Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Ten years of balanced sampling with the cube method: An appraisal

by Yves Tillé

December 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 215-226
Statistics Canada, Catalogue No. 12-001-X

Ten years of balanced sampling with the cube method: An appraisal

Yves Tillé [1]

Abstract

This paper presents a review and assessment of the use of balanced sampling by means of the cube method. After defining the notion of balanced sample and balanced sampling, a short history of the concept of balancing is presented. The theory of the cube method is briefly presented. Emphasis is placed on the practical problems posed by balanced sampling: the interest of the method with respect to other sampling methods and calibration, the field of application, the accuracy of balancing, the choice of auxiliary variables and ways to implement the method.

Key Words: Sampling; Balancing; Horvitz-Thompson estimator.

1. Introduction

While the idea of balanced sampling has been around since the early days of survey statistic development, applying the concept has been difficult because almost all the proposed methods have either been enumerative or rejective and required considerable computation time. The algorithm of the cube method was proposed in 1998 by Deville and Tillé, and a first implementation was written by three students of the École Nationale de la Statistique et de l'Analyse de l'Information of Rennes in France (see Bousabaa, Lieber and Sirolli 1999). Finally, the method was published in Tillé (2001) and Deville and Tillé (2004). Since this time, several implementations have been proposed and several survey managers have used the cube method to select samples, the most important applications being the New French Census and the French Master Sample.

Our aim is to assess 10 years of development and use of balanced sampling in order to better ascertain when and how the cube method can be used to select samples of householders or establishments. After discussing the concept of balanced sample and balanced sampling in Section 2, we give a list of particular cases in Section 3. In Section 4, we briefly trace the history of this concept for both the model-based and design-based frameworks. Next, in Section 5, we provide a brief overview of the cube method, which is a class of algorithms that allows us to select randomly balanced samples with given inclusion probabilities (see Deville and Tillé 2004; Tillé 2001, 2006b). We try to present the main principles of this algorithm without giving a detailed description of the technicalities of the method. Section 6 is devoted to the principles of variance estimation in balanced sampling. Finally, in Sections 7, we discuss the interest of balanced sampling in practice and compare balanced sampling with other sampling methods and calibration. We also give a list of recent applications. This Section also deals with the accuracy of balancing, the choice of auxiliary variables and ways to implement balanced sampling. The paper ends with an exhaustive bibliographical list of references on balanced sampling and their applications.

2. Balanced sampling

2.1 Definition of a balanced sample

Consider a sample s of size n that is a subset of a finite population U of size N. A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kp})'$,

$$\frac{1}{n}\sum_{k\in S}\mathbf{x}_k = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k, \tag{1}$$

which means that the sample means of the x-variables match their population means.

Brewer (1999) drew a distinction between a balanced selection of samples and a random selection of samples. However, a balanced sample may be selected randomly. If a random sample S is selected randomly, then each unit of the population has an inclusion probability π_k of being selected. In this case, a random sample must satisfy the following balancing equations:

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U}\mathbf{x}_k. \tag{2}$$

In other words, in a balanced sample, the total of the x-variables are estimated without error. Several authors like Cumberland and Royall (1981) and Kott (1986) would call a sample that satisfies Equation (2) a 'π-balanced sample', as opposed to a 'mean-balanced sample' defined by Equation (1). Nevertheless, in this paper, we will consider that (1) is only a particular case of (2) that occurs when $\pi_k = n/N$ or when the sample is not selected randomly. We refer to both cases as a balanced sample.

1. Yves Tillé, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel Switzerland. E-mail : yves.tille@unine.ch.

2.2 Balanced sampling design

Let $p(s)$ denote the sampling design, *i.e.*, the probability that sample s is selected, such that $p(s) = \Pr(S = s)$, where S is the random sample and $n(S)$ the size of the sample S. According to the definition of Deville and Tillé (2004), a sampling design $p(\cdot)$ is said to be *balanced* on auxiliary variables $x_1, \ldots, x_p$ if the Horvitz-Thompson estimator satisfies Equation (2). In a balanced sampling design, the inclusion probabilities are decided prior to sampling. A balanced sampling can be viewed as a kind of calibration that is directly integrated into the sampling design. The main problem is that the balancing equations (2) can rarely be exactly satisfied. We refer to this difficulty as the 'rounding problem'.

Example 1. If $N = 4$, $n = 2$, $\pi_k = 1/2$, for all $k \in U$ and $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 4$, then the balancing equations given in (2) becomes

$$\frac{1}{n}\sum_{k\in s} x_k = \frac{1}{N}\sum_{k\in U} x_k,$$

which is equivalent to

$$\sum_{k\in s} x_k = \frac{n}{N}\sum_{k\in U} x_k. \qquad (3)$$

Since

$$\frac{n}{N}\sum_{k\in U} x_k = \frac{2}{4}(0 + 1 + 2 + 4) = 3.5,$$

and the left side of (3) is always an integer, then an exactly balanced sample does not exist.

Indeed, sample selection is an integer problem. The cube method therefore aims to select a sample that exactly satisfies the inclusion probabilities π_k while remaining as balanced as possible.

3. Special cases of balanced sampling

3.1 Unequal probability sampling and stratification

Some well-known sampling designs are particular cases of balanced sampling:

1. Sampling with a fixed sample size is a particular case of balanced sampling. In this case, the only balancing variable is π_k. The balancing equations given in (2) become

$$\sum_{k\in S}\frac{\pi_k}{\pi_k} = \sum_{k\in S} 1 = \sum_{k\in U}\pi_k,$$

which means that the sample size must be fixed.

2. Stratification is a particular case of balanced sampling. Suppose that the population is partitioned in H strata U_h, $h = 1, \ldots, H$, of sizes N_h, $h = 1, \ldots, H$, and that a sample is selected in each stratum by means of simple random sampling without replacement with fixed sample size n_h, $h = 1, \ldots, H$. In this case, the balancing variables are the indicator variables of the strata

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

Under a stratified design, the Horvitz-Thompson estimators of the sizes of the strata exactly equal the sizes of the strata, which is a property of balancing on the indicator variables of the strata. Indeed, since the inclusion probabilities in stratum h are $\pi_k = n_h / N_h$, $k \in U_h$, the balancing equations become

$$\sum_{k\in S}\frac{N_h\delta_{kh}}{n_h} = \sum_{k\in U}\delta_{kh} = N_h,\ h = 1, \ldots, H,$$

and are exactly satisfied.

These two designs are well known and are commonly applied in official statistics in order to reduce variance. The more general concept of balancing allows more freedom to choose the most appropriate balancing variables that will improve the accuracy of the estimators.

3.2 Overlapping strata

Constructing a stratified sampling design is often a difficult exercise. Statisticians often try to stratify using several qualitative variables. However, in most cases, crossing all of the strata of all the variables will cause the cells to become too small for a sample to be selected in each cell. In the context of calibration, statisticians generally calibrate on marginal totals and not on all the cells contained in a contingency table. Since a balanced sampling can be viewed as a kind of calibration that is directly integrated in the sampling design, one would also like to balance using only marginal totals. Nevertheless, the usual theory of stratification does not allow overlapping strata since the stratification must be a partition of the population. Now, the cube method enables us to directly balance on totals of overlapping strata by simply using the indicators of the strata as balancing variables.

3.3 Balancing on a constant

Another interesting special case of balanced sampling occurs when a constant is used as a balancing variable. If $\mathbf{x}_k = 1$ for all $k \in U$, the balancing equations become

$$\sum_{k\in S}\frac{1}{\pi_k} = \sum_{k\in U} 1 = N.$$

Actually,

$$\sum_{k \in S} \frac{1}{\pi_k}$$

is the Horvitz-Thompson estimator of N. This means that, if a constant is used as a balancing variable, the estimated population size matches the known size N, which is far from being a given when the statistical units are selected with unequal inclusion probabilities.

4. History of the concept of balancing and existing methods

The idea of balanced sampling is very old and is linked to the vague concept of representativeness that was already used by Kiaer (1896, 1899, 1903, 1905). The first paper dedicated to the selection of a balanced sample is due to Gini (1928) and Gini and Galvani (1929) who selected a sample of 29 from the 214 Italian districts in order to match several population totals. Both Neyman (1952) and Yates (1960) condemned the paper of Gini and Galvani essentially because this sample was not randomly selected (see Langel and Tillé 2010). The first methods for selecting a random balanced sample were proposed by Yates (1946) and Thionet (1953), but these methods were rejective in the sense that they involved selecting samples or changing units randomly in the sample until a balanced enough sample was obtained.

In the model-based framework, Royall (1976a, b) advocated the use of balanced sampling in order to reach the optimal strategy and to protect against mis-specification of the model. (see also Royall and Pfeffermann 1982; Kott 1986; Cumberland and Royall 1988; Royall 1988; Tirari 2006; Nedyalkova and Tillé 2009). While several methods for selecting a balanced sample are presented in the book of Valliant, Dorfman and Royall (2000), these methods do not necessarily specify the inclusion probabilities of the sample. In the model-based framework, it is important to have a balanced sample. However, this sample does not always need to be randomly selected.

Hájek (1981) also advocated the use of balanced sampling. For Hájek, a balanced sampling is a particular case of representative strategy, a strategy being a couple made of a sampling design and an estimator. A representative strategy is a strategy that estimates the totals of auxiliary variables without error. In this sense, a balanced sampling design with the Horvitz-Thompson estimator is a representative strategy. Hájek (1981) proposes a rejective procedure that consists of selecting a sequence of samples until a balanced one is obtained. Rejective procedures have two drawbacks: if several balancing variables are used, the procedure can be very slow; secondly, the inclusion probabilities of rejective designs are not the same as the original design. The inclusion probabilities of statistical units that are close to the population means are increased to the detriment of the units that are far from the center (see for instance the simulations of Legg and Yu 2010).

Another method of selection consists of enumerating all the possible samples, and then constructing a sampling design only to select the samples that are adequately balanced. Such a design can be constructed by using linear programming. This technique was applied by Ardilly (1991) to select the primary units of the French master sample. Nevertheless, this method can only be applied on small population sizes because of the combinatory explosion of the number of samples when the size of the population is large.

Deville, Grosbras and Roth (1988) and Deville (1992) proposed multivariate methods for balanced sampling with equal inclusion probabilities. Hedayat and Majumdar (1995) have proposed the adaptation of an experimental design technique that would enable a balanced sampling design to be constructed. Again, this technique is restricted to equal inclusion probabilities. Finally, the cube method was proposed by Deville and Tillé (2004). This method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible.

Fuller (2009) studied a rejective procedure by fixing a tolerance interval outside of which the sample is rejected and proposed an estimator of variance. Even if the inclusion probabilities are changed with a rejective procedure, Fuller (2009) shows that efficient estimates are obtained by using the inclusion probabilities of the original design. Using a set of simulations, Legg and Yu (2010) compared this rejective procedure to the cube method and showed that both methods perform equally. Finally, Dudoignon and Vanheuverzwyn (2006) proposed a fast method of balanced sampling for marginal totals, whereas Périé (2008) proposed a method based on permanent random numbers that provides a balanced sample. With the Périé (2008) method, the inclusion probabilities are only approximately satisfied.

5. The cube method

5.1 Main ideas

The cube method (see Deville and Tillé 2004; Tillé 2001, 2006a, b; Ardilly 2006) is a class of sampling algorithms that selects a balanced sample and exactly satisfies a set of given inclusion probabilities. The cube method is an extension of the splitting method that was developed by Deville and Tillé (1998). It is based on a random transformation of the vector of inclusion probabilities until a sample is obtained such that:

(i) the inclusion probabilities are exactly satisfied,
(ii) the balancing equations are satisfied to the furthest extent possible.

The name of the method comes from the geometric representation of a sampling design. Indeed, a sample may be represented by a vector of samples indicators:

$$\mathbf{s} = (I[1 \in s] \ldots I[k \in s] \ldots I[N \in s])',$$

where $I[k \in s]$ takes value 1 if $k \in s$ and 0 if not. A sample may thus be viewed as a vertex of an N-cube as showed in Figure 1.

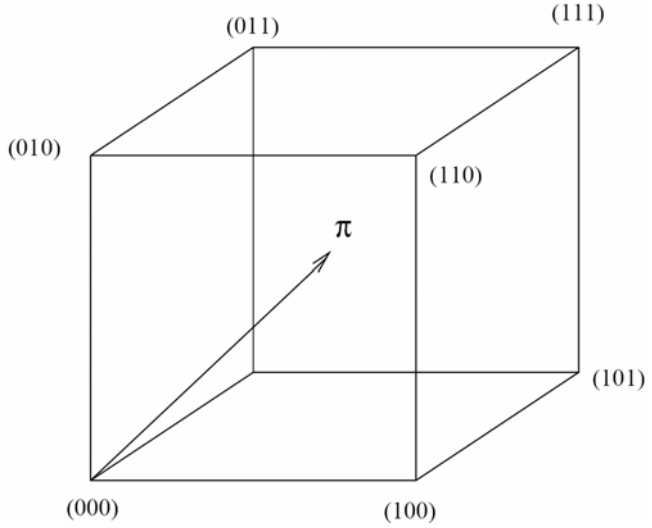

Figure 1 Possible samples in a population of size $N = 3$

Let us also define

$$\mathrm{E}(\mathbf{s}) = \sum_{s \in S} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi},$$

where $\boldsymbol{\pi} = [\pi_k]$ is the vector of inclusion probabilities. The balancing equations

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k,$$

may also be written

$$\sum_{k \in U} \breve{\mathbf{x}}_k \, s_k = \sum_{k \in U} \breve{\mathbf{x}}_k \, \pi_k, \qquad (4)$$

where $s_k \in \{0, 1\}$ and $\breve{\mathbf{x}}_k = \mathbf{x}_k / \pi_k$, $k \in U$. Expression (4) is a system of equations with unknowns values s_k that define an affine subspace in $\mathbb{R}^N$ of dimension $N - p$ denoted by Q, where

$$Q = \left\{ \mathbf{u} \in \mathbb{R}^N \,|\, \sum_{k \in U} \breve{\mathbf{x}}_k \, u_k = \sum_{k \in U} \mathbf{x}_k \right\}.$$

The problem of selecting a balanced sample may thus be reformulated. A balanced sampling design consists of choosing a vertex of the N-cube (a sample) that remains on the linear sub-space Q. Figures 2 and 3 respectively show two examples: the first one is a constraint of fixed sample size and the second one is a constraint that generates a rounding problem.

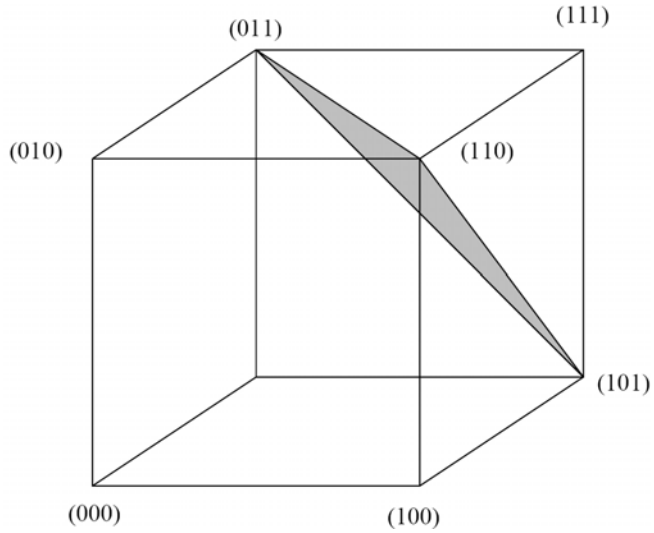

Figure 2 Possible samples in a population of size $N = 3$ with a constraint of fixed sample size $n = 2$

The Cube method (Deville and Tillé 2004) is divided into two phases: the flight phase and the landing phase. The flight phase is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. At the end of the flight phase, if a sample is not obtained, the landing phase entails in selecting a sample that is as close as possible to the constraint subspace.

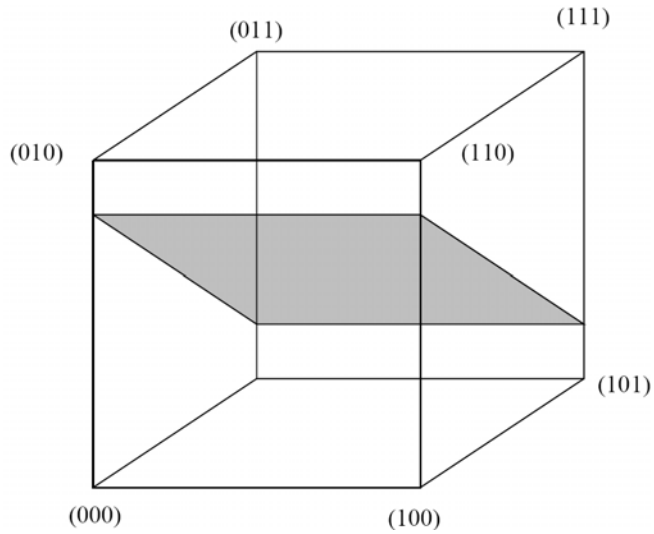

Figure 3 Possible samples in a population of size $N = 3$ with a constraint and a rounding problem

Example 2. If the constraint is the fixed sample size, the flight phase randomly transforms a vector of inclusion probabilities into a vector of 0 and 1. At each step of the algorithm, the vector of inclusion probabilities is transformed randomly, but the sum of inclusion probabilities must remain equal to n. For instance, with $\boldsymbol{\pi} = (0.5, 0.5, 0.5, 0.5)$ and $n = 2$, we are able to obtain the following sequence of vectors:

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6666 \\ 0.6666 \\ 0.6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{s}.$$

The algorithm ends when all the components of the vector are equal to 0 or 1.

Example 3. If the constraint is the fixed sample size, a rounding problem appears if the sum of inclusion probabilities is not an integer. If there is a rounding problem, then some components cannot be set to zero. For instance, with $\boldsymbol{\pi} = (0.5, 0.5, 0.5, 0.5, 0.5)$ and

$$\sum_{k \in U} \pi_k = 2.5,$$

we may observe the following sequence of vectors:

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.625 \\ 0 \\ 0.625 \\ 0.625 \\ 0.625 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.25 \\ 1 \\ 0.25 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix} = \boldsymbol{\pi}^*.$$

In this case, the flight phase cannot end with a vector of 0 or 1 of which the sum is equal to 2.5. In this case, the flight phase ends with a vector containing one non-integer component.

5.2 The flight phase

The first step of the flight phase is presented in Figure 4 for a very specific case: the population size $N = 3$. The only balancing constraint is the fixed sample size $n = 2$. At the first step, a vector $\mathbf{u}(0)$ must be chosen. This vector may be chosen freely but must be such that $\boldsymbol{\pi} + \mathbf{u}(0)$ remains in the subspace of constraints. Actually, the cube method is a family of methods that depends on the way the vector $\mathbf{u}(0)$ is chosen. This vector may be chosen randomly or not.

If, from $\boldsymbol{\pi}$, we follow the direction given by vector $\mathbf{u}(0)$, then we will necessarily cross a face of the cube. Let us consider this point denoted on Figure 4 by $\boldsymbol{\pi}(0) + \lambda_1^*(0)\mathbf{u}(0)$. Now, if, from $\boldsymbol{\pi}$, we follow the opposite direction, *i.e.*, the direction given by vector $-\mathbf{u}(0)$, we will also cross a face of the cube. Let us consider this point denoted on Figure 4 by $\boldsymbol{\pi}(0) - \lambda_2^*(0)\mathbf{u}(0)$. At the first step, vector $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ is modified randomly. Vector $\boldsymbol{\pi}(1)$ will be set to $\boldsymbol{\pi}(0) + \lambda_1^*(0)\mathbf{u}(0)$ or to $\boldsymbol{\pi}(0) - \lambda_2^*(0)\mathbf{u}(0)$. The choice is done randomly in such a way that $\mathrm{E}[\boldsymbol{\pi}(1)] = \boldsymbol{\pi}(0)$. At the end of the first step of the flight phase, we have thus jumped on a face of the cube, which means that at least one component of $\boldsymbol{\pi}(1)$ is equal to 0 or 1, *i.e.*, the problem is reduced from a problem of sampling from a population of size $N = 3$ to a population of size $N = 2$. In N steps at least, the flight phase is thus completed.

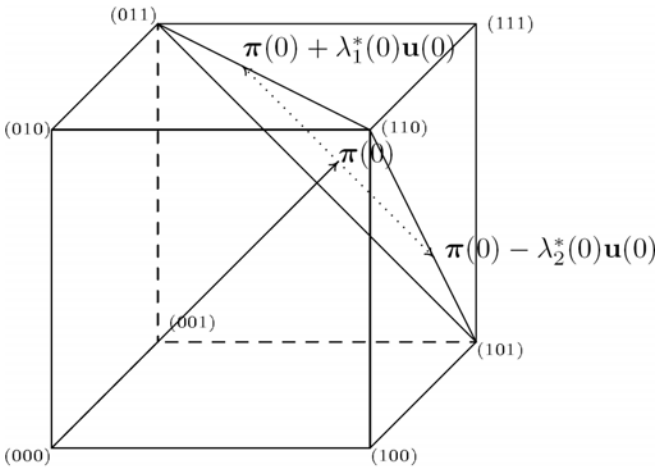

Figure 4 Flight phase in a population of size $N = 3$ with a sample size constraint $n = 2$

More generally, the flight phase is a random walk in the intersection of the balancing subspace and the cube. This random walk stops at a vertex of the intersection of the cube and the subspace. The flight phase is defined by the following class of algorithms. First initialize with $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Next, at time $t = 0, \ldots, T$,

1. Generate any vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ such that
   (i) $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_k/\pi_k, \ldots, \mathbf{x}_N/\pi_N)$, *i.e.*, $\mathbf{Au}(t) = 0$,
   (ii) $u_k(t) = 0$ if $\pi_k(t)$ is integer.
2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values such that
$$0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$$
$$0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1.$$
3. Compute
$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } q_2(t), \end{cases}$$
where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)$.

The flight phase stops when it is no longer possible to find a vector $\mathbf{u}(t) \neq 0$.

5.3 Landing phase

If, at the end of the flight phase, the balancing equations are not exactly satisfied, there is a need for a landing phase. Let $\boldsymbol{\pi}^* = [\pi_k^*]$ be the vector obtained at the last step of the flight phase. It is possible to prove (see Deville and Tillé 2004) that

$$\mathrm{card}(U^*) \leq p,$$

where

$$U^* = \{k \in U \mid 0 < \pi_k^* < 1\}$$

and p is the number of balancing variables. The aim of the landing phase is to find a sample $\mathbf{s}$ such that

$\mathrm{E}(\mathbf{s}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$, which is almost balanced. There are two ways of selecting such a sample:

1. *The flight phase by linear programming* consists of considering all the possible samples of U^*. A cost is assigned to each sample. This cost, is, for instance, the distance between the sample and the subspace of constraints. Next, one looks for a sampling design on U^* that minimizes the expected cost and that satisfies the inclusion probabilities $\boldsymbol{\pi}^*$. This problem can be solved because the number of samples to consider is reasonable due to the small size of U^*.
2. *The flight phase by suppression of variables* may be used when the number of balancing variables is too large for the linear program to be solved by a simplex algorithm ($p > 20$). With this method, an auxiliary variable is dropped at the end of the flight phase. Next, we can return to the flight phase until it is no longer possible to 'move' within the constraint subspace. The constraints are then relaxed successively according to an order of preference.

6. Variance and variance estimation

6.1 A residual technique

The variance of the Horvitz-Thompson estimator can be estimated by using a residual technique developed in Deville and Tillé (2005). The residual technique is comparable to the technique used to estimate the variance of the calibration estimator and has been validated by a set of simulations. The estimated variance of the Horvitz-Thompson estimator is thus very similar to the estimated variance of a generalized regression (GREG) estimator. Nevertheless, the variance of the GREG estimator is generally underestimated because it does not take into account the randomness of the weights. Indeed, if the usual variance of the GREG estimator is computed for the special case of poststratification, we obtain the variance of a stratified design with proportional allocation. The variance of the poststratified estimator is nevertheless larger than the variance in a stratified design with proportional allocation.

6.2 Approximation of variance

If the balanced sampling design has a large entropy, Hájek (1981) and Deville and Tillé (2005, method 4) have proposed the following approximation of the design variance given by:

$$\mathrm{var}_p(\hat{Y}_\pi) \cong \mathrm{var}_{app}(\hat{Y}_\pi) = \sum_{k\in U} d_k \frac{(y_k - \mathbf{x}_k'\mathbf{b})^2}{\pi_k^2}, \quad (5)$$

where the subscript p denotes the sampling design,

$$\mathbf{b} = \left(\sum_{k\in U} d_k \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k^2}\right)^{-1} \sum_{k\in U} d_k \frac{\mathbf{x}_k y_k}{\pi_k^2},$$

and the d_k are the solution of the nonlinear system

$$\pi_k(1-\pi_k) = d_k - \frac{d_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in U} d_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{d_k\mathbf{x}_k}{\pi_k}, \; k \in U. \quad (6)$$

The entropy of the sampling design depends on the way vectors $\mathbf{u}(t)$ are chosen during the flight phase. In order to increase the entropy, vector $\mathbf{u}(t)$ can be chosen randomly or the population can be randomly sorted before selecting the sample.

Expression (5), which only uses the first-order inclusion probabilities, was validated by Deville and Tillé (2005) under a variety of balanced samples regardless of how the y-values were generated. An approximation very close to Expression (5) was obtained by Fuller (2009) and Legg and Yu (2010) for a balanced sampling design obtained by a rejective procedure in the case of an initial design that uses Poisson sampling. These approximations do not take the rounding problem into account.

6.3 Estimation of variance

Deville and Tillé (2005) proposed a family of variance estimators for balanced sampling, of the form

$$\widehat{\mathrm{var}}(\hat{Y}_\pi) = \sum_{k\in S} c_k \frac{(y_k - \mathbf{x}_k'\hat{\mathbf{b}})^2}{\pi_k^2}, \quad (7)$$

where

$$\hat{\mathbf{b}} = \left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell y_\ell}{\pi_\ell^2}$$

and the c_k are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{c_k\mathbf{x}_k}{\pi_k}, \quad (8)$$

which can be solved by a fixed point algorithm.

In Deville and Tillé (2005), simpler variants of c_k were also proposed. For instance, one can use the alternative values,

$$\tilde{c}_k \approx \frac{n}{n-p}(1-\pi_k),$$

that are very close to c_k. The estimator $\widehat{\mathrm{var}}(\hat{Y}_\pi)$ is approximately design-unbiased because it is an estimator by substitution of the approximation given in expression (5), (for more information regarding estimators obtained by substitution, see Deville 1999), which is a reasonable approximation of the variance under the sampling design.

It is not easy to use bootstrap method to estimate the variance in the context of balanced sampling. Balanced samples with replacement should be selected from the original sample. A generalization of the cube method for balanced sampling with replacement has not yet been described. A solution, proposed by Chauvet (2007), consists of reconstructing an artificial population from the sample.

Next, bootstrap samples are selected by using balanced sampling. Another solution was proposed by Fuller (2010) for balanced rejective sampling. Breidt and Chauvet (2010a) have proposed an alternative method where a martingale difference representation of the cube method is used in order to approximate second-order inclusion probabilities, which enables us to construct a nearly unbiased variance estimator.

7. Balanced sampling in practice

7.1 Interest of balanced sampling

In the model-assisted and the model-based frameworks, a balancing sampling design with the Horvitz-Thompson estimator is often the optimal strategy (see Nedyalkova and Tillé 2009). Indeed, when the sample is balanced, the variances of the Horvitz-Thompson estimators of the auxiliary variables are equal to zero. Under a linear model, the variance of the Horvitz-Thompson estimator of the interest variable will only depend on the residuals of the model.

The advantages of balanced sampling are as follows:

(i) Balanced sampling increases the accuracy of the Horvitz-Thompson estimator. This point has been developed in Section 6. Indeed, the variance of the Horvitz-Thompson estimator only depends on the residuals of the regression of the interest variable by the balancing variables.

(ii) Balanced sampling protects against large sampling errors. Indeed, the most unfavourable samples have a null probability of being selected.

(iii) If the variable of interest is well explained by the auxiliary information, in model-based inference, balanced sampling protects against a mis-specification of the model. This point is largely developed by Royall (1976b, a) and Valliant *et al.* (2000). A recent discussion of this important question is given in Nedyalkova and Tillé (2009, 2010).

(iv) Balanced sampling can ensure that the sample sizes in planned domains are not too small or - much worse - equal to zero. Indeed, if an indicator variable of the domain is added in the list of auxiliary variables, the size of the domain is then fixed in the sample.

(v) Balanced sampling allows us to avoid random weights. With balanced sampling, the Horvitz-Thompson weights can be used. If the sampling design does not contain any balancing constraints (for instance with Poisson sampling) the weighting system obtained by a calibration procedure becomes very random, which increases the variance of the estimators. If the sample is balanced, the weights will be less random even if a calibration procedure is used after balancing.

The availability of easy to use packages contributed to the large use of the cube method in several important statistical processes. The first main application of the cube method is selection of the rotation groups for the French census. (See Desplanques 2000; Dumais, Bertrand and Kauffmann 2000; Durr and Dumais 2001, 2002; Dumais and Isnard 2000; Bertrand, Christian, Chauvet and Grosbras 2004; da Silva, da Silva Borges, Aires Leme and Moura Reis Miceli 2006). For the municipalities with fewer than 10,000 inhabitants, five non-overlapping rotation groups of municipalities are selected using a balanced sampling design with equal inclusion probabilities (1/5). Each year, a fifth of the municipalities are surveyed. So after 5 years, all the small municipalities are selected. For the municipalities with more than 10,000 inhabitants, in each municipality, five non-overlapping balanced samples of addresses are selected with inclusion probabilities 8%. So, after 5 years, 40% of the addresses are visited. The balancing variables are socio-demographic variables taken from the last census.

In the French master sample, the primary units are geographical areas that are selected using a balanced sampling design (see Wilms 2000; Christine and Wilms 2003; Christine 2006). The master sample is a self-weighted multi-stage sampling. So the primary units are selected with unequal probabilities that are proportional to their sizes. The balancing variables are socio-demographic variables taken from the last census. Bardaji (2001) and Even (2002) have also used balanced sampling to select a sample of beneficiaries of subsidized jobs. Seven populations are surveyed, a balanced sample of beneficiaries is selected in each of the populations by using between two and five balancing variables according to the populations.

In the company Électricité de France (EDF), new electricity meters allow electricity consumption for each household to be measured on a continuous basis. The amount of information collected is so large that it is impossible to archive all the data. Dessertaine (2006, 2007) used balanced sampling to select the time series of consumption that must be archived in order to ensure that they represent the consumption of the entire French population as accurately as possible. Biggeri and Falorsi (2006) used balanced sampling to improve the quality of the consumer price index in Italy. Gismondi (2007) tested balanced sampling to estimate the number of tourist nights spent in Italy. D'Alò, Di Consiglio, Falorsi and Solari (2006) and Falorsi and Righi (2008) also proposed using a balanced sampling design to estimate totals in small domains. Simulations were run by Marí, Barbará, Mitas and Passamonti (2007b, a) in Argentina and Chipperfield (2009) in Australia to assess the interest of balanced sampling for the master sample.

At Statistics Canada, Fecteau and Jocelyn (2006) and Jocelyn (2006) tested balanced sampling to select a sample of businesses. Canadian unincorporated businesses complete their income tax returns either on paper or electronically. More than half of the returns are submitted electronically. Balanced sampling was used to select a sample from businesses that responded electronically so that, for some key variables that are known for the whole population, the sample means matched the known population means.

Balanced sampling can also be used to impute a missing value in case of item nonresponse. Indeed, using a model to predict an imputation allocates central values, which will lead to a biased inference on quantiles. In contrast, a random imputation generally increases the variances of the estimators. In order to solve this dilemma, Deville (1998, 2005, 2006) and Chauvet, Deville and Haziza (2010c, b) have proposed using imputation by prediction and to add a residual that is chosen amongst the residuals of the respondent according to a balanced sampling design. This is done to avoid adding a term of variance to the total of the imputed variable.

7.2 Balanced sampling versus other sampling techniques

Unequal probability sampling is a particular case of the cube method. Indeed, when the only auxiliary variable is the inclusion probability, the sample has a fixed sample size. The cube method is a generalization of the splitting method (see Deville and Tillé 1998), which includes several sampling algorithms with unequal probabilities (Brewer's method, pivotal method, corrected Sunter method, see Brewer 1975; Sunter 1977; Deville and Tillé 1998; Tillé 2006b). Stratification is also a particular case of balanced sampling. With the cube method, one can balance on overlapping strata and use qualitative and quantitative variables together. Systematic sampling can even be seen as a balanced sampling design on the order statistic related to the variable on which the population is ordered.

Almost all the other sampling techniques are particular cases of balanced sampling (except multistage sampling). In fact, balanced sampling is simply more general, in the sense that all the other methods of sampling can be implemented with the cube method. The cube method allows us to use any variable for balancing with a reasonable computation time. With the more general concept of balancing, strata can overlap, quantitative and qualitative variables can be used together, and the inclusion probabilities can be chosen freely.

It is well known that the ratio estimator and the post-stratified estimator are particular cases of the regression estimator. The regression estimator is also a particular case of the calibration estimator (which includes a non-linear adjustment). In the same way, balanced sampling is a more general method of sampling that includes almost all the other methods. The algorithm of the cube method may seem complicated but, once implemented, it enables us to run a function with two arguments: the vector of inclusion probabilities and the matrix of balancing variables.

7.3 Choice of the balancing strategy

The main recommendation is to choose balancing variables that are closely correlated to the interest variables. As with any regression problem, the balancing variables must be chosen parsimoniously: one must not choose too many balancing variables because, accuracy no longer improves with a large number of variables and the instability of the variance estimator increases with each additional variable. Practically, the aim is not to estimate one variable but a set of interest variables. Thus, the set of auxiliary variables must be correlated to all the interest variables. Moreover, the auxiliary variables should not be too correlated amongst themselves.

Lesage (2008) has proposed a method to balance a sample on complex statistics rather than simply using population totals. The main idea consists in balancing on the linearized value (or influence function) of the parameter of interest. Breidt and Chauvet (2010b) have proposed using penalized balanced sampling in order to possibly relax some balancing constraints, which can be useful for instance in small domain estimation.

In many cases, the balancing variables contain measurement errors. For example, in most registers, one can suspect errors in the data. Missing values can obviously occur and auxiliary variables are often corrected by a method of imputation. As for calibration, the fact of having errors in the auxiliary variables is not very important as long as the calibration is done on the total of the auxiliary variables of the register. Indeed, with balanced sampling, the Horvitz-Thompson estimator is used and is unbiased even if the auxiliary variables are false. The gain in efficiency only depends on the correlation between the balancing variables and the interest variables. This correlation is rarely affected by errors in the balancing variables.

Several variables can be used to improve small domain estimates. To ensure that a domain D is not empty, one can simply add the auxiliary variable:

$$x_k = \begin{cases} \pi_k & \text{if } k \in D \\ 0 & \text{otherwise,} \end{cases}$$

which implies that the number of sampled units that belong to D is equal to

$$n_D = \sum_{k \in U} x_k = \sum_{k \in D} \pi_k,$$

if n_D is integer, or one of the closest two integers to n_D if n_D is not an integer.

In some cases, it is interesting to balance on auxiliary variables in subgroups, domains or strata. An interesting procedure described in Chauvet (2009) consists of separately running the flight phase in each stratum. A rounding problem will then occur in each stratum. These rounding problems can then be merged and a flight phase can be run again on the whole population. Finally, the landing phase is applied only to the whole population. This procedure enables us to roughly satisfy the balancing equations in each strata without cumulating the rounding problems.

The inclusion probabilities must be computed prior to sampling. When a linear model is assumed, these probabilities should in principle be proportional to the errors of the model in order to minimize variance (see Tillé and Favre 2005; Chauvet, Bonnery and Deville 2010a; Nedyalkova and Tillé 2009, 2010). This choice generalizes Neyman's allocation for stratified sampling (Neyman 1934). However, the inclusion probabilities often need to be chosen on others constraints. For instance, in order to construct the rotation groups of the French census, the inclusion probabilities must all be equal to a fifth.

7.4 Balancing versus calibration

Stratification is a particular case of balancing, while post-stratification is a particular case of calibration. In stratification and balancing, the weights do not become random. It is thus generally a better strategy. Nevertheless, more auxiliary information is needed for balancing. Indeed, for balanced sampling, the auxiliary variables must be known for all the units of the population, whereas, for calibration, only the population totals are needed. Balancing is a very interesting method for small population sizes. It is thus a very good method for selecting primary units in a multistage sampling design.

Both techniques can be used together. They are not contradictory. The best strategy consists of using balanced sampling and calibration together. Indeed calibration can resolve the small rounding problem that may remain after balancing. At the estimation stage, more auxiliary variables are often available because, in order to balance a sample, the auxiliary information must be known at the individual level while, in order to calibrate the sample, only the population totals are necessary.

Generally, it is recommended to re-calibrate on the balancing variables at the estimation stage even if more calibration variables are available. If only new variables are used in calibration, the effect of balancing can be lost. There is, however, one case where calibration can be used without re-calibrating on the balancing variables: when, conditionally on the calibration variables, we can reasonably assume that the balancing variables are no longer correlated to the variables of interest. This can occur when the balancing and the calibration variables are the same variables measured at different moments, and the calibration variables are more recent.

When the determination coefficient between the interest variable and the auxiliary variables is equal to or close to one, then calibration is more efficient because of the rounding problem of balanced sampling. Anyway the most efficient strategy always consists of using balanced sampling and calibration together (see the simulation in Deville and Tillé 2004).

7.5 Accuracy of the balancing equations

It is possible to prove, under realistic assumptions (see Deville and Tillé 2004), that with the cube method

$$\left| \frac{\widehat{X}_j - X_j}{X_j} \right| < O(p/n),$$

where p is the number of variables, and $O(x)/x$ is a quantity that remains bounded when x tends to infinity. With simple random sampling

$$\left| \frac{\widehat{X}_j - X_j}{X_j} \right| = O_p(\sqrt{1/n}),$$

where $O_p(x)/x$ is a quantity that remains bounded in probability when x tends to infinity.

The gains in accuracy are therefore considerable. The small rounding problem can be fixed by a small calibration. The rounding problem comes from the fact that selecting a sample is an integer problem. It also occurs in stratification, which is a particular case of balancing. In stratification with proportional allocation, the sums of the inclusion probabilities in the strata are generally not integers. So, the sample sizes in the strata are obtained by rounding the sum of inclusion probabilities in the strata. The cube method does this rounding automatically and randomly in such a way as to ensure that the inclusion probabilities are exactly satisfied.

7.6 Balanced sampling in repeated surveys

An important difficulty occurs in repeated sampling. The problem comes from the fact that, when a balanced sample is selected with unequal inclusion probabilities, the complementary sample is not necessarily balanced. Indeed, the equality

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k$$

does not imply that

$$\sum_{k \in U \setminus S} \frac{\mathbf{x}_k}{1 - \pi_k} = \sum_{k \in U} \mathbf{x}_k.$$

This problem occurred in the French master sample. In this sampling design, the primary units, which are geographical

areas, are selected with unequal probabilities that are proportional to the size. After selecting the sample, some regions asked for complementary samples of areas that were not already selected. This question is intricate, because the complementary sample of a balanced sample is no longer balanced, and the aim is thus to select a balanced sample from a part of the population that is no longer balanced. Tillé and Favre (2004) gave a few methods to co-ordinate balanced samples, which were selected with unequal inclusion probabilities. More generally, the coordination (in the sense of managing overlap) of balanced samples can be difficult when the sampling design is balanced.

While challenging, it is possible to organize rotations if all the samples are selected together and the samples are selected with equal inclusion probabilities. Indeed, in this case the complementary $\overline{S} = U \setminus S$ of the samples S is also a balanced sample. A second balanced sample can be directly selected from $\overline{S}$ and so on. This method was used to create five rotation groups in the French master sample. The five groups are five balanced samples of municipalities.

If the samples are selected with unequal inclusion probabilities, some solutions are described in Tillé and Favre (2004). An interesting particular case can easily be solved: when two non-overlapping samples must be selected with the same unequal inclusion probabilities $\pi_k < 0.5$ from the same population. First, a sample S_A balanced on $\mathbf{x}_k$ must be selected with inclusion probabilities $\pi_{kA} = 2\pi_k$ such that

$$\sum_{k \in S_A} \frac{\mathbf{x}_k}{2\pi_k} = \sum_{k \in U} \mathbf{x}_k .$$

Next, a sample S_1 can be selected from S_A. This sample must be selected with inclusion probability $\pi_{kB} = 0.5$ and must be balanced on $\mathbf{x}_k / 2\pi_k$, which gives the following balancing equations:

$$\sum_{k \in S_2} \frac{\mathbf{x}_k / (2\pi_k)}{1/2} = \sum_{k \in S_A} \frac{\mathbf{x}_k}{2\pi_k} = \sum_{k \in U} \mathbf{x}_k .$$

The sample $S_2 = S_A \setminus S_1$ is also balanced.

If the population changes over times (deaths and births), the organization of a rotation becomes much more difficult. This difficulty already occurs with stratified samples. Nevertheless, for stratification, several reasonable solutions exist (see, amongst others, De Ree 1999; Hesse 1998; Rivière 1999; Nedyalkova, Péa and Tillé 2006).

7.7 Main implementations of balanced sampling

An SAS/IML® implementation was first programmed by three students of the École nationale de la statistique et de l'analyse de l'information (Ensai) (Bousabaa *et al.* 1999). An official version of the *Institut National de la Statistique et des Études Économiques* done by Tardieu (2001) and Rousseau and Tardieu (2004) is now available on the Insee Web site. Another SAS/IML® version done by Chauvet and Tillé (2005b, a, 2006) is also available on the University of Neuchâtel Web site. In R language, the sampling package (Tillé and Matei 2007) allows us to use the cube method. These software programs are free, available over the Internet and are easy to use.

The available programs written using R language or SAS/IML® have no limit as far as population size is concerned. An application with 40 balanced variables is possible. In order to select the sample, the computation times increase with $N \times p^2$, where N is the population size and p the number of balancing variables. It is thus possible to select a sample in a population of several million statistical units.

Acknowledgements

This paper has been written in response to an invitation to speak at the Demographic Statistical Methods Division Seminar of the U.S. Census Bureau in June 2008. The author would like to thank the U.S. Census Bureau and particularly Patrick Flanagan without whom this paper would never have been written. The author is also grateful to an associate editor and two anonymous reviewers for valuable comments and corrections that helped to improve this paper.

References

Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique*, 23, 91-113.

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Bardaji, J. (2001). Un an après la sortie d'un contrat emploi consolidé : près de six chances sur dix d'avoir un emploi. *Premières Informations Synthèses*, *Direction de l'Animation de la Recherche des Etudes et des Statistiques* (*DARES*) *du Ministère du Travail des relations sociales et de la solidarité*, 43, 3, 1-8.

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

Biggeri, L., and Falorsi, P.D. (2006). A probability sample strategy for improving the quality of the consumer price index survey using the information of the business register. In *Proceedings of the Conference of European Statisticians Group of Experts on Consumer Price Indices*, Eighth Meeting, Geneva, 10-12 May 2006.

Bousabaa, A., Lieber, J. and Sirolli, R. (1999). La macro cube. Technical report, Ensai, Rennes.

Breidt, F.J., and Chauvet, G. (2010a). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Breidt, F.J., and Chauvet, G. (2010b). Penalized balanced sampling. Working paper, Ensai.

Brewer, K.R.W. (1975). A simple procedure for πpswor. *Australian Journal of Statistics*, 17, 166-172.

Brewer, K.R.W. (1999). Design-based or prediction-based inference? Stratified random *vs* stratified balanced sampling. *International Statistical Review*, 67, 35-47.

Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. PhD thesis, Université Rennes 2.

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

Chauvet, G., Bonnery, D. and Deville, J.-C. (2010a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 2, 984-994.

Chauvet, G., Deville, J. and Haziza, D. (2010b). Adapting the cube algorithm for balanced random imputation in surveys. Technical report, Ensai, Rennes.

Chauvet, G., Deville, J. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*.

Chauvet, G., and Tillé, Y. (2005a). *Fast SAS Macros for balancing Samples*: *user's guide*. Software Manual, University of Neuchâtel, http://www2.unine.ch/statistics/page10890.html.

Chauvet, G., and Tillé, Y. (2005b). New SAS macros for balanced sampling. In *Journées de Méthodologie Statistique*, Insee, Paris.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21, 9-31.

Chipperfield, J. (2009). An evaluation of cube sampling for ABS household surveys. Technical report, Australian Bureau of Statistics.

Christine, M. (2006). Use of balanced sampling in the framework of the master sample for french household surveys. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Christine, M., and Wilms, L. (2003). Theoretical and practical problems related to the development of "EMEX": How to improve the precision of the regional supplements of National Surveys with an Additional Sample? In *Proceedings*: *Symposium 2003*, *Challenges in Survey Taking for the Next Decade*, Statistics Canada, Ottawa.

Cumberland, W.G., and Royall, R.M. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

Cumberland, W.G., and Royall, R.M. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, B, 50, 118-124.

da Silva, A.D., da Silva Borges, A., Aires Leme, R. and Moura Reis Miceli, A.P. (2006). Modalidades alternativas de censo demográfico: o cenário internacional a partir das experiências dos estados unidos, França, holanda, israel e alemanha. Technical report, Instituto Brasileiro de Geografia e Estatística.

D'Alò, M., Di Consiglio, L., Falorsi, S. and Solari, F. (2006). Small area estimation of the italian poverty rate. *Statistics in Transition*, 7, 771-784.

De Ree, S.J.M. (1999). Co-ordination of business samples using measured response burden. In *Contributed paper*, *52th Session of the ISI Helsinki*.

Desplanques, G. (2000). La rénovation du recensement de la population. In *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 2-5.

Dessertaine, A. (2006). Sondages et séries temporelles : une application pour la prévision de la consommation électrique. In *Actes des journées Françaises de Statistique 2006*, Clamart, France.

Dessertaine, A. (2007). Sampling and data-stream: Some ideas to built balanced sampling using auxiliary Hilbertian informations. In *Proceedings of 56th the International Statistical Institute Conference*: *IPM56 - New methods of sampling*, Lisboa, Portugal.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro (Sweden).

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. In *Recueil de la Section des méthodes d'enquêtes des communications présentées au 26ème congrès de la Société Statistique du Canada*, 103-110, Sherbrooke.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Deville, J.-C. (2005). Imputation stochastique et échantillonnage équilibré. Technical report, École Nationale de la Statistique et de l'Analyse de l'Information.

Deville, J.-C. (2006). Stochastic imputation using balanced sampling. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Deville, J.-C., Grosbras, J.-M. and Roth, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT*, *Proceedings in Computational Statistics*, Heidelberg. Physica Verlag, 255-266.

Deville, J.-C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

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.

Dudoignon, L., and Vanheuverzwyn, A. (2006). Tirage d'un échantillon à probabilités inégales : application au panel Médiamat. In *Actes de des Journées de Méthodologie Statistique*, 1-10.

Dumais, J., Bertrand, P. and Kauffmann, B. (2000). Sondage, estimation et précision dans la rénovation du recensement de la population. In *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 6-26.

Dumais, J., and Isnard, M. (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*, Paris. Insee, 100, 37-76.

Durr, J.-M., and Dumais, J. (2001). Redesign of the french census of population. In *Proceedings*: *Symposium 2001*, *Achieving Data Quality in a Statistical Agency*: *A Methodological Perspective*, Statistics Canada, Ottawa.

Durr, J.-M., and Dumais, J. (2002). Redesign of the french census of population. *Survey Methodology*, 28, 43-49.

Even, K. (2002). Improved tool for evaluating employment and vocational training policy: Panel of beneficiaries. *Premières Informations Synthèses*, *Direction de l'Animation de la Recherche des Études et des Statistiques (DARES) du Ministère du Travail des relations sociales et de la solidarité*, 33, 1, 1-7.

Falorsi, P.D., and Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34, 223-234.

Fecteau, S., and Jocelyn, W. (2006). Une application de l'échantil-lonnage équilibré : le plan de sondage des entreprises non incoprporées. In *Méthodes d'enquêtes et sondages* : *pratiques européenne et nord-américaine*, (Eds., P. Lavallée and L.-P. Rivest), Paris. Dunod, 405-410.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A. (2010). Replication variance estimation for rejective sampling. In *Seminar of Statistics Canada*, June 2010, Ottawa.

Gini, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne (1er décembre 1921). *Bulletin of the International Statistical Institute*, 23, 2, 198-215.

Gini, C., and Galvani, L. (1929). Di una applicazione del metodo rappresentative all'ultimo censimento Italiano della popolazione (1° decembri, 1921). *Annali di Statistica*, Series 6, 4, 1-107.

Gismondi, R. (2007). Quick estimation of tourist nights spent in italy. *Statistical Methods and Applications*, 16, 141-168.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Hedayat, A.S., and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44, 237-247.

Hesse, C. (1998). Sampling co-ordination: A review by country. Technical Report E9908, Direction des Statistique d'Entreprises, Insee, Paris.

Jocelyn, W. (2006). Sampling and estimation strategies for the canadian unincorporated business population. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Kiaer, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9, 2, 176-183.

Kiaer, A. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 1, 180-185.

Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 13, 1, 66-78.

Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 1, 119-134.

Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40, 202-204.

Langel, M., and Tillé, Y. (2010). Corrado Gini, a pioneer in balanced sampling and inequality theory. Technical report, University of Neuchatel.

Legg, J.C., and Yu, C.L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, 36, 69-79.

Lesage, E. (2008). Contraintes d'équilibrage non linéraires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Paris. Dunod, 285-289.

Marí, G., Barbará, G., Mitas, G. and Passamonti, S. (2007a). Construcción de un estimador de variancia para muestras balanceadas estratificadas. In *XXXV Coloquio Argentino de Estadística. Mar del Plata*, Argentina. 22, 23 y 24 de Octubre de 2007.

Marí, G., Barbará, G., Mitas, G. and Passamonti, S. (2007b). Muestras equilibradas en poblaciones finitas: un estudio comparativo en muestras de explotaciones agropecuarias. In *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, noviembre de 2007*, Universidad Nacional de Rosario, Argentina.

Nedyalkova, D., Péa, J. and Tillé, Y. (2006). A review of some current methods of coordination of stratified samples. introduction and comparison of new methods based on microstrata. Technical report, Université de Neuchâtel.

Nedyalkova, D., and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.

Nedyalkova, D., and Tillé, Y. (2010). Bias robustness and efficiency in model-based inference. Technical report, University of Neuchâtel.

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

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School; U.S. Department of Agriculture, Washington.

Périé, P. (2008). Échantillonnage à entropie maximale sous contraintes : un algorithme rapide basé sur l'optimisation linéaire en nombres binaires. In *Méthodes d'enquêtes* : *applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Paris. Dunod, 294-299.

Rivière, P. (1999). Coordination of samples: The microstrata methodology. In *13th International Roundtable on Business Survey Frames*, Paris. Insee.

Rousseau, S., and Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, Insee, Paris.

Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. *Biometrika*, 63, 605-614.

Royall, R.M. (1976b). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R.M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics Volume 6*: *Sampling*, (Eds., P.R. Krishnaiah and C.R. Rao), Amsterdam. Elsevier/North-Holland, 399-413.

Royall, R.M., and Pfeffermann, D. (1982). Balanced samples and robust bayesian inference in finite population sampling. *Biometrika*, 69, 401-409.

Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

Tardieu, F. (2001). Échantillonnage équilibré: de la théorie à la pratique. Technical report, Insee, Paris.

Thionet, P. (1953). *La théorie des sondages*. Insee, Imprimerie nationale, Paris.

Tillé, Y. (2001). *Théorie des sondages* : *échantillonnage et estimation en populations finies*. Dunod, Paris.

Tillé, Y. (2006a). Balanced sampling by means of the cube method. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Tillé, Y. (2006b). *Sampling Algorithms*. New York: Springer.

Tillé, Y., and Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91, 913-927.

Tillé, Y., and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Tillé, Y., and Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, http://cran. r-project. org/, Manual of the Contributed Packages.

Tirari, M. (2006). Le plan de sondage équilibré et l'estimation du total d'une population finie. In *Méthodes d'enquêtes et sondages* : *pratiques européenne et nord-américaine*, (Eds., P. Lavallée and L.-P. Rivest), Paris, Dunod, 411-416.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York: John Wiley & Sons, Inc.

Wilms, L. (2000). Présentation de l'échantillon-maître en 1999 et application au tirage des unités primaires par la macro cube. In *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109, 12-43.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Charles Griffin, London, England, third edition.