thompson estimator of the total for the balancing variables

Variable	Horvitz-Thompson estimator of the total	Real total	Relative difference (%)
NLOG	251,199	251,199	0.00%
NLOGCO	251,199	251,198	0.00%
H0019	46,833	46,759	0.16%
H2039	74,681	74,729	-0.06%
H4059	45,733	45,763	-0.07%
H6074	20,878	20,870	0.04%
H7599	10,357	10,316	0.40%
F0019	45,806	45,852	-0.10%
F2039	83,965	84,022	-0.07%
F4059	51,572	51,455	0.23%
F6074	28,803	28,739	0.22%
F7599	21,573	21,484	0.41%
ACTIFS	360,085	359,984	0.03%
INACTIFS	70,115	70,005	0.16%
NATFN	323,584	323,219	0.11%
NATFA	17,453	17,450	0.02%
NATHE	8,963	8,990	-0.30%
NATUE	80,199	80,330	-0.16%

balanced by French department on the number of households, in order to ensure that a reasonable number of municipalities of each department can be found in any of the five rotation groups. The technique may also be used in big municipalities, to ensure reasonable sample size and good accuracy in districts, see section 5.4.

4.2. Drawbacks of direct balancing by strata

Stratified balanced sampling can be performed by selecting a sample directly from the whole population. Indeed, the condition of balance by strata is equivalent to the expression

$$\sum_{k\in U}\frac{S_k\left(x_k 1_{k\in U_h}\right)}{\pi_k}=\sum_{k\in U}x_k 1_{k\in U_h} \text{ for all } h=1,\cdots,H$$

We thus only need to select a sample in U, balanced on the variables equal to the product of the balancing variables $x_1,\cdots,x_p$ and the indicator variables:

$$1_{k\in U_h}=\begin{cases}1 & \text{if } k\in U_h\\ 0 & \text{otherwise}\end{cases}$$

which means balancing on $H\times p$ variables. This method has several drawbacks:

- If H × p is too big, we cannot perform the landing phase by searching for a sample that causes a small difference to the balancing state, because the number of possible samples is too large. The only landing option available is the first one that consists in progressively removing the constraints.
-
- All strata do not have the same balancing quality. With the first option for the landingphase, the balancing is worst for the stratum corresponding to the variables removed first.

Table 4. Relative difference between the real total and the Horvitz-Thompson estimator of the total for the balancing variables

Variable	Real total	HT estimator of the total given by the 1st sample	Relative difference (%)	HT estimator of the total given by the 2nd sample	Relative difference (%)	HT estimator of the total given by the 3d sample	Relative difference (%)
NLOG	251,199	251,199	0,00%	251,199	0,00%	251,199	0,00%
NLOGCO	251,198	251,198	0,00%	251,199	0,00%	251,198	0,00%
H0019	46,759	46,722	0,08%	46,676	0,18%	46,895	-0,29%
H2039	74,729	74,842	-0,15%	74,540	0,25%	74,724	0,01%
H4059	45,763	45,516	0,54%	45,670	0,20%	45,375	0,85%
H6074	20,870	20,786	0,40%	20,899	-0,14%	21,088	-1,04%
H7599	10,316	10,397	-0,79%	10,198	1,14%	10,495	-1,74%
F0019	45,852	46,229	-0,82%	45,991	-0,30%	46,193	-0,74%
F2039	84,022	84,463	-0,52%	83,957	0,08%	84,389	-0,44%
F4059	51,455	51,855	-0,78%	51,341	0,22%	51,308	0,29%
F6074	28,739	28,720	0,07%	28,685	0,19%	28,789	-0,17%
F7599	21,484	21,544	-0,28%	21,429	0,26%	21,455	0,13%
ACTIFS	359,984	360,602	-0,17%	359,280	0,20%	360,585	-0,17%
INACTIFS	70,005	70,472	-0,67%	70,106	-0,14%	70,127	-0,17%
NATFN	323,219	324,352	-0,35%	322,307	0,28%	324,228	-0,31%
NATFA	17,450	17,153	1,70%	17,692	-1,39%	17,474	-0,14%
NATHE	8,990	9,057	-0,75%	9,006	-0,18%	8,862	1,42%
NATUE	80,330	80,513	-0,23%	80,381	-0,06%	80,148	0,23%

- The fixed size cannot be obtained in each stratum.

The program we develop here draws its inspiration from a remark on the treatment of big databases due to Rousseau and Tardieu (2004). The idea is the following:

- We first try to balance by strata: we perform a flight phase independently on each stratum, in order to balance in the stratum, on the auxiliary variables.

- When it is not possible to balance by strata any more, we look for a global balancing: we gather the units that have not been sampled or have been rejected during the flight phases in the strata, then we perform a last flight phase on all these units before landing.

- The landing phase is then applied on the units that are not selected nor rejected.

The justification can be found in Chauvet (2006).

5. A SAS macro for stratified balancing

5.1. Description of the macro

The macro ECHANT_STRAT enables us to select a sample which is globally balanced on the whole population and approximately balanced on strata. This macro may be found on the website (http://www2.unine.ch/statistics/page10891.html) of the Neuchatel University and is available with this article.

5.2. The Input Data

There must be as many input SAS tables as strata in the population: each of these tables contains, for one particular stratum, the data relative to its units, and at least :

- the variable with inclusion probabilities,

- the balancing variables.

This table may not contain missing values for these variables. The variable with inclusion probabilities as well as the balancing variables must be of numerical type.

5.3. Syntax of the macro

5.3.1. Parameters relative to the Data Base and Output

All these parameters are compulsory.

- BASE = name of SAS library
  Name of the SAS library containing the SAS tables of Input data.

- DATA = SAS table(s)
  Name(s) of the SAS table(s) containing the Input data. The names must be spaced with blanks. Each table contains the units of one stratum.

  For example, suppose that the population is stratified into 4 strata U_1, U_2, U_3, U_4. 4 tables are created, say STRAT1 for stratum U_1, gathering the units of U_1, STRAT2 for stratum U_2, gathering the units of U_2, and so on. The syntax will be:
  DATA= STRAT1 STRAT2 STRAT3 STRAT4.

- PI = variable
  Name of the variable with the inclusion probabilities

- CONTR =variable(s)
  Names of the variables on which the sample will be balanced. The names must be spaced with blanks.

- SORT = name of SAS table
  Name of the SAS table containing the Output data. This table belongs to the library quoted in BASE. It contains all the units of the population, and a variable ECH equal to 1 if the unit has been selected in the sample, and 0 otherwise.

5.4. A numerical example

Once again, we use the former population of 26 471 units. This city is divided into 36 strata (variable ZONE). By means of the ECHANT_STRAT macro, we select a sample with equal inclusion probabilities $1/5$ which is balanced on the variables quoted in Table 1. We thus require a sample that is

- globally balanced on the whole city,

- approximately balanced within each stratum,

- of fixed size within each stratum.

We get a sample of 5,296 units in a few seconds. Table 5 compares the sample sizes we get in each stratum with those we wanted to get. If we round the sample sizes wanted, the condition of fixed size is satisfied perfectly within the strata. The estimations on the whole city are presented in Table 6. The global balancing is satisfied perfectly.

Table 5. Comparison between the sample sizes obtained and the wanted sample sizes by stratum

Stratum	1	2	3	4	5	6	7	8	9	10	11	12
Sample size wanted	155.8	137.8	148.6	141	148.8	142.4	145.6	149.6	141.8	147.6	146.4	144.8
Sample size obtained	155	138	148	141	148	142	146	150	142	148	147	145
Stratum	13	14	15	16	17	18	19	20	21	22	23	24
Sample size wanted	153.4	140	141.8	136.8	144	147.2	150.6	147.2	150.4	153.4	145.4	153.4
Sample size obtained	153	140	142	137	144	148	150	147	150	154	146	153
Stratum	25	26	27	28	29	30	31	32	33	34	35	36
Sample size wanted	151.2	146	144	155.8	145.8	147.4	151.4	143.8	153.8	142.4	152	146.8
Sample size obtained	151	146	144	156	145	148	152	144	154	143	152	147

Table 6. Relative difference between the real total and the Horvitz-Thompson estimator of the total for the balancing variables

Variable	Horvitz-Thompson estimator of the total	Real total	Relative difference (%)
NLOG	251,275	251,199	0.03%
NLOGCO	251,275	251,198	0.03%
H0019	46,790	46,759	0.07%
H2039	74,735	74,729	0.01%
H4059	45,800	45,763	0.08%
H6074	20,875	20,870	0.02%
H7599	10,315	10,316	-0.01%
F0019	45,850	45,852	0.00%
F2039	84,045	84,022	0.03%
F4059	51,475	51,455	0.04%
F6074	28,770	28,739	0.11%
F7599	21,465	21,484	-0.09%
ACTIFS	360,065	359,984	0.02%
INACTIFS	70,055	70,005	0.07%
NATFN	323,305	323,219	0.03%
NATFA	17,460	17,450	0.06%
NATHE	8,995	8,990	0.06%
NATUE	80,360	80,330	0.04%

The balancing by strata is also respected with high quality (see Table 7). As mentioned before, we could perform a similar sampling with the other macro EXE_CUBE. We would draw one sample directly in the whole population. The following balancing variables should then be used :

- the inclusion probability (to get a fixed sample size) and the 18 variables quoted above to obtain a global balancing. If we take collinearities into account, that means: 17 variables,
- a variable indicating the fact that a unit belongs to a stratum, to get a fixed sample size by stratum. That means: 35 balancing variables,
- variables equal to the product of the socio-demographic variables (18) and the variables indicating the belonging of a unit to a stratum (36) to get a stratified balancing. If we take collinearities into account, that means: $16 \times 35 = 560$ balancing variables.

For the same kind of sampling, we would have needed 612 balancing variables. The sampling would have been much slower, and the balancing would have been very badly performed in some strata.

Table 7. Indicators of the quality of balancing by stratum for the balancing variables

Stratum	1	2	3	4	5	6	7	8	9	10	11	12
Maximum relative difference (modulus) in percentage	3.2	2.4	3.8	5.7	1.7	3.1	4.0	2.2	3.9	3.2	3.5	6.3
Average relative difference (modulus) in percentage	0.7	0.8	1.1	1.0	0.7	1.0	0.9	0.7	1.2	0.9	0.9	0.8
Stratum	13	14	15	16	17	18	19	20	21	22	23	24
Maximum relative difference (modulus) in percentage	3.1	4.9	2.6	3.2	2.2	5.1	2.6	2.1	3.9	3.2	2.1	4.3
Average relative difference (modulus) in percentage	1.3	1.4	0.8	0.7	0.5	1.2	1.3	0.5	1.1	1.0	0.7	1.2
Stratum	25	26	27	28	29	30	31	32	33	34	35	36
Maximum relative difference (modulus) in percentage	6.8	1.5	1.7	1.8	4.5	3.9	4.0	1.5	1.5	1.8	2.8	2.7
Average relative difference (modulus) in percentage	1.0	0.5	0.5	0.6	1.4	1.4	0.9	0.6	0.4	0.6	0.6	0.9

Appendix

SAS program for drawing samples

```
/*****************************************/
/* Drawing balanced samples
*/
/*****************************************/

%include "C:\fastcube\fast_cube.sas";
%include
"C:\fastcube\fast_cube_stratification.sas";

libname base "C:\fastcube\files";

%let liste = pi nlogco h0019 h2039 h4059
h6074 h7599 f0019 f2039 f4059 f6074 f7599
actifs inactifs natfn natfa nathe natue;
%let nb = 18;

data municipality ; set base.municipality ;
run;

/* Example 1 */
data municipality ; set base.municipality ;
pi=1/5 ; run;

%exe_cube ( base = work, data =
municipality, id = id_adresse, pi = pi,
contr = &liste, sort = ech1, atter = 1,
compeq = 0);
data ech1 ; set ech1 ; if ech = 1;run;

/* Example 2 */
data _null_ ; set municipality end=fin ;
totlog + nlog ;
if fin then call symput("totlog",totlog) ;
run;

data municipality ; set base.municipality ;
pi=1500 * nlog / &totlog ; run;
%exe_cube(base=work , data = municipality ,
id = id_adresse, pi = pi, contr = &liste,
sort = ech2, atter = 3, compeq = 0);
data ech2 ; set ech2 ; if ech = 1 ; run;

/* Example 3 */

%macro example3;

%let liste_bis=;

data _null_ ; set municipality end=fin ;
totlog + nlog ;
if fin then call symput("totlog",totlog) ;
run;

data municipality ; set municipality ; pi =
500 * nlog / &totlog ; run;

%exe_cube (base = work, data = municipality,
id = id_adresse, pi = pi, contr = &liste,
sort = ech3_1, atter = 1, compeq = 1);
data comp ; set ech3_1; if ech=0;

%do blacksad = 1 %to &nb;
      %let var = %scan(&liste,&blacksad);
      %let liste_bis = &liste_bis &var._bis;
      &var._bis = &var/(1-pi);
%end;
run;

%exe_cube ( base = work, data = comp, id =
id_adresse, pi = pi_bis, contr = &liste_bis,
sort = ech3_2, atter = 1, compeq = 1);

data comp ; set ech3_2 ; if ech = 0 ;
%do blacksad = 1 %to &nb;
      %let var = %scan(&liste,&blacksad);
      &var._bis = &var/(1-2*pi);
%end;
run;

%exe_cube (base = work, data = comp, id =
id_adresse, pi = pi_bis, contr = &liste_bis,
sort = ech3_3, atter = 1,compeq = 0);

data ech3_1 ; set ech3_1 ; if ech = 1 ; run;
data ech3_2 ; set ech3_2 ; if ech = 1 ; run
;
data ech3_3 ; set ech3_3 ; if ech = 1 ; run;

%mend example3;

%example3;

/* Example 4 : stratified balancing */

%macro example4;

data municipality ; set municipality ;
pi=1/5 ; run;

data
      %do j=1 %to 35;
            zone&j
      %end;
      zone36;
set municipality;
      %do i=1 %to 36;
            if zone=&i then output zone&i;
      %end;
run;

%echant_strat ( base = work, data = %do i=1
%to 36; zone&i %end;, id = id_adresse, pi =
pi, contr = &liste, sort = ech_strat);

data ech_strat ; set ech_strat ; if ech=1 ;
run;

%mend example4;

%example4;
```

REFERENCES

Bertrand, P., B. Christian, G. Chauvet and J.M. Grosbras. 2004. Plans de sondage pour le recensement rénové de la population. In *Séries INSEE Méthodes: Actes des Journées de Méthodologie Statistique*, Paris. Paris: INSEE, to appear (in French).

Caron, N. and M. Christine. 2005. Du nouveau recensement au futur système d'échantillonnage des enquêtes ménages. In *Séries INSEE Méthodes: Actes des Journées de Méthodologie Statistique*, Paris. Paris: INSEE (in French).

Chauvet, G. 2006. De nouvelles macros SAS d'échantillonnage équilibré. Technical report, ENSAI, Rennes (in French).

Chauvet, G. and Y. Tillé. 2006. A fast algorithm of balanced sampling. *Computational Statistics*, 21:53–61.

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

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

Dumais, J. and M. Isnard. 2000. Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. In *Séries INSEE Méthodes: Actes des Journées de Méthodologie Statistique*, volume 100. pp. 37-76, Paris: INSEE (in French).

Rousseau, S. and Tardieu, F. (2004). La macro SAS cube d'échantillonnage équilibré – documentation de l'utilisateur. Technical report, INSEE (in French).

Tillé, Y. 2006. Sampling algorithms. *Springer*, New-York.

Tillé, Y. and A.C. Favre. 2004. Coordination, combination and extension of balanced samples. *Biometrika*, 91:913–928.