Biometrika (2004), **91,** 4, *pp.* 893–912

Printed in Great Britain

Efficient balanced sampling: The cube method

BY JEAN-CLAUDE DEVILLE

Laboratoire de Statistique d'Enquête, CREST–ENSAI,
École Nationale de la Statistique et de l'Analyse de l'Information, rue Blaise Pascal,
Campus de Ker Lann, 35170 Bruz, France

deville@ensai.fr

AND YVES TILLÉ

Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 805,
2002 Neuchâtel, Switzerland

yves.tille@unine.ch

SUMMARY

A balanced sampling design is defined by the property that the Horvitz–Thompson estimators of the population totals of a set of auxiliary variables equal the known totals of these variables. Therefore the variances of estimators of totals of all the variables of interest are reduced, depending on the correlations of these variables with the controlled variables. In this paper, we develop a general method, called the cube method, for selecting approximately balanced samples with equal or unequal inclusion probabilities and any number of auxiliary variables.

Some key words: Calibration; Poststratification; Quota sampling; Sampling algorithm; Stratification; Sunter's method; Unequal selection probabilities.

1. INTRODUCTION

The use of auxiliary information is a central issue in survey sampling from finite populations. The classical techniques that use auxiliary information in a sampling design are stratification (Neyman, 1934; Tschuprow, 1923) and unequal probability sampling or sampling proportional to size (Hansen & Hurwitz, 1943; Madow, 1949).

The problem of balanced sampling is an old one and has not yet been solved. Kiaer (1896), founder of modern sampling, argued for samples that match the means of known variables to obtain what he called 'representative samples'. He advocated purposive methods before the development of the idea of probability sampling proposed by Neyman (1934, 1938). Yates (1949) also insisted on the idea of respecting the means of known variables in probability samples because the variance is then reduced. Yates (1946) and Thionet (1953, pp. 203–7) have described limited and heavy methods of balanced sampling. Hájek (1964; 1981, p. 157) gives a rigorous definition of a representative strategy and its properties. According to Hájek, a strategy is a pair composed of a sampling design and an estimator, the strategy being representative if it estimates exactly the total of an auxiliary variable. He showed that a representative strategy could be achieved by regression, but he did not succeed in finding a representative sampling method associated

with the Horvitz–Thompson estimator other than the rejective procedure, which consists of selecting new samples until a balanced sample is found. Royall & Herson (1973) stressed the importance of balancing a sample in order to protect the inference against a misspecified model. They called this idea 'robustness'. Since no method existed for achieving a multivariate balanced sample, they proposed the use of simple random sampling, which is 'mean-balanced' with large samples. Several partial solutions were proposed by Deville et al. (1988), Deville (1992), Ardilly (1991) and Hedayat & Majumdar (1995), but a general solution for balanced sampling was never found. Recently, Valliant et al. (2000) surveyed some existing methods.

In this paper, we propose a general method, the cube method, that allows the selection of approximately balanced samples, in that the Horvitz–Thompson estimates for the auxiliary variables are equal, or nearly equal, to their population totals. The method is appropriate for a large set of qualitative or quantitative balancing variables, it allows unequal inclusion probabilities, and it permits us to understand how accurately a sample can be balanced. Moreover, the sampling design respects any fixed, equal or unequal, inclusion probabilities. The method can be viewed as a generalisation of the splitting procedure (Deville & Tillé, 1998) which allows easy construction of new unequal probability sampling methods.

Since its conception, the cube method has aroused great interest amongst survey statisticians at the Institut National de la Statistique et des Études Économiques (INSEE), the French Bureau of Statistics. A first application of the method was implemented in SAS-IML by A. Bousabaa, J. Lieber, R. Sirolli and F. Tardieu. This macro allows the selection of samples with unequal probabilities of up to 50 000 units and 30 balancing variables. The INSEE has adopted the cube method for its most important statistical projects. In the redesigned census in France, a fifth of the municipalities with fewer than 5000 inhabitants are sampled each year, so that after five years all the municipalities will be selected. All the households in these municipalities are surveyed. The five samples of municipalities are selected with equal probabilities using the cube method and are balanced on a set of demographic variables (Dumais & Isnard, 2000).

The demand for such sampling methods is very strong. In the French National Statistical Institute, the use of balanced sampling in several projects improved efficiency dramatically, allowing a reduction of the variance by 20 to 90% in comparison to simple random sampling.

2. Formulation of the problem, with examples

Consider a finite population U of size N whose units can be identified by labels $k \in \{1, \ldots, N\}$. The aim is to estimate the total $Y = \sum_{k \in U} y_k$ of a variable of interest y that takes the values y_k $(k \in U)$ for the units of the population. Suppose also that the vectors of values $x_k = (x_{k1} \ldots x_{kj} \ldots x_{kp})'$ taken by p auxiliary variables are known for all the units of the population. The p vectors $(x_{1j} \ldots x_{kj} \ldots x_{Nj})'$, for $j = 1, \ldots, p$, are assumed without loss of generality to be linearly independent.

A sample is denoted by a vector $s = (s_1 \ldots, s_k \ldots, s_N)'$, 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 $\mathscr{S} = \{0, 1\}^N$ of all the possible samples. The random sample S takes the value s with probability $\mathrm{pr}(S = s) = p(s)$. The inclusion probability of unit k is the probability $\pi_k = \mathrm{pr}(S_k = 1)$ that unit k is in the sample and the joint inclusion probability is the probability $\pi_{k\ell} = \mathrm{pr}(S_k = 1 \text{ and } S_\ell = 1)$ 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 jth auxiliary total $X_j = \sum_{k \in U} x_{kj}$ is $\hat{X}_j = \sum_{k \in U} S_k x_{kj} / \pi_k$. The Horvitz–Thompson estimator vector, $\hat{X} = \sum_{k \in U} S_k x_k / \pi_k$, estimates without bias the totals of the auxiliary variables, $X = \sum_{k \in U} x_k$.

The aim is to construct a balanced sampling design, defined as follows.

DEFINITION 1. *A sampling design $p(s)$ is said to be balanced on the auxiliary variables, $x_1, \ldots, x_p$, if and only if it satisfies the balancing equations given by*

$$\hat{X} = X, \tag{1}$$

which can also be written as

$$\sum_{k \in U} \frac{S_k x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

for all $s \in \mathscr{S}$ such that $p(s) > 0$.

Remark. If the y_k are linear combinations of the x_k, that is $y_k = x_k' b$ for all k, where b is a vector of constants, then $\hat{Y} = Y$. More generally, if the y_k are well predicted by a linear combination of the x_k, one can expect $\operatorname{var}(\hat{Y})$ to be small.

Next consider the following three particular cases of balanced sampling.

Example 1. A sampling design of fixed sample size n is balanced on the variable $x_k = \pi_k$ $(k \in U)$ because

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

Example 2. Suppose that the design is stratified and that, from each stratum U_h $(h = 1, \ldots, H)$ of size N_h, a simple random sample of size n_h is selected. Then the design is balanced on the variables δ_{kh}, where

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

In this case, we have

$$\sum_{k \in U} \frac{S_k \delta_{kh}}{\pi_k} = \sum_{k \in U} S_k \delta_{kh} \frac{N_h}{n_h} = N_h \quad (h = 1, \ldots, H).$$

Example 3. In sampling with unequal probabilities, when all the inclusion probabilities are different, the Horvitz–Thompson estimator $\hat{N} = \sum_{k \in U} S_k / \pi_k$ of the population size N is generally random. When the population size is known before selecting the sample, it could be important to select a sample such that

$$\sum_{k \in U} \frac{S_k}{\pi_k} = N. \tag{2}$$

Equation (2) is a balancing equation, in which the balancing variable is $x_k = 1$ $(k \in U)$. Until now, there has been no method by which (2) can be approximately satisfied for arbitrary inclusion probabilities, but we will see that this balancing equation can be satisfied by means of the cube method.

Stratification and unequal probability sampling are thus special cases of balanced sampling. In § 6, we present new cases, but the main practical interest of balanced sampling lies in its generality. Nevertheless, in most cases, the balancing equations (1) cannot be exactly satisfied, as the following example shows.

Example 4. Suppose that $N = 10$, $n = 7$, $\pi_k = \frac{7}{10}$ $(k \in U)$ and that the only auxiliary variable is $x_k = k$ $(k \in U)$. Then a balanced sample satisfies

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

so that $\sum_{k \in U} kS_k$ has to be equal to $55 \times \frac{7}{10} = 38{\cdot}5$, which is impossible because 38·5 is not an integer. The problem arises because $1/\pi_k$ is not an integer and the population size is small.

Consequently, our objective is to construct a sampling design which satisfies the balancing equations (1) exactly if possible, and to find the best approximation if this cannot be achieved. The rounding problem becomes negligible when the expected sample size is large.

3. CUBE REPRESENTATION OF BALANCED SAMPLES

The cube method is based on a geometric representation of the sampling design. The 2^N possible samples correspond to 2^N vectors of $\mathbb{R}^N$ in the following way. Each vector s is a vertex of an N-cube, and the number of possible samples is the number of vertices of an N-cube. A sampling design with inclusion probabilities π_k $(k \in U)$ consists of assigning a probability $p(s)$ to each vertex of the N-cube such that

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

where $\pi = (\pi_k)$ is the vector of inclusion probabilities. Geometrically, a sampling design consists of expressing the vector π as a convex combination of the vertices of the N-cube.

A sampling algorithm can thus be viewed as a 'random' way of reaching a vertex of the N-cube from a vector π in such a way that the balancing equations (1) are satisfied. Figure 1 shows the geometric representation of the possible samples from a population of size $N = 3$.

The cube method is composed of two phases called the flight phase and the landing phase. In the flight phase, the constraints are exactly satisfied. The objective is to round

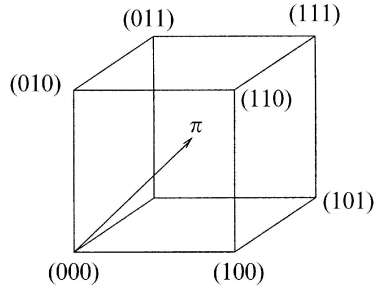

Fig. 1. Geometric representation of possible samples in a population of size $N = 3$.

off randomly to 0 or 1 almost all the inclusion probabilities. The landing phase consists of coping as well as possible with the fact that the balancing equations (1) cannot always be satisfied exactly.

The balancing equations (1) can also be written

$$\sum_{k \in U} a_k s_k = \sum_{k \in U} a_k \pi_k, \quad s_k \in \{0, 1\}, \quad k \in U, \tag{3}$$

where $a_k = x_k/\pi_k$ ($k \in U$) and s_k equals 1 if unit k is in the sample and 0 otherwise. The first equation of (3) with given a_k and coordinates s_k defines a hyperplane Q in $\mathbb{R}^N$ of dimension $N - p$. Note that $Q = \pi + \ker A$, where $\ker A$ is the kernel or null-space of the $p \times N$ matrix A given by $A = (a_1 \ldots a_k \ldots a_N)$. The main idea in obtaining a balanced sample is to choose a vertex of the N-cube that remains in the hyperplane Q or near to Q if that is not possible.

If $C = [0, 1]^N$ denotes the N-cube in $\mathbb{R}^N$ whose vertices are the samples of U, the intersection between C and Q is nonempty, because π is in the interior of C and belongs to Q. The intersection between an N-cube and a hyperplane defines a polytope $K = C \cap Q$, which is of dimension $(N - p)$ because it is the intersection of an N-cube and a plane, of dimension $(N - p)$, that has a point in the interior of C.

DEFINITION 2. *Let D be a convex polyhedron. A vertex, or extremal point, of D is defined as a point that cannot be expressed as a convex linear combination of other points of D. The set of all the vertices of D is denoted by* $\mathrm{Ext}(D)$.

DEFINITION 3. *A sample s is said to be exactly balanced if* $s \in \mathrm{Ext}(C) \cap Q$.

Note that a necessary condition for finding an exactly balanced sample is that $\mathrm{Ext}(C) \cap Q \neq \varnothing$.

DEFINITION 4. *A balancing equation system is*

(i) *exactly satisfied if* $\mathrm{Ext}(C) \cap Q = \mathrm{Ext}(C \cap Q)$,
(ii) *approximately satisfied if* $\mathrm{Ext}(C) \cap Q = \varnothing$,
(iii) *sometimes satisfied if* $\mathrm{Ext}(C) \cap Q \neq \mathrm{Ext}(C \cap Q)$ *and* $\mathrm{Ext}(C) \cap Q \neq \varnothing$.

Whether the balancing equation system is exactly satisfied, approximately satisfied or sometimes satisfied depends on the values of π and A.

PROPOSITION 1. *If $r = (r_k)$ is a vertex of K then $\#\{k | 0 < r_k < 1\} \leqslant p$, where p is the number of auxiliary variables, and $\#(B)$ denotes the cardinality of a set B.*

Proof. Let A^* be the submatrix of A consisting of the columns corresponding to non-integer components of the vector r. If $q = \#(U^*) > p$, then $\ker A^*$ has dimension $q - p > 0$, and r is not an extreme point of K. □

The following three examples show that the rounding problem can be viewed geometrically. Indeed, the balancing equations cannot be exactly satisfied when the vertices of K are not vertices of C, that is when $q > 0$.

Example 5. In Fig. 2(a), a sampling design for a population of size $N = 3$ is considered. The only constraint consists of fixing the sample size $n = 2$, and thus $p = 1$ and $x_k = \pi_k$ ($k \in U$). The inclusion probabilities satisfy $\pi_1 + \pi_2 + \pi_3 = 2$, so that the balancing equation is exactly satisfied.

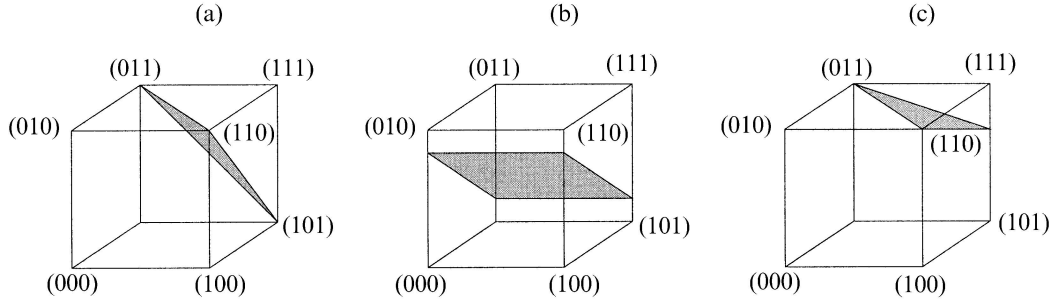

Fig. 2. Geometric representation of rounding problem. (a) shows fixed size constraint when all the vertices of K are vertices of the cube. In (b) none of vertices of K is a vertex of the cube. In (c) some vertices of K are vertices of the cube and others are not.

Example 6. Figure 2(b) exemplifies the case where the constraint hyperplane does not pass through any vertices of the cube. The inclusion probabilities are $\pi_1 = \pi_2 = \pi_3 = 0{\cdot}5$. The only constraint is given by the auxiliary variables $x_1 = 0$, $x_2 = 6\pi_2$ and $x_3 = 4\pi_3$. It is then impossible to satisfy the balancing equation exactly, but the balancing equation is always satisfied approximately.

Example 7. Figure 2(c) exemplifies the case where the constraint hyperplane passes through two vertices of the cube but one vertex of the intersection is not a vertex of the cube. The inclusion probabilities are $\pi_1 = \pi_2 = \pi_3 = 0{\cdot}8$. The only constraint, $p = 1$, is given by the auxiliary variable $x_1 = \pi_1$, $x_2 = 3\pi_2$ and $x_3 = \pi_3$. The balancing equation is only sometimes satisfied. In this case there exist balanced samples, but there does not exist an exactly balanced sampling design for the given inclusion probabilities. In other words, although exactly balanced samples exist, one must accept selection of only approximately balanced samples in order to satisfy the given inclusion probabilities.

It seems hard to find a general method for detecting when the balancing equations can be satisfied exactly. It depends on complex patterns in $\mathbb{R}^N$. These problems are already very intricate in $\mathbb{R}^3$ and are treated in crystallography; for a short introduction to this topic see for instance Doubrovine et al. (1979, pp. 173–95).

4. THE FLIGHT PHASE

At the end of the flight phase, a vertex of K is chosen randomly in such a way that the inclusion probabilities π_k ($k \in U$) and the balancing equations (1) are exactly satisfied. The landing phase is necessary only if the attained vertex of K is not a vertex of C, and consists of relaxing the constraints (1) as little as possible in order to select a sample, i.e. a vertex of C.

The general method for completing the flight phase is to use a balancing martingale.

DEFINITION 5. *A discrete time stochastic process* $\pi(t) = \{\pi_k(t)\}$, *for* $t = 0, 1, \ldots,$ *in* $\mathbb{R}^N$ *is said to be a balancing martingale for an inclusion probability vector* π *and auxiliary variables* $x_1, \ldots, x_p$ *if*

(i) $\pi(0) = \pi$;

(ii) $E\{\pi(t)|\pi(t-1), \ldots, \pi(0)\} = \pi(t-1)$, *for* $t = 1, 2, \ldots$;

(iii) $\pi(t) \in K = C \cap Q$, *where* A *is the* $p \times N$ *matrix given by* $A = (x_1/\pi_1 \ldots x_k/\pi_k \ldots x_N/\pi_N)$.

PROPOSITION 2. *If $\pi(t)$ is a balancing martingale, then we have the following*:
(i) $E\{\pi(t)\} = E\{\pi(t-1)\} = \ldots = E\{\pi(0)\} = \pi$;
(ii) $\sum_{k \in U} a_k \pi_k(t) = \sum_{k \in U} a_k \pi_k = X$, *for* $t = 0, 1, 2, \ldots$;
(iii) *when the balancing martingale reaches a face of C, it does not leave it.*

Proof. Part (i) is obvious. Part (ii) holds because $\pi(t) \in K$. To prove (iii), note that $\pi(t-1)$ belongs to a face; it is the mean of the possible values of $\pi(t)$ that therefore must also belong to this face. □

Part (iii) of Proposition 2 directly implies that (a) if $\pi_k(t) = 0$ then $\pi_k(t+h) = 0$, for $h \geqslant 0$; (b) if $\pi_k(t) = 1$ then $\pi_k(t+h) = 1$, for $h \geqslant 0$; and (c) the vertices of K are absorbing states.

The practical problem is to find a method that rapidly reaches a vertex. The following procedure allows us to attain a vertex of K in at most N steps. First initialise at $\pi(0) = \pi$. Next, at time $t = 1, \ldots, T$, repeat the following three steps.

Step 1. Generate any vector $u(t) = \{u_k(t)\} \neq 0$, such that $u(t)$ is in the kernel of the matrix A, and $u_k(t) = 0$ if $\pi_k(t-1)$ is an integer. The vector $u(t)$ can be chosen randomly or deterministically but $u(t)$ must be independent of $\pi(t-1), \ldots, \pi(1)$.

Step 2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that

$$0 \leqslant \pi(t-1) + \lambda_1(t)u(t) \leqslant 1, \quad 0 \leqslant \pi(t-1) - \lambda_2(t)u(t) \leqslant 1.$$

Note that $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$.

Step 3. Select

$$\pi(t) = \begin{cases} \pi(t-1) + \lambda_1^*(t)u(t), & \text{with probability } q(t), \\ \pi(t-1) - \lambda_2^*(t)u(t), & \text{with probability } 1 - q(t), \end{cases} \tag{4}$$

where $q(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$.

This general procedure is repeated until it is no longer possible to carry out Step 1.

The above procedure defines a balancing martingale. Clearly, $\pi(0) = \pi$. Also, from expression (4) we obtain $E\{\pi(t)|\pi(t-1), \ldots, \pi(0)\} = \pi(t-1)$, for $t = 1, 2, \ldots$, because

$$E\{\pi(t)|\pi(t-1), u(t)\} = \pi(t-1) \quad (t = 1, 2, \ldots).$$

Finally, since $u(t)$ is in the kernel of A, from (4) we obtain that $\pi(t)$ always remains in $K = C \cap Q$.

At each step, at least one component of the process is rounded to 0 or 1. Thus $\pi(1)$ is on a face of the N-cube, i.e. on a cube of dimension $N-1$ at most, $\pi(2)$ is on a cube of dimension $N-2$ at most, and so on. Let T be the time when the flight phase has stopped. The fact that Step 1 is no longer possible shows that the balancing martingale has attained a vertex of K, and thus by Proposition 1 that $\#U_T \leqslant p$, where $U_t = \{k|0 < \pi_k(t) < 1\}$.

The flight phase was implemented in a SAS-IML macro. For generating the vector $u(t)$, we first generate any, random or not, vector $v(t) = \{v_k(t)\}$ in $\mathbb{R}^N$, that is independent of $\pi(t-1), \ldots, \pi(1)$. Next $v(t)$ is projected on to the constraint hyperplane. Let $W_t = \text{diag}\{w_k(t)\}$ where $w_k(t) = 0$ if $k \notin U_t$. Then we choose

$$u(t) = W_t v(t) - W_t A'(AW_t A')^- AW_t v(t), \tag{5}$$

where B^- denotes a generalised inverse of a matrix B. The weights $w_k(t)$ allow a change of metric that is used, for example, in § 7·4 to generalise Sunter's method.

5. The landing phase

At the end of the flight phase, the balancing martingale has reached a vertex of K, which is not necessarily a vertex of C. This vertex is denoted by $\pi^* = (\pi_k^*) = \pi(T)$. Let q be the number of non-integer components of this vertex. If $q = 0$, the algorithm is completed. If $q > 0$ then only some constraints were exactly attained. Thus we shall seek a sampling design that minimises the function

$$\operatorname{var}(\hat{X}) = \sum_{k \in U} \sum_{\ell \in U} \frac{x_k x_k'}{\pi_k \pi_\ell} \Delta_{k\ell} = A \Delta A', \tag{6}$$

where A is the $p \times N$ matrix given by $A = (x_1/\pi_1 \ldots x_k/\pi_k \ldots x_N/\pi_N)$,

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell, & \text{if } k \neq \ell, \\ \pi_k(1 - \pi_k), & \text{if } k = \ell, \end{cases}$$

and $\Delta = (\Delta_{k\ell}) = \operatorname{var}(S)$. The matrix Δ, can be split into two parts, $\Delta = \Delta_F + E\Delta_{L|\pi^*}$, where Δ_F is the part related to the flight phase, $\Delta_F = \operatorname{var} E(S|\pi^*) = \operatorname{var}(\pi^*)$, and $\Delta_{L|\pi^*}$ is related to the landing phase:

$$\Delta_{L|\pi^*} = \operatorname{var}(S|\pi^*) = \sum_{s \in S} p(s|\pi^*)(s - \pi^*)(s - \pi^*)',$$

where $p(s|\pi^*)$ is the probability of selecting the sample s given that the flight phase has ended at π^*.

Since $A\Delta_F A' = 0$,

$$\operatorname{var}(\hat{X}) = E \operatorname{var}(\hat{X}|\pi^*) = E(A\Delta_{L|\pi^*} A').$$

At the end of the flight phase we must therefore find the sampling design $p(s|\pi^*)$ that minimises a function of the matrix, $\operatorname{var}(\hat{X}|\pi^*)$. In the convex cone of symmetric positive semidefinite matrices, there does not exist a unique minimal element generated by a design with expectation π^*. In practice, we have to limit ourselves to positive linear forms on the ordered vector space of symmetric matrices. Therefore, let $M = (m_{ij})$ be any $p \times p$ positive semidefinite matrix. Any such positive linear form is the M-trace of $\operatorname{var}(\hat{X}|\pi^*)$, which is given by

$$\begin{aligned} M - \operatorname{tr} \operatorname{var}(\hat{X}|\pi^*) &= \operatorname{tr}\{M \times \operatorname{var}(\hat{X}|\pi^*)\} \\ &= \sum_{s \in S} p(s|\pi^*)(s - \pi^*)' A' M A (s - \pi^*). \end{aligned} \tag{7}$$

DEFINITION 6. *A sample s is said to be compatible with a vector π^* if $\pi_k^* = s_k$ for all k such that π_k^* is an integer. Let $\mathscr{C}(\pi^*)$ denote the set with 2^q elements of samples compatible with π^*.*

It is clear that we can limit ourselves to finding a design with mean value π^* and whose support is included in $\mathscr{C}(\pi^*)$.

If

$$C(s) = (s - \pi^*)' A' M A (s - \pi^*)$$

defines the 'cost' associated with sample s, minimising the M-trace consists of minimising the conditional mean cost with respect to π^*. The choice of the cost function is discussed in the Appendix. The conditional mean cost function is minimised if we solve the following linear program:

$$\min_{p(.|\pi^*)} \sum_{s \in \mathscr{C}(\pi^*)} C(s)p(s|\pi^*), \tag{8a}$$

subject to

$$\sum_{s \in \mathscr{C}(\pi^*)} p(s|\pi^*) = 1, \quad \sum_{s \in \mathscr{C}(\pi^*)|s \ni k} p(s|\pi^*) = \pi_k^* \quad (k \in U, 0 \leqslant p(s|\pi^*) \leqslant 1, s \in \mathscr{C}(\pi^*)). \tag{8b}$$

Let $U^* = \{k \in U | 0 < \pi_k^* < 1\}$, for $q = \#(U^*)$, and $\mathscr{S}^* = \{0, 1\}^q$. then (8) can also be written as

$$\min_{p^*(.)} \sum_{s^* \in \mathscr{S}^*} C(s^*)p^*(s^*), \tag{9a}$$

subject to

$$\sum_{s^* \in \mathscr{S}^*} p^*(s^*) = 1, \quad \sum_{s^* \in \mathscr{S}^*|s^* \ni k} p^*(s^*) = \pi_k^* \quad (k \in U^*, 0 \leqslant p^*(s^*) \leqslant 1, s^* \in \mathscr{S}^*). \tag{9b}$$

Since $q \leqslant p$, this linear program no longer depends on the population size but only on the number of balancing variables. In fact, it is restricted to 2^q possible samples, and with a modern computer it can be applied without difficulty to a balancing problem with a score of auxiliary variables.

A linear program always produces particular sampling designs defined in the following way.

Definition 7. *Let $p(.)$ be a sampling design for a population U with inclusion probabilities π_k, and let $\mathscr{B} = \{s|p(s) > 0\}$. A sampling design $p(.)$ is said to be defined on a minimal support if and only if there does not exist a subset $\mathscr{B}_0 \subset \mathscr{B}$ such that $\mathscr{B}_0 \neq \mathscr{B}$ and*

$$\sum_{s \in \mathscr{B}_0} p_0(s)s_k = \pi_k \quad (k \in U) \tag{10}$$

has a solution in $p_0(s)$.

Wynn (1977) has studied sampling designs defined on minimal supports, and Deville & Tillé (1998) developed a general method for selecting unequal probability sampling designs defined on minimal supports.

Proposition 3. *The linear program* (8) *has at least one solution defined on a minimal support.*

The proof follows directly from the fundamental theorem of linear programming; see for instance Luenberger (1973, p. 18).

If the number of auxiliary variables is too large for the linear program to be solved by a simplex algorithm, $q > 10$ in our experience, then, at the end of the flight phase, an auxiliary variable can be dropped. One constraint is thus relaxed, allowing return to the flight phase until it is no longer possible to 'move' within the constraint hyperplane. The constraints are thus relaxed successively. For this reason, it is necessary to order the auxiliary variables according to their importance so that the least important constraints are relaxed first. This naturally depends on the context of the survey.

6. QUALITY OF THE APPROXIMATION

The rounding problem can arise with any balanced sampling design. For instance, in stratification, the rounding problem arises when the sums of the inclusion probabilities within the strata are not integers, which is almost always the case in proportional stratification or optimal stratification. In practice the stratum sample sizes n_h are rounded either deterministically or randomly. Random rounding is used so as to satisfy the values of n_h in expectation. The purpose of the random rounding is to respect the initial inclusion probabilities.

The cube method also uses random rounding. In the particular case of stratification, it provides exactly the well-known method of random rounding of the sample sizes in the strata. With any variant of the landing phase, the difference of the Horvitz–Thompson from the total is bounded, because the rounding problem only depends on $q \leqslant p$ values.

PROPOSITION 4. *For any application of the cube method,*

$$|\hat{X}_j - X_j| \leqslant p \times \max_{k \in U} \left| \frac{x_{jk}}{\pi_k} \right|. \tag{11}$$

Proof. We have that

$$|\hat{X}_j - X_j| = \left| \sum_{k \in U} S_k \frac{x_{kj}}{\pi_k} - \sum_{k \in U} \frac{x_{kj}}{\pi_k} \pi_k \right| = \left| \sum_{k \in U} S_k \frac{x_{kj}}{\pi_k} - \sum_{k \in U} \frac{x_{kj}}{\pi_k} \pi_k^* \right|$$

$$\leqslant \sup_{F | \#F = q} \sum_{k \in F} \left| \frac{x_{kj}}{\pi_k} \right| \leqslant p \times \max_{k \in U} \left| \frac{x_{kj}}{\pi_k} \right|. \qquad \square$$

PROPOSITION 5. *If the sum of the inclusion probabilities is an integer, and if the sampling design has a fixed sample size, i.e. the auxiliary variables include the variable* $x_{1k} = \pi_k$, *then, for any application of the cube method,*

$$|\hat{X}_j - X_j| \leqslant (p-1) \times \max_{k \in U} \left| \frac{x_{kj}}{\pi_k} - \frac{N \bar{X}_j}{n} \right|, \tag{12}$$

where $\bar{X}_j = N^{-1} \sum_{k \in U} x_{kj}$.

Proof. With the cube method, we can always satisfy the fixed sample size constraint when the sum of the inclusion probabilities is an integer, which can be written as

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

Thus, at the end of the flight phase, at most $p-1$ values of π^* are not integers. We obtain

$$|\hat{X}_j - X_j| = \left| \sum_{k \in U} S_k \frac{x_{kj} - c\pi_k}{\pi_k} - \sum_{k \in U} \frac{x_{kj} - c\pi_k}{\pi_k} \pi_k \right| \leqslant (p-1) \times \max_{k \in U} \left| \frac{x_{kj} - c\pi_k}{\pi_k} \right|,$$

for any $c \in \mathbb{R}$. If $c = N\bar{X}_j/n$, we obtain Proposition 5. $\square$

This bound is a conservative bound of the rounding error, because we consider the worst case. Moreover, the bound is computed for a total, and must be considered relatively to the population size. Let $\alpha_k = \pi_k N/n$ $(k \in U)$. For almost all the sampling designs in common use, we can assume that $1/\alpha_k$ is bounded when $n \to \infty$, $N \to \infty$ and $n/N \to f$.

Note that, for a fixed sample size,

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

The bound for the estimation of the mean can thus be written

$$\frac{|\hat{X}_j - X_j|}{N} \leqslant \frac{p}{n} \times \max_{k\in U}\left|\frac{x_{kj}}{\alpha_k}\right| = O(p/n),$$

where $O(p/n)$ is a quantity that remains bounded when multiplied by n/p. The bound thus very quickly becomes negligible, if the sample size is large with respect to the number of balancing variables.

For comparison note that with a single-stage sampling design, such as simple random sampling or Bernoulli sampling, we have generally that

$$\frac{|\hat{X}_j - X_j|}{N} = O_p(1/\sqrt{n});$$

see for example Rosén (1972) and Isaki & Fuller (1982).

Despite the over-statement of the bound, the gain obtained by balanced sampling is very important. The rate of convergence is much faster for balanced sampling than for a usual sampling design. Moreover, in balanced sampling, the convergence is not only in probability but also deterministic. In practice, except for the case of very small sample sizes, the rounding problem is thus negligible.

Example 8. Suppose that $N = 100$, $n = 25$, $p = 2$, $\pi_k = 0{\cdot}25$ $(k \in U)$ and that two balancing variables are used, namely $x_{k1} = 1$ $(k \in U)$, that is fixed sample size, and $x_{k2} = k$ $(k \in U)$. Since the sum of the inclusion probabilities is an integer, the first constraint can be exactly satisfied. Now $\bar{X}_2 = N^{-1}\sum_{k\in U} x_{k2} = 50{\cdot}5$ and, if the sample is balanced,

$$\left|\frac{\hat{X}_{2\text{bal}} - X_2}{N}\right| \leqslant (p-1) \times \max_{k\in U}\left|\frac{x_{kj} - \bar{X}_2}{n}\right| = \frac{49{\cdot}5}{25} = 1{\cdot}98.$$

A set of simulations has been run to estimate the variance under balanced sampling, and we obtained

$$\sqrt{\{\text{var}_{\text{sim}}(\hat{X}_{2\text{bal}}/N)\}} = 1{\cdot}1746.$$

If a simple random sample of fixed sample size is selected,

$$S_{x2}^2 = \frac{1}{N-1}\sum_{k\in U}(x_{k2} - \bar{X}_2)^2 = 841{\cdot}667,$$

$$\sqrt{\{\text{var}\,(\hat{X}_{2\text{SRS}}/N)\}} = \sqrt{\left(\frac{N-n}{Nn}S_{x2}^2\right)} = 5{\cdot}0241.$$

Balanced sampling is thus much more accurate. The design effect is extremely small:

$$\text{Deff} = \frac{\text{var}_{\text{sim}}(\hat{X}_{\text{2bal}})}{\text{var}\,(\hat{X}_{\text{2SRS}})} = 0{\cdot}05465.$$

7. Particular cases of balanced sampling

7·1. *Poisson sampling*

The cube method can be used without an auxiliary variable. An interesting unbalanced sampling procedure is the Poisson sampling design, which consists of selecting unit k with inclusion probability π_k, independently of the other units. The sample size is thus random. Poisson sampling can be implemented using the cube method by defining $u(t)$ such that $u_t(t) = 1$ and $u_k(t) = 0$, if $k \neq t$. Next, $\lambda_1(t) = 1 - \pi_t$, $\lambda_2(t) = \pi_t$, and we define

$$\pi(t+1) = \begin{cases} (\pi_1(t) \ldots \pi_{t-1}(t)\ 1\ \pi_{t+1} \ldots \pi_N)', & \text{with probability } q(t), \\ (\pi_1(t) \ldots \pi_{t-1}(t)\ 0\ \pi_{t+1} \ldots \pi_N)', & \text{with probability } 1 - q(t), \end{cases}$$

where $q(t) = \pi_t$. Each unit is thus selected independently of the others.

7·2. *Simple random sampling*

Simple random sampling is a particular case of the cube method. Let

$$\pi = (n/N, \ldots, n/N, \ldots, n/N)',$$

and $x_k = n/N$ $(k \in U)$. At the first step, the projection of any vector $v(1)$ given in (5) with $w_k(t) = 1$ if $k \in U_t$ becomes

$$u_k(1) = v_k(1) - \frac{1}{N} \sum_{\ell \in U} v_\ell(1).$$

There are at least three ways of selecting a simple random sample without replacement. The first way begins the first step by projecting the vector

$$v(1) = (1\ 0\ \ldots\ 0)',$$

which gives $u(1) = ((N-1)/N, -1/N, \ldots, -1/N)'$. Then $\lambda_1(1) = (N-n)/(N-1)$, $\lambda_2(1) = n/(N-1)$ and

$$\pi(1) = \begin{cases} \left(1\ \dfrac{n-1}{N-1} \cdots \dfrac{n-1}{N-1}\right)', & \text{with probability } q(1), \\ \left(0\ \dfrac{n}{N-1} \cdots \dfrac{n}{N-1}\right)', & \text{with probability } 1 - q(1), \end{cases}$$

where $q(1) = \pi_1 = n/N$. This first step corresponds exactly to the classical sequential method, described in Fan et al. (1962), that produces a simple random sample without replacement. The second step consists of taking $v(1) = (0\ 1\ 0\ \ldots\ 0)'$.

The second way consists of sorting the data randomly before applying the cube method with any vectors $v(t)$. Indeed, any choice of $v(t)$ provides a fixed size sampling with inclusion probabilities $\pi_k = n/N$. A random sort applied before any equal probability sampling produces a simple random sampling (Sunter, 1977).

The third way consists of using a random vector $v = (v_k)$, where the v_k are N independent identically distributed variables. Note that for such v_k it is obvious that a preliminary sorting of the data will not change the sampling design, which is thus a simple random sampling design. In this case, the preliminary sorting is thus not necessary.

An interesting problem occurs when the design has equal inclusion probabilities $\pi_k = \pi$ $(k \in U)$ such that $N\pi$ is not an integer. If there is only one constraint, implying a fixed sample size, that is $x_k = 1$ $(k \in U)$, then the balancing equation can only be approximately satisfied. Nevertheless the flight phase of the cube method works until $N - p = N - 1$ elements of $\pi^* = \pi(N-1)$ are integers. The landing phase consists of deciding randomly whether the last unit is drawn or not. The sample size is therefore equal to one of the two nearest integers to $N\pi$. The cube method therefore automatically solves the rounding problem for stratum sample sizes so as to ensure that the given inclusion probabilities are exactly satisfied.

7·3. *Stratification*

Stratification can be achieved by taking $x_{kh} = \delta_{kh} n_h / N_h$ $(h = 1, \ldots, H)$, where N_h is the size of stratum U_h, n_h is the sample stratum size, and

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

In the first step, the projection of $v(t)$ by (5) gives $u = (u_k)$, where

$$u_k(1) = v_k(1) - \frac{1}{N_h} \sum_{\ell \in U_h} v_\ell(1) \quad (k \in U_h).$$

The three strategies described in § 7·2 for simple random sampling allow us to obtain a stratified random sample with simple random sampling within the strata.

An interesting property of the cube method is that the stratification can be generalised to overlapping strata, which can be called 'quota random design' or 'cross-stratification' (Deville, 1991). Suppose that two stratification variables are available, such as 'activity sector' and 'region' in a business survey. The strata defined by the first variable are denoted by $U_{h.}$ $(h = 1, \ldots, H)$ and the strata defined by the second variable are denoted by $U_{.i}$ $(i = 1, \ldots, K)$. Next define $p = H + K$ auxiliary variables,

$$x_{kj} = \pi_k \times \begin{cases} I(k \in U_{j.}) & (j = 1, \ldots, H), \\ I(k \in U_{.(j-H)}) & (j = H+1, \ldots, H+K), \end{cases}$$

where $I(.)$ is an indicator variable that takes the value 1 if the condition is satisfied and 0 otherwise. The sample can now be selected directly by means of the cube method. Generalisation to multiple quota random design follows immediately. It can be shown (Deville & Tillé, 2000) that the quota random sampling can be satisfied exactly.

Another interesting case of overlapping strata is triangular stratification. Let U_1, U_2 and U_3 be three subsets of U such that $\cup_{i=1}^3 U_i = U$, $\cap_{i=1}^3 U_i = \varnothing$, $U_i \cap U_j \neq \varnothing$, for $i \neq j$. Suppose that the inclusion probabilities are such that $\sum_{U_i} \pi_i$ is an integer. The auxiliary variables are defined by $x_{ki} = I(k \in U_i)$ $(i = 1, 2, 3)$. In this case it is possible to show that, although these variables only take the value 0 or 1, the balancing equations cannot be satisfied exactly. They can only be sometimes satisfied; see Fig. 2(c).

7·4. *Unequal probability sampling with fixed sample size*

Sampling with unequal inclusion probabilities can be carried out by means of the cube method. Suppose that the objective is to select a sample of fixed size n with inclusion probabilities π_k ($k \in U$) such that $\sum_{k \in U} \pi_k = n$. In this case, the only auxiliary variable is $x_k = \pi_k$. In order to satisfy this constraint, expression (5) implies that

$$\sum_{k \in U} u_k(t) = 0. \tag{13}$$

Each choice, random or not, of vectors $u(t)$ that satisfy (13) produces another method for sampling with unequal probability. Nearly all existing methods, except the rejective ones and the variations of systematic sampling, can easily be implemented by means of the cube method. In the case of sampling with unequal probabilities, the cube method is identical to the splitting method described in Deville & Tillé (1998). An interesting procedure emerges when projecting vector $u(t)$ used for Poisson sampling on the fixed size constraint.

Example 9. The first step can be implemented by projecting the vector

$$v(1) = (1 \;\; 0 \;\; \ldots \;\; 0)'$$

on to the fixed size constraint by means of (5) with $w_k(t) = \pi_k(t)$ ($k \in U_t$). The first step is thus as follows:

$$u(1) = \pi_1 \left\{ v(1) - \frac{\pi}{n} \right\} = \pi_1 \left(\frac{n - \pi_1}{n} \;\; \frac{-\pi_2}{n} \;\; \ldots \;\; \frac{-\pi_N}{n} \right)'.$$

Two cases must be distinguished

Case 1. If $n\pi_k/(n - \pi_1) \leqslant 1$, for all $k \neq 1$, then $\lambda_1(1) = (1 - \pi_1)n/\{\pi_1(n - \pi_1)\}$, $\lambda_2(1) = n/(n - \pi_1)$, and we then select

$$\pi(1) = \begin{cases} \pi^a = (\pi_1^a, \ldots, \pi_k^a, \ldots, \pi_N^a)', & \text{with probability } q = \pi_1, \\ \pi^b = (\pi_1^b, \ldots, \pi_k^b, \ldots, \pi_N^b)', & \text{with probability } 1 - q, \end{cases}$$

where

$$\pi_k^a = \begin{cases} 1 & (k = 1), \\ \pi_k(n - 1)/(n - \pi_1) & (k \neq 1), \end{cases}$$

$$\pi_k^b = \begin{cases} 0 & (k = 1), \\ \pi_k n/(n - \pi_1) & (k \neq 1). \end{cases}$$

Case 2. If there is at least one π_k such that $n\pi_k/(n - \pi_1) > 1$, then we write $\pi_m = \max_{k \neq 1} \pi_k$. We find that $\lambda_1(1) = (1 - \pi_1)n/\{\pi_1(n - \pi_1)\}$, $\lambda_2(1) = (1 - \pi_m)n/(\pi_1 \pi_m)$, and we select

$$\pi(1) = \begin{cases} \pi^a = (\pi_1^a, \ldots, \pi_k^a, \ldots, \pi_N^a)', & \text{with probability } q = \lambda_2(1)/\{\lambda_1(1) + \lambda_2(1)\}, \\ \pi^b = (\pi_1^b, \ldots, \pi_k^b, \ldots, \pi_N^b)', & \text{with probability } 1 - q, \end{cases}$$

where

$$\pi_k^a = \begin{cases} 1 & (k = 1), \\ \pi_k(n - 1)/(n - \pi_1) & (k \neq 1), \end{cases}$$

$$\pi_k^b = \begin{cases} \pi_1 - (1 - \pi_m)(n - \pi_1)/\pi_m & (k = 1), \\ \pi_k/\pi_m & (k \neq 1). \end{cases}$$

In fact, this method generalises and corrects Sunter's procedure (Sunter, 1977, 1986). The generalisation comes from the fact that Sunter's method does not handle the case where $n\pi_k/(n-\pi_1) \leqslant 1$, for all $k \neq 1$, at each step of the algorithm. In order to deal with this case, Sunter proposes to sort the units in decreasing order and to equalise the inclusion probabilities of the last units of the file in order that $\pi_k n/(n-\pi_1) \leqslant 1$ $(k \neq 1)$ at each step. By considering both cases, we generalise Sunter's method and can apply it to any vector of inclusion probabilities whose units are sorted in an aribitrary order.

The techniques of unequal probability sampling can always be improved. Indeed, in all the available methods for sampling with unequal probabilities with a fixed sample size, the design is only balanced on a single variable. However, two auxiliary variables are always available, namely $x_{k1} = \pi_k$ $(k \in U)$ and $x_{k2} = 1$ $(k \in U)$. The first variable implies a fixed sample size, and the second variable implies that $\hat{N} = \sum_{k \in U} S_k/\pi_k = N$. In all the available methods for selection with arbitrary inclusion probabilities, the sample is balanced on x_{k1} but not on x_{k2}. The balanced cube method allows us to satisfy both constraints approximately and to benefit at the same time from the Horvitz–Thompson and Hájek ratio estimators.

8. Balanced sampling and calibration

The comparison of calibration and balanced sampling may seem unnatural because the former is an estimation technique whereas the latter is a sampling technique. Balanced sampling and calibration can, however, be used together. Balanced sampling tends to achieve the same result as calibration at the sampling stage of the survey. Generally, the use of auxiliary variables greatly improves the precision of estimators of totals when they are highly correlated with the variables of interest. However, balanced sampling requires more auxiliary information: the values of the variables have to be known for all the population units, whereas, for calibration, only population totals must be known. In calibration, the weights of the units are modified in such way that the estimators are exactly equal to the population totals of the auxiliary variables. However, very often the weights can become very unstable with extreme or negative weights, which is very problematic particularly with small sample sizes as in small domain estimation.

Balanced sampling does not produce perfect calibration, but the deviations are negligible when the sample size is large. Balanced sampling has several advantages. If the design is exactly balanced, the Horvitz–Thompson weights are not random, and the estimator is unbiased. Furthermore, if the design is sometimes or approximately unbiased, still the calibration weights are less random, because the calibration only adjusts the rounding problem. Thus, even with calibration, balanced sampling protects against instability of the weights, and is therefore more robust.

An advantage of calibration is the opportunity to change the auxiliary variables for each study variable, and to decide about the auxiliary variables or transformations of them after the sample has been selected and the form of their relationship with the target response variable is identified. However, the use of different auxiliary variables for different response variables may produce different weights for each estimator, which is often problematic in official statistics. Calibration is also a very good way of dealing with nonresponse after the selection of the sample; in this scenario the weights must be modified in any case. Balanced sampling is thus particularly important in cases where the weights should not be changed and nonresponse does not occur. Examples include the sampling

of primary sampling units for a self-weighting two-stage or multi-stage sampling design, the sampling of clusters, as in the new French census described in § 1, sampling for quality control and sampling in a census.

In practice, the application of calibration and balanced sampling are quite different. Calibration is a weighting technique that deals with ultimate sampling units and allows treatment of nonresponse. Balanced sampling is well suited to select clusters, like the selection of primary units in a two-stage sampling scheme. The best way is to use these methods together at different stages of the survey, and we argue for the use of calibration after the selection of a balanced sample. This argument is supported by the simulation results in § 9 that show that the most accurate strategy is the use of calibration after selecting a balanced sample. Thus, if the auxiliary information is available at the unit level, it is always advantageous to select a balanced sample. Moreover, when the sample is balanced, calibration weights tend to be variable. It therefore becomes possible to use more auxiliary variables in the calibration process.

In a further paper (Deville & Tillé, 2004) we propose an approximation for the variance under balanced sampling. This variance is almost the same as for the optimal regression estimator (Montanari, 1987). The variance of the calibration estimator is generally approximated by using the linearisation technique. The variance is therefore expected to be underestimated. In the case of post-stratification at least, a second term can be computed, and it can be shown that the variance of the post-stratified estimator is larger than in stratification (Särndal et al., 1992, p. 267, expression 7.6.6).

9. Simulation study

A set of simulations has been carried out in order to evaluate and compare the following four strategies.

Strategy 1. Nonbalanced sampling with the Horvitz–Thompson estimator.
Strategy 2. Balanced sampling with the Horvitz–Thompson estimator.
Strategy 3. Nonbalanced sampling with a calibration estimator.
Strategy 4. Balanced sampling with a calibration estimator.

We have used the familiar MU284 population of Särndal et al. (1992, pp. 252–9). The four biggest municipalities have been removed from the population, because, for these municipalities, $nz_k/Z > 1$, where z_k is the size of the municipality, and Z is the population total of the z_k. Next the municipalities were regrouped in 50 clusters. We have thus used a modified version of the 'Clustered MU284 population' of Särndal et al. (1992, pp. 660–1). For both the balanced and nonbalanced designs, samples of size 20 were selected with inclusion probabilities proportional to the variable P75, the population in 1975, and fixed sample sizes. What we call the 'nonbalanced design' is actually an unequal probability design balanced on only one variable, the inclusion probabilities, i.e. variable P75. Thus, the 'nonbalanced samples' were also selected by means of the cube method. The balanced samples were selected with the balancing variables P75, RMT85, SOC82 and ME84; see Särndal et al. (1992, pp. 652–61) for a description of these variables. Since P75 is a balancing variable, the balanced samples also have a fixed sample size. Two estimators are computed for each sample, the Horvitz–Thompson estimator and the calibration estimator using the same auxiliary variables, P75, RMT85, SOC82 and ME84. The calibration estimator is defined as

$$\hat{Y}_R = \hat{Y} + (X - \hat{X})'b,$$

where

$$b = \left(\sum_{k \in U} s_k \frac{x_k x_k'}{\pi_k} \right)^{-1} \sum_{k \in U} s_k \frac{x_k y_k}{\pi_k}$$

is the 'standard' probability weighted estimator. The empirical 'true' mean squared errors were calculated by sampling independently 1000 times from the population under each of the two sampling schemes.

The results of the simulation study are presented in Table 1. The mean squared errors are relative to those obtained under Strategy 1. If the cube method provides an exactly balanced sample, the estimator and the calibration estimator are equal. However, when several continuous variables are used, an exact balanced sample can rarely be selected. Table 1 shows that variable P75 is exactly balanced while variables RMT85, SOC82 and ME3 are approximately balanced. If the cube method provides an approximately balanced sample, then the regression estimator adjusts the weights to obtain the exact calibration, and thus in this case the calibration estimator is different from the Horvitz–Thompson estimator. In this example, four balancing variables are used for a sample size equal to 20. Thus the rounding problem can be substantial. It is not surprising that, except for the variable CS82, the best strategy is the use of balanced sampling with the calibration estimator. For CS82, the best strategy is to use balanced sampling with the Horvitz–Thompson estimator. The simulations suggest also that balanced sampling is always more accurate for both estimators. Thus, if the necessary auxiliary information is available, it is always more accurate to select a balanced sample regardless of which estimation procedure is used.

The use of balanced sampling has another important advantage. When the calibration estimator was used with a nonbalanced sample, in 32% of the simulations there was at least one negative weight, while, when the calibration estimator was used with a balanced sample, in only one case did a negative weight appear. Balanced sampling protects against extreme or negative weights, which, as mentioned before, can be very problematic, particularly with small samples.

Table 1: *Simulation results. Mean squared errors relative to the values for nonbalanced sampling with the Horvitz–Thompson estimator. For full description of the variables see Särndal et al.* (1992, pp. 652–61)

	Horvitz–Thompson		Regression	
Variable	Nonbalanced	Balanced	Nonbalanced	Balanced
P75	0	–	–	–
RMT85	1	0·12	0	0
SOC82	1	0·14	0	0
ME84	1	0·17	0	0
R85	1	0·90	0·82	0·76
P85	1	0·91	1·02	0·87
CS82	1	0·80	0·92	0·82
S82	1	0·21	0·11	0·11
REV84	1	0·15	0·21	0·08
SIZE	1	0·26	0·15	0·14
S82-CS82-SS82	1	0·34	0·28	0·27
CS82-SS82	1	0·29	0·14	0·13

10. IMPLEMENTATION OF THE METHOD AND FURTHER DEVELOPMENT

The cube method generalises all methods that use auxiliary information for the sampling design including stratification, quota random design and unequal probability sampling. Moreover, several of these methods can be improved. The cube method can use overlapping strata, it corrects Sunter's procedure and it permits sample selection with unequal inclusion probabilities, while balanced on the population size. Moreover, the cube algorithm facilitates the joint use of these methods, such as quota random sampling with unequal probabilities.

Nevertheless, the importance of the cube method is not its use in particular cases but in its generality. For instance, municipalities can be selected with unequal inclusion probabilities that are proportional to the number of inhabitants. The sample can be balanced on qualitative variables such as 'regions' and a classification into 'urban/large urban/rural', and continuous variables such as 'age' and 'income'.

A quick examination of the method shows that the number of computational operations increases no faster than the square of the population size. Indeed, the method can be applied directly to several thousands of population units. The cube method was first implemented in Matlab in order to carry out a simulation study for testing variance estimation procedures. Another implementation, written in SAS-IML, allows the selection of balanced samples in populations of up to 50 000 units with 30 auxiliary variables. For larger populations, the cube algorithm can be applied in strata or other subpopulations.

A simple way of reducing the number of operations is to apply the first step of the flight phase to the first M units of the population, where $p < M < N$. Next, the second step is applied to the first M units that have non-integer elements of $\pi(1)$, and so on. This simple modification generalises the moving stratification algorithm of Tillé (1996). For very large files and when the units are selected with equal probabilities, M clusters can be created randomly. Next, the cube method is applied to the clusters using the cluster totals as auxiliary variables. When the flight phase is completed, the q clusters with non-integer inclusion probabilities are split in order to recreate M subclusters. The flight phase is then applied to these subclusters again, and so on until the clusters are split into units of interest. With some adjustments, the cube method can thus be applied to any sampling frame, even with millions of units and a large number of auxiliary variables.

The computation of the joint inclusion probabilities seems not to be feasible for the general case. In some particular cases, certain joint inclusion probabilities can be equal to zero, and a design-based estimator of the variance does not exist. A referee pointed out that, for skewed populations, the joint inclusion probabilities can be very unstable and change drastically by a change in the auxiliary variable. Consider the following example.

Example 10. Let $N = 100$, $n = 25$, $p = 2$, $x_{k1} = 1$, for all $k \in U$, and $x_{12} = 1{\cdot}000\,001$, $x_{22} = 0{\cdot}999\,999$, $x_{32} = 1{\cdot}000\,002$, $x_{42} = 0{\cdot}999\,998$, $x_{52} = 1{\cdot}000\,003$, $x_{62} = 0{\cdot}999\,997$, $x_{72} = 1{\cdot}000\,004$, $x_{82} = 0{\cdot}999\,996$, and $x_{k2} = 0$, for $k = 9, \ldots, 100$. In this case, a balanced sample either includes units 1 and 2 together or not, and thus $\pi_{12} = 0{\cdot}25$. Several joint inclusion probabilities are equal to 0. If x_{12} changes from $1{\cdot}000\,001$ to $1{\cdot}000\,002$ and x_{32} changes from $1{\cdot}000\,002$ to $1{\cdot}000\,001$, then π_{12} changes from 0·25 to 0. A small change of the auxiliary variable leads to a change in π_{12} from 0·25 to 0.

Nevertheless, even when a design-based estimator of the variance does not exist, it is still possible to propose an accurate estimator of the variance. We propose in a further paper (Deville & Tillé, 2004) to approximate the variance without using the joint inclusion

probabilities. A variance approximation is proposed for balanced sampling based on regression residuals, which is validated by a theoretical development and a large set of simulations. This variance estimator is similar to the variance estimator of a calibration estimator. Moreover we have calculated the matrix of joint inclusion probabilities by simulation for several examples and, except in very special cases, the joint inclusion probabilities are strictly positive.

Acknowledgement

The authors are grateful to Jean Dumais and David Leblanc for interesting comments on a previous version of this paper and to Carl Särndal for his continuous encouragement. The authors also thank Abdelkader Bousabaa, Joseph Lieber and Rémi Sirolli, who wrote the SAS-IML program. Jean Dumais also made the first real application of the cube method by testing the program for the selection of balanced samples of municipalities for the INSEE continuous census. The second application was carried out by the much lamented Benoît Merlat, who tested the cube method in order to select the cantons of the INSEE master sample. The authors also thank the associate editor and a reviewer for their constructive comments, which helped to improve the paper significantly, particularly §§ 8 and 9.

Appendix

Choice of the cost function

It can be argued that the choice of $C(.)$ in the landing phase is an arbitrary decision that depends on the priorities of the survey manager. As we have seen in expression (7), the cost is defined by a trace matrix M. A simple cost could be defined by the sum of squares

$$C_1(s) = \sum_j \frac{\{\hat{X}_j(s) - X_j\}^2}{X_j^2},$$

where $\hat{X}_j(s)$ is the value taken by $\hat{X}_j$ on sample s. The function $C_1(.)$ is an M-trace where M is a diagonal matrix with the jth diagonal element equal to $1/X_j^2$.

Instead we might take $M = (m_{k\ell}) = (AA')^{-1}$, and define

$$C_2(s) = (s - \pi^*)'A'(AA')^{-1}A(s - \pi^*).$$

The choice of $C_2(.)$ has a natural interpretation as a distance in $\mathbb{R}^N$, as shown by the following result.

Proposition A1. *The square of the distance between a sample s and its Euclidean projection on to the constraint hyperplane is given by*

$$C_2(s) = (s - \pi^*)'A'(AA')^{-1}A(s - \pi^*). \tag{A1}$$

Proof. The projection of a sample s on to the constraint hyperplane is

$$s - A'(AA')^{-1}A(s - \pi).$$

The Euclidean distance between s and its projection is thus

$$(s - \pi)'A'(AA')^{-1}A(s - \pi) = (s - \pi^* + \pi^* - \pi)'A'(AA')^{-1}A(s - \pi^* + \pi^* - \pi)$$

and, since $A(\pi - \pi^*) = 0$, (A1) follows directly. □

REFERENCES

ARDILLY, P. (1991). Echantillonnage représentatif optimum à probabilités inégales. *Ann. Écon. Statist.* **23**, 91–113.

DEVILLE, J.-C. (1991). A theory of quota surveys. *Survey Methodol.* **17**, 163–81.

DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop Auxiliary Information in Surveys*, pp. 21–40. Örebro, Sweden: Statistics Sweden.

DEVILLE, J.-C. & TILLÉ, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89–101.

DEVILLE, J.-C. & TILLÉ, Y. (2000). Selection of several unequal probability samples from the same population. *J. Statist. Plan. Infer.* **86**, 215–27.

DEVILLE, J.-C. & TILLÉ, Y. (2004). Variance approximation under balanced sampling. *J. Statist. Plan. Infer.* To appear.

DEVILLE, J.-C., GROSBRAS, J.-M. & ROTH, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT, Proceeding in Computational Statistics*, Ed. R. Payne and P. Green, pp. 255–66. Heidelberg: Physica Verlag.

DOUBROVINE, B., NOVIKOV, S. & FOMENKO, A. (1979). *Géométrie Contemporaine, Méthodes et Applications.* Moscow: MIR.

DUMAIS, J. & 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*, vol. **100**, Ed. M. Christine, pp. 37–76. Paris: INSEE.

FAN, C., MULLER, M. & REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *J. Am. Statist. Assoc.* **57**, 387–402.

HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* **35**, 1491–523.

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

HANSEN, M. & HURWITZ, W. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* **14**, 333–62.

HEDAYAT, A. & MAJUMDAR, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *J. Statist. Plan. Infer.* **44**, 237–47.

ISAKI, C. & FULLER, W. (1982). Survey design under a regression population model. *J. Am. Statist. Assoc.* **77**, 89–96.

KIAER, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bull. Inst. Int. Statist.* **9**, livre 2, 176–83.

LUENBERGER, D. (1973). *An Introduction to Linear and Non-linear Programming.* New York: Addison-Wesley.

MADOW, W. (1949). On the theory of systematic sampling, II. *Ann. Math. Statist.* **20**, 333–54.

MONTANARI, G. (1987). Post sampling efficient QR-prediction in large sample surveys. *Int. Statist. Rev.* **55**, 191–202.

NEYMAN, J. (1934). On the two different aspects of representative method: the method of stratified sampling and the method of purposive selection. *J. R. Statist. Soc.* **97**, 558–606.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *J. Am. Statist. Assoc.* **44**, 101–16.

ROSÉN, B. (1972). Asymptotic theory for successive sampling I. *Ann. Math. Statist.* **43**, 373–97.

ROYALL, R. & HERSON, J. (1973). Robust estimation in finite populations I. *J. Am. Statist. Assoc.* **68**, 880–9.

SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. (1992). *Model Assisted Survey Sampling.* New York: Springer Verlag.

SUNTER, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Appl. Statist.* **26**, 261–8.

SUNTER, A. (1986). Solutions to the problem of unequal probability sampling without replacement. *Int. Statist. Rev.* **54**, 33–50.

THIONET, P. (1953). *La Théorie des Sondages.* Paris: INSEE, Imprimerie Nationale.

TILLÉ, Y. (1986). A moving stratification algorithm. *Survey Methodol.* **22**, 85–94.

TSCHUPROW, A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation. *Metron* **3**, 461–93, 646–80.

VALLIANT, R., DORFMAN, A. & ROYALL, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, New York: Wiley.

WYNN, H. (1977). Convex sets of finite population plans. *Ann. Statist.* **5**, 414–8.

YATES, F. (1946). A review of recent statistical developments in sampling and sampling surveys (with Discussion). *J. R. Statist. Soc.* A **109**, 12–43.

YATES, F. (1949). *Sampling Methods for Censuses and Surveys.* London: Griffin.

[*Received September* 2002. *Revised March* 2004]