Survey Methodology, December 2008
Vol. 34, No. 2, pp. 223-234
Statistics Canada, Catalogue No. 12-001-X

A balanced sampling approach for multi-way stratification designs for small area estimation

Piero Demetrio Falorsi and Paolo Righi [1]

Abstract

The present work illustrates a sampling strategy useful for obtaining planned sample size for domains belonging to different partitions of the population and in order to guarantee the sampling errors of domain estimates be lower than given thresholds. The sampling strategy that covers the multivariate multi-domain case is useful when the overall sample size is bounded and consequently the standard solution of using a stratified sample with the strata given by cross-classification of variables defining the different partitions is not feasible since the number of strata is larger than the overall sample size. The proposed sampling strategy is based on the use of balanced sampling selection technique and on a GREG-type estimation. The main advantages of the solution is the computational feasibility which allows one to easily implement an overall small area strategy considering jointly the sampling design and the estimator and improving the efficiency of the direct domain estimators. An empirical simulation on real population data and different domain estimators shows the empirical properties of the examined sample strategy.

Key Words: Planning sampling size of small domains; Controlled selection; Balanced sampling.

1. Introduction

The small area problem is usually considered to be treated via estimation. However, if the domain indicator variables are available for each unit in the population there are opportunities to be exploited at the survey design stage. This condition is usually met in the business survey context where the domain indicator variables are available in the business register. As noted by Singh, Gambino and Mantel (1994), there is a need to develop an *overall strategy* that deals with small area problems, involving both planning sample design and estimation aspects. In this framework, it is crucial to control the sample size for each domain of interest, so that the domain is treated as a planned domain, at design stage, for which it is possible to produce direct estimates with a prefixed level of precision. In general, with a design-based approach to the inference, the presence of sample units in each domain allows one to compute domain estimates although not always reliably. Furthermore, in the model-based or model-assisted approach, the presence of sample units in each estimation domain allows one to use models with specific small area effects, giving more accurate estimates of the parameters of interest at small area level (Lehtonen, Särndal and Veijanen 2003). Marker (1999, 2001) deals with the problems of sampling design issues in small area context suggesting sample strategies, based on stratification and over-sampling, increasing the number of small areas for which accurate direct estimation is possible. These strategies are feasible in case of nested domains, but they may be unfeasible when the aim of the survey is to produce estimates for two or more partitions of the population. A standard solution to obtain planned sample sizes for the domains of two or more partitions is to use a stratified sample in which strata are identified by cross-classification of variables defining the different partitions. In the following, this design will be denoted as *cross-classification design*. In many practical situations, however the cross-classification design is unsuitable since it needs the selection of at least a number of sampling units as large as the product of the number of categories of the stratification variables. Cochran well illustrates (1977, page 124) this problem giving a clear example in which the cross-classification design is unfeasible.

The above background is typical of the business survey context. The European *Council Regulation* on Structural Business Statistics establishes that the parameters of interest refer to estimation domains defined by three different partition subsets of the population of enterprises. For instance, as we may note by table 1.1, in Italy the total number of estimation domains is 1,821; while the number of non-empty strata of the cross-classification design is larger than 37,000.

In order to overcome some problems of cross-classification designs, an easy strategy is to drop one or more stratifying variables or to group some of the categories. Nevertheless, some planned domains become unplanned and some of them can have small or null sample size.

Many methods have been proposed in the literature to keep under control the sample size in all the categories of the stratifying variables without using cross-classification design. These methods are generally referred to as *multi-way stratification techniques,* and have been developed under two

[1] Piero Demetrio Falorsi, Italian National Statistical Institute. E-mail: falorsi@istat.it; Paolo Righi, Italian National Statistical Institute. E-mail: parighi@istat.it.

main approaches: (i) Latin Squares or Latin Lattices schemes (Jessen 1970); (ii) controlled rounding problems via linear programming (Lu and Sitter 2002). Both approaches have some drawbacks which have limited the use of multi-way stratification techniques as a standard solution for planning the survey sampling designs in real survey contexts. Indeed, as described in Falorsi, Orsini and Righi (2006), it is not possible to implement the Latin Lattices schemes in many real survey contexts; as for example if there are no population units in one or more cross-classification strata. The main weakness of the linear programming approach is the computational complexity. The sampling strategy considered in this paper does not suffer from the disadvantages of the above mentioned methods and allows one the control of the sample sizes for domains of interest, which are defined by different partitions of the reference population. Furthermore it guarantees that the sampling errors of domain estimates are lower than the given thresholds.

The proposed sampling strategy is based on the use of both a *balanced sampling* selection technique (Deville and Tillé 2004) and a *GREG-type* estimation (Lehtonen *et al.* 2003). As shown in the study on empirical data herein illustrated and in Falorsi and Righi (2008), the main advantages of this solution is the computational feasibility and the efficiency, that is the sampling errors for multi-domain-multivariate case are reasonably close to those defined by the optimal univariate solutions. This allows one to fairly implement an overall small-area strategy considering jointly the sampling design and the estimator and improving the efficiency of the direct domain estimators.

In some survey context, the proposed sampling strategy might define a too large overall sample size for assuring the prefixed bound of the direct domain estimates sampling errors. This may happen due to a too large number of domains of a given population partition. If the overall sample size is bounded by budget constraints, then the proposed sampling strategy with direct estimators may be not feasible. Therefore, it could be necessary to adopt an indirect small-area estimator in order to control the mean square errors of partition domain estimates. However, the proposed approach may be easily extended to a strategy using the direct estimator and the indirect small area estimators for the partitions requiring a too large overall sample size for bounding the sampling errors.

The paper is organised as follows. Section 2 states the problem, introduces the essential notation and describes the overall sampling strategy. Section 3 shows the algorithms for finding the inclusion probabilities and the corresponding planned domain sample sizes. Sections 4 and 5 illustrate two extensions of the sampling strategy. In section 4 the case in which the variance criterion is represented by the anticipated variance is studied. An extension to the case of a simple small area indirect estimator is presented in section 5. The main results of an empirical study on a real population of Italian enterprises are shown in section 6. Some brief conclusions are finally underlined in section 7.

2. The sampling strategy

2.1 Parameters of interest

In order to define formally the problem, let us denote with U a population of N elements and with b a specific partition of $U(b = 1, ..., B)$ in which b^{th} partition defines M_b different non overlapping domains, U_{bd} $(d = 1, ..., M_b)$, of size N_{bd} being $\sum_{d=1}^{M_b} N_{bd} = N$ and, finally let $\sum_{b=1}^{B} M_b = Q$ be the overall number of domains.

Let $y_{r,k}$ and ${}_{bd}\delta_k$ denote respectively the value of the r^{th} $(r = 1, ..., R)$ variable of interest in the k^{th} population unit and the domain membership indicator, being ${}_{bd}\delta_k = 1$ if $k \in U_{bd}$ and ${}_{bd}\delta_k = 0$, otherwise. Let us suppose that the ${}_{bd}\delta_k$ values are known for each unit in the population.

The parameters of interest are the $M = Q \times R$ domains totals

$$ {}_{bd}t_r = \sum_{k\in U} y_{r,k}\,{}_{bd}\delta_k = \sum_{k\in U_{bd}} y_{r,k} $$

$$ (r = 1, ..., R;\ b = 1, ..., B;\ d = 1, ..., M_b). \quad (2.1.1) $$

The expression (2.1.1) defines a *multivariate-multi-domain* problem since there are R variables of interest (multivariate aspect) and $Q > 1$ domains (multi-domain aspect).

Table 1.1
Number of domains of the Italian Structural Business Statistics Survey by partition

Partitions	Number of domains
Economic activity class (4-digits of the NACE rev.1 classification)	465
Economic activity group (3-digits of the NACE rev.1 classification) by Size class[1]	395
Economic activity division (2-digits of the NACE rev.1 classification) by Region[1]	961
Total number of estimation domains	1,821

[1] Size classes are defined in terms of number of persons employed.

[2] Regions are 21 including autonomous provinces.

2.2 A concise description of the sampling strategy

Let us suppose that, in order to estimate the ${}_{bd}t_r$ parameters, a sample s of fixed size n is selected from population U, with inclusion probabilities $\pi_k (k \in U)$. Let $s_{bd} = s \cap U_{bd}$ be the sample of n_{bd} units belonging to the U_{bd} domain (with $\sum_{d=1}^{M_b} n_{bd} = n$), being

$$n_{bd} = \sum_{k \in U_{bd}} \lambda_k = \sum_{k \in U_{bd}} \pi_k, \qquad (2.2.1)$$

with $\lambda_k = 1$ if $k \in s$ and $\lambda_k = 0$ otherwise.

The sample is selected by a *multi-way stratification technique* developed under the *balanced sampling* framework guaranteeing that the selected sample respects the following *balancing equations*

$$\hat{t}_{\mathbf{z},\, ht} = t_{\mathbf{z}} \qquad (2.2.2)$$

where $\hat{t}_{\mathbf{z},\, ht} = \sum_{k \in U} \mathbf{z}_k \lambda_k a_k$ denote the Horvitz-Thompson estimates of $t_{\mathbf{z}} = \sum_{k \in U} \mathbf{z}_k$, being $\mathbf{z}_k$ a value vector of auxiliary variables known for each population unit at the design stage and $a_k = 1/\pi_k$. A suitable specification of the $\mathbf{z}_k$ vectors can assure that the realized sample sizes, n_{bd}, are equal to fixed quantities known in advance, as described in section 2.3.

The estimates of ${}_{bd}t_r$, denoted with ${}_{bd}\hat{t}_{r,\, \mathrm{greg}}$, are obtained with the *modified GREG* estimator (Rao 2003, page 20), given by:

$${}_{bd}\hat{t}_{r,\, \mathrm{greg}} = \sum_{k \in s} {}_{bd}w_k \; y_{r,\, k} \qquad (2.2.3)$$

where

$${}_{bd}w_k = a_k \, {}_{bd}\delta_k$$
$$+ ({}_{bd}t_{\mathbf{x}} - {}_{bd}\hat{t}_{\mathbf{x},\, ht})' \left[\sum\nolimits_{k \in s} (a_k \mathbf{x}_k \mathbf{x}'_k / c_k)\right]^{-1} a_k \, \mathbf{x}_k / c_k$$

denote the sampling weights, $\mathbf{x}_k$ indicates a value vector of the auxiliary variables, c_k is a known constant, being ${}_{bd}t_{\mathbf{x}} = \sum_{k \in U_{bd}} \mathbf{x}_k$ and ${}_{bd}\hat{t}_{\mathbf{x},\, ht} = \sum_{k \in s_{bd}} \mathbf{x}_k a_k$. The estimator (2.2.3), may be derived under the following *working* super population model

$$y_{r,\, k} = \mathbf{x}'_k \boldsymbol{\beta}_r + \varepsilon_{r,\, k} \qquad (2.2.4)$$

where $\boldsymbol{\beta}_r$ is an unknown vector of fixed regression parameters and $\varepsilon_{r,\, k}$ is the random residual. The model expectation, E_m, and model variances, V_m, are respectively given by $E_m({}_r\varepsilon_k) = 0$; $V_m(\varepsilon_{r,\, k}) = c_k \sigma_r^2$; $E_m(\varepsilon_{r,\, k}, \varepsilon_{r,\, i}) = 0$ if $k \neq i$.

The approximated sampling variance of the modified GREG estimator under balanced sampling is:

$$V_p({}_{bd}\hat{t}_{r,\, \mathrm{greg}} \,|\, \hat{t}_{\mathbf{z},\, ht} = t_{\mathbf{z}}) = \frac{N}{N - Q} \sum_{k \in U} \left(\frac{1}{\pi_k} - 1\right) {}_{bd}\eta_{r,\, k}^2, \qquad (2.2.5)$$

being

$${}_{bd}\eta_{r,\, k} = \begin{cases} \varepsilon_{r,\, k} - \mathbf{z}'_k \, {}_{bd}\mathbf{B}_{\mathbf{z},\, \varepsilon} & \text{for } k \in U_{bd} \\ -\mathbf{z}'_k \, {}_{bd}\mathbf{B}_{\mathbf{z},\, \varepsilon} & \text{for } k \in U_{b\bar{d}} \end{cases}, \qquad (2.2.6)$$

where

$${}_{bd}\mathbf{B}_{\mathbf{z},\, \varepsilon} = \left[\sum\nolimits_{k \in U} \mathbf{z}_k \mathbf{z}'_k (1/\pi_k - 1)\right]^{-1} \sum_{k \in U} \mathbf{z}_k \, \varepsilon_{r,\, k} \, {}_{bd}\delta_k (1/\pi_k - 1),$$

being $U_{b\bar{d}}$ the subset of U complementary to U_{bd}. A proof of (2.2.5) is given in section 2.5.

The inclusion probabilities, π_k, and the domain sample sizes, n_{bd}, are determined with a procedure which attempts to minimize the overall sample size, n, guaranteeing that the sampling variances are lower than prefixed level of precision thresholds, ${}_{bd}\bar{V}_r$: $V_p({}_{bd}\hat{t}_{r,\, \mathrm{greg}} \,|\, \hat{t}_{\mathbf{z},\, ht} = t_{\mathbf{z}}) \leq {}_{bd}\bar{V}_r$ ($b = 1, \ldots, B$; $d = 1, \ldots, M_b$; $r = 1, \ldots, R$). The technical details are described in section 3.

Let us note that two different sets of covariates have been introduced in order to underline that the set of covariates available at the design stage ($\mathbf{z}$ variables) could be different from the set available at the estimation stage ($\mathbf{x}$ variables) even if in many practical situations they could be the same. As for example the covariates at estimation stage could be updated with respect to those available at the design stage. In our context (see section 2.3) the $\mathbf{z}_k$ vectors are characterized as specified by the expression (2.3.2) being defined only by the domain membership indicator variables and by the inclusion probabilities, while the $\mathbf{x}_k$ vectors could contain the values of some other variables more explicative of the phenomena of interest. For instance, in the business survey context the $\mathbf{x}$ variables could include, among others, the number of employees or the turnover.

2.3 The balanced sampling for marginal stratification

Multi-way stratification designs can be treated in the context of the *balanced sampling*.

The definition of a balanced sample depends on the assumed inferential framework. In the model based approach, a sample is defined as *balanced* on a set of auxiliary variables if there is the equality between the sample and the known population means of the auxiliary variables (Valliant, Dorfman and Royall 2000). Following the design based (or model assisted approach) considered in this paper, a sample is *balanced* when the Horvitz-Thompson estimates of the auxiliary variables totals are equal to their known population totals (Deville and Tillé 2004).

For defining the balanced sampling in the design or model assisted approach, let us introduce the general definition of sampling design as a probability distribution $p(\cdot)$ on the set S of all the subset s of the population U

such that $\sum_{s \in \mathrm{S}} p(s) = 1,$ where $p(s)$ is the probability of the sample s to be drawn. Each set s may be represented by the outcome $\boldsymbol{\lambda}' = (\lambda_1, ..., \lambda_k, ..., \lambda_N)$ of a vector of N random variables. Let $\boldsymbol{\pi}' = (\pi_1, ..., \pi_k, ..., \pi_N)$ be the vector of inclusion probabilities, where $\boldsymbol{\pi} = E_p(\boldsymbol{\lambda}) = \sum_{s \in \mathrm{S}} p(s)\boldsymbol{\lambda},$ being $E_p(\cdot)$ the expected value over repeated sampling. Let $\mathbf{z}'_k = (z_{1k}, ..., z_{hk}, ..., z_{Qk})$ be a vector of Q auxiliary variables available for each population unit. The sampling design $p(s)$ with inclusion probabilities $\boldsymbol{\pi}$ is said to be *balanced* with respect to the Q auxiliary variables if and only if it satisfies the balancing equations given by (2.2.2) for all $s \in S$ such that $p(s) > 0.$

Let us suppose that a vector of inclusion probabilities $\boldsymbol{\pi},$ consistent with the marginal sampling distributions n_{bd} $(b = 1, ..., B;\ d = 1, ..., M_b),$ is available, that is

$$n_{bd} = \sum_{k \in U_{bd}} \pi_k (b = 1, ..., B;\ d = 1, ..., M_b). \quad (2.3.1)$$

Multi-way stratification design represents a special case of balanced design where for unit k the auxiliary variable vector is given by

$$\mathbf{z}'_k = (\underbrace{\overbrace{0, ..., \pi_k, ..., 0}^{b=1}, ..., \overbrace{0, ..., \pi_k, ..., 0}^{b=B}}_{Q})$$

$$= \pi_k({}_{11}\delta_k, ..., {}_{bd}\delta_k, ..., {}_{BM_B}\delta_k). \quad (2.3.2)$$

The expression (2.3.2) defines the $\mathbf{z}_k$ as vectors of $(Q - B)$ zeros and with B entries equal to π_k in the places indicating the domains which the unit k belongs to. When defining the $\mathbf{z}_k$ vector as (2.3.2), if condition (2.3.1) holds, the selection of sample satisfying the system of *balancing equations* (2.2.2), $\sum_{k \in U}(\mathbf{z}_k \lambda_k)/\pi_k = \sum_{k \in s} \mathbf{z}_k,$ guarantees that the n_{bd} values are non random quantities. The left hand-side of the balancing equation (2.2.2) is $\sum_{k \in U}(\pi_k\ {}_{bd}\delta_k \lambda_k)/\pi_k = \sum_{k \in U_{bd}} \lambda_k = n_{bd},$ while the right hand-side is $\sum_{k \in U} \pi_k\ {}_{bd}\delta_k = \sum_{k \in U_{bd}} \pi_k = n_{bd}.$

Deville and Tillé (2004) proposed the *cube method* that allows one the selection of balanced (or approximately balanced) samples for a large set of auxiliary variables and with respect to different vectors of inclusion probabilities. In particular, Deville and Tillé (2000) show that with specification (2.3.2) of the $\mathbf{z}_k$ vectors, the balancing equations (2.2.2) can be exactly satisfied. The cube method is implemented by an enhanced algorithm for large data sets (Chauvet and Tillé 2006) available in a free software code that may be downloaded in the website *http://www.insee.fr/fr/nom_df_met/outils_stat/cube/accueil_cube.htm* .

2.4 The modified direct GREG estimator

Following Lehtonen *et al.* (2003), the estimator (2.2.3), may be expressed under the general form

$${}_{bd}\hat{t}_{r,\,\mathrm{greg}} = \sum_{k \in U_{bd}} \tilde{y}_{r,k} + \sum_{k \in s_{bd}} a_k (y_{r,k} - \tilde{y}_{r,k}) \quad (2.4.1)$$

where $\tilde{y}_{r,k}$ denotes the prediction of $y_{r,k}$ under the assumed super population model. The predictions $\{\tilde{y}_{r,k};\ k \in U\}$ differ from one model specification to another, depending on the functional form and from the choice of the auxiliary variables. The estimator (2.2.3) is derived under the working super-population model (2.2.4). The predictions $\tilde{y}_{r,k}$ are then obtained by

$$\tilde{y}_{r,k} = \mathbf{x}'_k \hat{\boldsymbol{\beta}}_r, \quad (2.4.2)$$

being

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}'_k\, a_k / c_k\right)^{-1} \sum_{k \in s} \mathbf{x}_k\, y_{r,k}\, a_k / c_k. \quad (2.4.3)$$

Let us observe that the linear model (2.2.4) allows one to define the estimator only knowing the domain totals of the auxiliary information and the $\mathbf{x}_k$ values for the sampling units. However, knowing the $\mathbf{x}_k$ values for every $k \in U,$ it is possible to build an estimators with more efficient predictions $\tilde{y}_{r,k}$ obtained by generalized linear models (Lehtonen and Veijanen 1998) or non parametric regression techniques (Montanari and Ranalli 2003).

As noted by Rao (2003, page 20) the estimator (2.2.3) is approximately design unbiased as the overall sample size increases, even if the domain sample size n_{bd} is small. Moreover, the sum of the ${}_{bd}\hat{t}_{r,\,\mathrm{greg}}$ estimates over all the domains of a partitions is benchmarked to the usual GREG estimate of the total, $\sum_{d=1}^{M_b} {}_{bd}\hat{t}_{r,\,\mathrm{greg}} = \sum_{k \in s} y_{r,k}\, a_k[1 + (\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} \mathbf{x}_k a_k)' (\sum_{k \in s} \mathbf{x}_k \mathbf{x}'_k\, a_k / c_k)^{-1} \mathbf{x}_k / c_k].$

2.5 Sampling variances

In order to derive the expression of the variance (2.2.5), consider the results given by Deville and Tillé (2005). They have proposed approximating the variance of the Horvitz-Thompson estimator $\hat{t}_{r,\,ht} = \sum_{k \in s} y_{r,k}\, a_k$ of the total $t_r = \sum_{k \in U} y_{r,k},$ by supposing that the balanced sampling can be viewed as a conditional Poisson sampling and assuming that, at least for large sample sizes, the inclusion probabilities π_k well approximate the inclusion probabilities of the Poisson design. Assuming that, through Poisson sampling, the vector $(\hat{t}_{r,\,ht},\ \hat{t}'_{\mathbf{z},\,ht})'$ has approximately a multi-normal distribution, the authors suggest a good approximation of the sampling variance given by

$$\begin{aligned} V_p(\hat{t}_{r,ht} | \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) &= V_p(\hat{t}_{r,\,ht} + (t_{\mathbf{z}} - \hat{t}_{\mathbf{z},\,ht})'\, \mathbf{B}_{\mathbf{z},y}) \\ &= V_p(\hat{t}_{r,\,ht} - \hat{t}'_{\mathbf{z},\,ht}\, \mathbf{B}_{\mathbf{z},\,y}) \\ &= V_p\left(\sum_{k \in s} a_k (y_{r,k} - \mathbf{z}'_k\, \mathbf{B}_{\mathbf{z},\,y})\right) \\ &\cong \frac{N}{N - Q} \sum_{k \in U} \left(\frac{1}{\pi_k} - 1\right)(y_{r,k} - \mathbf{z}'_k\, \mathbf{B}_{\mathbf{z},\,y})^2 \quad (2.5.1) \end{aligned}$$

where $\mathbf{B}_{\mathbf{z},\,y} = [\sum_{k \in U} \mathbf{z}_k \mathbf{z}'_k (1/\pi_k - 1)]^{-1} \sum_{k \in U} \mathbf{z}_k \, y_{r,\,k} (1/\pi_k - 1)$. The expression (2.5.1) has been validated by a set of simulations.

Let us consider, now, the linear approximation, ${}_{bd}\hat{t}^{*}_{r,\,\text{greg}}$, of the GREG estimator, the derivation of which may be obtained according to Särndal, Swensson and Wretman (1992, pages 450-451)

$$
\begin{aligned}
{}_{bd}\hat{t}_{r,\,\text{greg}} \cong {}_{bd}\hat{t}^{*}_{r,\,\text{greg}} &= \sum_{k \in U_{bd}} \mathbf{x}'_k \boldsymbol{\beta}_r + \sum_{k \in s_{bd}} a_k \, \varepsilon_{r,\,k} \\
&= \sum_{k \in U_{bd}} \mathbf{x}'_k \boldsymbol{\beta}_r + \sum_{k \in s} a_k \, \varepsilon_{r,\,k} \; {}_{bd}\delta_k. \quad (2.5.2)
\end{aligned}
$$

On the basis of expressions (2.5.1) and (2.5.2), it is possible to derive the following result

$$
\begin{aligned}
V_p({}_{bd}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) &\cong V_p({}_{bd}\hat{t}^{*}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \\
&= V_p\Big(\sum_{k \in s} a_k \; \varepsilon_{r,k} \; {}_{bd}\delta_k \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}\Big) \\
&= V_p\Big(\sum_{k \in s} a_k \, \varepsilon_{r,k} \; {}_{bd}\delta_k + (t_{\mathbf{z}} - \hat{t}_{\mathbf{z},ht})' \; {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\Big) \\
&= V_p\Big(\sum_{k \in s} a_k \; (\varepsilon_{r,k} \; {}_{bd}\delta_k - \mathbf{z}'_k \; {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\Big) \\
&= V_p\Big(\sum_{k \in s} a_k \; {}_{bd}\eta_{r,k}\Big) \\
&\cong \frac{N}{N-Q} \sum_{k \in U} \Big(\frac{1}{\pi_k} - 1\Big) {}_{bd}\eta^2_{r,k},
\end{aligned}
$$

where ${}_{bd}\eta^2_{r,\,k}$ is defined in (2.2.6).

The approximated sampling variance of ${}_{bd}\hat{t}_{r,\,\text{greg}}$ depends on the residuals of the whole set of units, because of balanced selection. Therefore, the units not belonging to U_{bd} have an influence on the sampling variance of the estimator.

Let us examine now the univariate unidomain case and assume that the survey has an unique target parameter, ${}_{bd}t_r$. Furthermore, let us suppose that the selected sample respects the *balancing equations,* $\hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}$, being fixed the overall sample size *n.*

Following the arguments proposed by Särndal *et al.* (1992; Result 12.2.1, page 452), it is trivial to prove that, in this sampling context, each unit *k* could be selected with $(Q \times R)$ different optimal inclusion probabilities, ${}_{bd}\ddot{\pi}_{r,\,k}$ $(b = 1, \ldots, B;\ d = 1, \ldots, M_b;\ r = 1, \ldots, R)$

$$
{}_{bd}\ddot{\pi}_{r,\,k} = n \,|\, {}_{bd}\eta_{r,k} \,| \Big/ \sum_{i \in U} |\, {}_{bd}\eta_{r,i} \,|,
$$

which allow one to attain the $(Q \times R)$ different lower bounds, ${}_{bd}V^{*}_{r|n}$, of the approximated variances:

$$
V_p({}_{bd}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},\,ht} = t_{\mathbf{z}}) \geq {}_{bd}V^{*}_{r|n} =
$$

$$
\frac{N}{N-Q}\left[\frac{1}{n}\Big(\sum_{k \in U} |\, {}_{bd}\eta_{r,k} \,|\Big)^2 - \sum_{k \in U} {}_{bd}\eta^2_{r,\,k}\right].
$$

Let us finally underline that in Tillé and Favre (2005) is given a criterion for obtaining a prediction ${}_{bd}\hat{\eta}_{r,\,k}$ of the ${}_{bd}\eta_{r,\,k}$ values, that may be used in repeated sampling contexts.

3. Sampling algorithms for the determination of the domain sample sizes

The inclusion probabilities π_k and the derived domain sample sizes, $n_{bd} = \sum_{k \in U_{bd}} \pi_k$, are obtained with a two steps procedure: (i) in the first step, denoted as *optimization*, the preliminary inclusion probabilities, π'_k, are determined solving a minimum constrained problem; (ii) in the second step, denoted as *calibration*, the inclusion probabilities, π_k, are obtained as a slight modification of the π'_k; the calibration problem is implemented for assuring that the domain sample sizes n_{bd} are integers.

As illustrated in the following, the π_k values may be expressed as implicit functions of the unknown residuals ${}_{bd}\eta^2_{r,\,k}$. But, in real survey context, the determination of the inclusion probabilities π_k may be done using the predictions ${}_{bd}\hat{\eta}^2_{r,\,k}$ instead of ${}_{bd}\eta^2_{r,\,k}$. This is a general problem concerning the planning the sampling designs, because the variances are generally unknown quantities that may be suitably estimated. In repeated survey contexts the effect of using the estimates ${}_{bd}\hat{\eta}^2_{r,\,k}$ as a replacement for ${}_{bd}\eta^2_{r,\,k}$ may be tested by computing the sampling variances after the data collection. The empirical results may then be used for introducing proper adjustments in planning the next survey design. However, as illustrated in the empirical analysis and in Falorsi and Righi (2008), the proposed strategy seems to be efficient and sufficiently robust with respect to small departures of ideal conditions.

The sections 3.1 and 3.2 respectively describe the two steps of the algorithm for the determination of the domain sample sizes. A *simplified allocation rule*, which seems to be worthwhile in many real survey contexts, is described in section 3.3.

3.1 First step: Optimization

The inclusion probabilities π'_k can be defined as solution of the following non linear programming problem with N unknowns, π'_k, and $(N + Q \times R)$ constraints

$$\begin{cases} \operatorname{Min}\left(\sum_{k\in U}\pi'_k\right) \\ \frac{N}{N-Q}\sum_{k\in U}\left(\frac{1}{\pi'_k}-1\right){}_{bd}\eta^2_{r,k} \le {}_{bd}\bar{V}_r \\ \qquad (b=1,\ldots,B;\ d=1,\ldots,M_b;\ r=1,\ldots,R) \\ 0<\pi'_k\le 1\ \ (k=1,\ldots,N). \end{cases} \tag{3.1.1}$$

A numerical solution to (3.1.1) may be derived considering the algorithms developed for the multivariate allocation in stratified surveys. Such algorithms allow one to find the unknown values $v_h > 0\ (h = 1, 2, \ldots)$ which represent the solution of the following non linear problem $\operatorname{Min}(\sum_h \nu_h)$ under the constraints $\sum_h A_{rh}/\nu_h \le \bar{A}_r$, where A_{rh} and $\bar{A}_r\ (r = 1, 2, \ldots)$ are known positive quantities.

Bethel (1989) invokes the Kuhn-Tucker theorem to show that there exists a solution to the above problem. He describes a simple algorithm and discusses its convergence properties. Chromy (1987) develops an algorithm, suitable for automated spreadsheets but without an explicit proof that always converges. A slight modification of the Chromy's algorithm – able to solve the problem (3.1.1) guaranteeing the inequalities $0 < \pi'_k \le 1\ \ (k = 1, \ldots, N)$ are respected – is described herein in the following. After the *Initialization*, the algorithm finds the π'_k values by iterating the two actions of *Calculus* and *Check.* As far as the convergence issue is concerned, it is worthwhile to note that the Chromy's algorithm have been mostly used for stratified sampling design, and indeed, the documentation refers to stratified samples. In the applied sampling literature, there is a lot of empirical proofs of the successful use of the algorithm in this sampling context. Let us note that the modification of the Chromy's algorithm, herein proposed, treats the sampling units as strata and the resulting allocation, being fractional, defines the inclusion probabilities. Also in this case there is no formal proof that the proposed modified algorithm converges. Nevertheless, in all the different empirical experiments developed by the authors the algorithm has always converged and no critical conditions have been encountered.

Initialization: at initial iteration $(\tau = 0)$, set ${}^{\tau}\gamma_k = 1$ $(k = 1, \ldots, N)$.
Calculus: the generic iteration $(\tau = 1, 2, \ldots)$ consists of a sequence of steps denoted with $u = (0, 1, 2, \ldots)$.

- At initial step $(u = 0)$, set ${}^{\tau,u}_{bd}\phi_r = 1$ and calculate

$$ {}^{\tau}_{bd}V_{0r} = \frac{N}{N-Q}\sum_{k\in U}\ {}_{bd}\eta^2_{r,k}\ {}^{\tau}\gamma_k .$$

- At subsequent steps $(u = 1, 2, \ldots)$, calculate the values of the following equations

$$ {}^{\tau,u}\pi_k = \left[(1-{}^{\tau}\gamma_k)+{}^{\tau}\gamma_k\frac{N}{N-Q}\sum_{b=1}^{B}\sum_{d=1}^{M_b}\sum_{r=1}^{R}{}^{\tau,u}_{bd}\phi_r\,{}_{bd}\eta^2_{r,k}\right]^{1/2}. \tag{3.1.2}$$

$$ {}^{\tau,u}_{bd}V_r = \frac{N}{N-Q}\sum_{k\in U}\frac{1}{{}^{\tau,u}\pi_k}\ {}_{bd}\eta^2_{r,k}\ {}^{\tau}\gamma_k,$$

and (3.1.3)

$$ {}^{\tau,u}_{bd}V'_r = {}^{\tau,u}_{bd}V_r + {}^{\tau}_{bd}V_{0r}.$$

- If the following two conditions:

$$ {}^{\tau,u}_{bd}V'_r \le {}_{bd}\bar{V}_r \text{ and } {}^{\tau,u}_{bd}\phi_r\,({}^{\tau,u}_{bd}V'_r - {}_{bd}\bar{V}_r) = 0, \tag{3.1.4}$$

are respected (for all $b = 1, \ldots, B; d = 1, \ldots, M_b; r = 1, \ldots, R$) then the action of Calculus stops and the inclusion probabilities ${}^{\tau}\pi_k$ are those calculated in equation (3.1.2). Otherwise, the updated quantities ${}^{\tau,u+1}_{bd}\phi_r$ are computed

$$ {}^{\tau,u+1}_{bd}\phi_r = {}^{\tau,u}_{bd}\phi_r\,[{}^{\tau,u}_{bd}V_r/({}^{\tau,u}_{bd}V'_r - {}^{\tau}_{bd}\bar{V}_r)]^2 \tag{3.1.5}$$

and the equations (3.1.2) and (3.1.3) are calculated at $u + 1$, over and over again with ${}^{\tau,u+1}_{bd}\phi_r$ replacing ${}^{\tau,u}_{bd}\phi_r$ until conditions (3.1.4) are respected.

Check: if the condition ${}^{\tau}\pi_k \le 1$ is true for all k, then the algorithm stops and the π'_k values are set equal to $\pi'_k = {}^{\tau}\pi_k$. Otherwise the ${}^{\tau}\gamma_k$ values are updated as ${}^{\tau+1}\gamma_k = 1$ if ${}^{\tau}\pi_k \le 1$ and ${}^{\tau+1}\gamma_k = 0$ if ${}^{\tau}\pi_k > 1$. The calculus is iterated at $\tau + 1$ with ${}^{\tau+1}\gamma_k$ replacing ${}^{\tau}\gamma_k$. A SAS macro that allows one to solve the problem (3.1.1) has been developed by the authors of this paper and may be released on demand.

3.2 Second step: Calibration

The quantities n_{bd} are defined, first, by rounding the results of the Q sums, $\sum_{k\in U_{bd}}\pi'_k(b=1,\ldots,B;\ d=1,\ldots,M_b)$. Sometimes a further data manipulation could be necessary in order to assure the condition $\sum_{d=1}^{M_b} n_{bd} = n$ for each b. The probabilities π_k are then obtained as solution of *calibration problem*

$$\begin{cases} \operatorname{Min}\left(\sum_{k\in U} G(\pi_k;\pi'_k)\right) \\ \sum_{k\in U}\pi_k = n,\ \sum_{k\in U_{bd}}\pi_k = n_{bd} \\ \qquad (b=1,\ldots,B;\ d=1,\ldots,M_b-1), \end{cases} \tag{3.2.1}$$

where, $G(\pi_k;\pi'_k)$ is a distance function between π_k and π'_k. Note that (3.2.1) may be solved by the well known *Iterative Proportional Fitting* algorithm (Bishop, Fienberg and Holland 1975) or the *Generalized Iterative Proportional Fitting* algorithm (GIPF; Dykstra and Wollan 1987) procedures. The logarithmic distance function $G(\pi_k;\pi'_k) = \pi_k \ln(\pi_k / \pi'_k) - (\pi_k + \pi'_k)$ avoids to define the π_k probabilities lower than 0, while GIPF prevents to obtain π_k values larger than 1.

3.3 A simplified allocation rule

In many real survey contexts in which the overall sample size n is fixed and there is not enough information to obtain good predictions ${}_{bd}\hat{\eta}^2_{r,k}$ of the ${}_{bd}\eta^2_{r,k}$ values, the following procedure may be implemented. Firstly the marginal sample sizes n_{bd} are determined by a quite simple rule

$$n_{bd} = \alpha_b\, n(N_{bd} / N) + (1 - \alpha_b) n / M_b, \qquad (3.3.1)$$

being $\alpha_b (0 \le \alpha_b \le 1)$ a fixed constant which have to be properly defined. The (3.3.1) turns out to be a compromise between the allocation proportional to population size $(\alpha_b = 1)$ and the allocation uniform for each domain of a given partition $(\alpha_b = 0)$.

The probabilities π_k are then obtained as solution of the calibration problem (3.2.1) where the marginal sample sizes are computed as above indicated and the initial probabilities π'_k are set uniformly equal to $\pi'_k = n / N$. The resulting inclusion probabilities are no more *optimal*, in the sense above described and do not guarantee that the sampling variances are lower than prefixed level of precision thresholds. However they are computed with a reasonable procedure, which may be fairly implemented and thus representing an interesting point of reference with respect to any real survey context.

4. The anticipated variance

A frequently used criterion for planning the sampling strategies is that of controlling the anticipated variance, which may be defined as:

$$\text{AV}({}_{bd}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) = E_m\, E_p({}_{bd}\hat{t}_{r,\,\text{greg}} - {}_{bd}t_r \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}})^2. \quad (4.1)$$

The following result may be derived under the assumptions of the model (2.2.4) and using the results given in section (2.5):

$$\begin{aligned}
&\text{AV}({}_{bd}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \\
&\quad \cong E_m\, V_p({}_{bd}\hat{t}^{*}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \\
&\quad = E_m \Bigg[\frac{N}{N-Q} \sum_{k \in U_{bd}} \left(\frac{1}{\pi_k} - 1\right) \varepsilon^2_{r,k} \\
&\qquad\quad \frac{N}{N-Q} \sum_{k \in U} \left(\frac{1}{\pi_k} - 1\right) (\mathbf{z}'_k\; {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 \\
&\qquad\quad - 2\, \frac{N}{N-Q} \sum_{k \in U_{bd}} \left(\frac{1}{\pi_k} - 1\right) \varepsilon_{r,k}\; \mathbf{z}'_k\; {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon} \Bigg] \\
&\quad = \frac{N}{N-Q} \sum_{k \in U} \left(\frac{1}{\pi_k} - 1\right) {}^{a}_{bd}\eta^2_{r,k} \qquad (4.2)
\end{aligned}$$

being

$${}^{a}_{bd}\eta^2_{r,k} = \begin{cases} \sigma^2_r\, c_k\, (1 - g_{kk})^2 + \sigma^2_r \displaystyle\sum_{j(\neq k) \in U_{bd}} g^2_{kj} c_j & \text{if } k \in U_{bd} \\ \sigma^2_r \displaystyle\sum_{j \in U_{bd}} g^2_{kj}\, c_k & \text{otherwise} \end{cases},$$

where: $(g_{k1}, \ldots, g_{kj}, \ldots, g_{kN}) = \mathbf{g}'_k = \mathbf{z}'_k(\mathbf{Z}'_U\, \mathbf{\Omega}^{-1}_U \mathbf{Z}_U)^{-1} \mathbf{Z}'_U\, \mathbf{\Omega}^{-1}_U$, $\mathbf{Z}_U = \text{col}\{\mathbf{z}'_k\}^N_{k=1}$, $\mathbf{\Omega}^{-1}_U = \text{diag}\{1/\pi_k - 1\}^N_{k=1}$. The expression (4.2) has been derived using the following two results

$$\begin{aligned}
&E_m(\mathbf{z}'_k\; {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 \\
&\quad = \sigma^2_r\, \mathbf{z}'_k (\mathbf{Z}'_U \mathbf{\Omega}^{-1}_U \mathbf{Z}_U)^{-1} \mathbf{Z}'_U\, \mathbf{\Omega}^{-1}_U\; {}_{bd}\mathbf{V}_r \mathbf{\Omega}^{-1}_U \mathbf{Z}_U (\mathbf{Z}'_U \mathbf{\Omega}^{-1}_U \mathbf{Z}_U)^{-1} \mathbf{z}_k \\
&\quad = \sigma^2_r\, \mathbf{g}'_k\; {}_{bd}V_r\; \mathbf{g}_k = \sigma^2_r \sum_{j \in U_{bd}} g^2_{kj}\, c_k,
\end{aligned}$$

$$\begin{aligned}
&E_m(\varepsilon_{r,k}\; \mathbf{z}'_k\; {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}) \\
&\quad = \mathbf{z}'_k (\mathbf{Z}'_U \mathbf{\Omega}^{-1}_U \mathbf{Z}_U)^{-1} \mathbf{Z}'_U\, \mathbf{\Omega}^{-1}_U\, E_m \\
&\quad (\varepsilon_{r,k} \mathbf{I}_N\, (\varepsilon_{r,1}\; {}_{bd}\delta_1, \ldots, \varepsilon_{r,k\,bd}\delta_k, \ldots, \varepsilon_{r,N}\; {}_{bd}\delta_N)') \\
&\quad = \sigma^2_r\, g_{kk}\, c_k\; {}_{bd}\delta_k,
\end{aligned}$$

where ${}_{bd}\mathbf{V}_r = \text{diag}\{c_k\; {}_{bd}\delta_k\}^N_{k=1}$, and $\mathbf{I}_N = \text{diag}\{1\}^N_{k=1}$.

The result (4.2) shows that it is possible to define a sampling strategy which aims at controlling the anticipated variances. Indeed, if the quantities ${}^{a}_{bd}\eta^2_{r,k}$ (or their proper predictions ${}^{a}_{bd}\hat{\eta}^2_{r,k}$) are used as a replacement for the residuals ${}_{bd}\eta^2_{r,k}$, the problem (3.1.1) defines a sampling design which allows one to guarantee the following conditions $\text{AV}({}_{bd}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \le {}_{bd}\bar{V}_r$ $(b = 1, \ldots, B;\ d = 1, \ldots, M_b;\ r = 1, \ldots, R)$.

An interesting result is the following. In the special case of a single partition, if the inclusion probabilities, π_k, and the heteroschedastic factors, c_k, are quite constant in each domain, then the selection of a balanced sample decreases the anticipated variance. This result is demonstrated in Falorsi and Righi (2008).

5. Brief extension to the case of a simple small area indirect estimator

If a given population partition defines a too large number of domains, it could happen that the budget constraints oblige to define a too large prefixed sampling errors of the direct estimators of the domains of the partition; in this situation, it could be necessary to adopt an indirect small-area estimator, in order to control the mean square errors of partition domain estimates. Herein in the following we will show as the sampling strategy, described in sections 2 and 3, may be extended to the case of a simple small area indirect estimator. Let us consider the enough general case in which the vector $\mathbf{x}_k$ of the auxiliary covariates has an intercept, such as $N_{bd} = \sum_{k \in s\, bd} w_k$.

Let $\ddot{b}$ denote the partition for which it is necessary to adopt a small area indirect estimator and let us consider the model (7.1.1) described in Rao (2005, page 116). In the herein studied context, the model for direct estimator, ${}_{\ddot{b}d}\hat{\bar{t}}_{r,\,\text{greg}} = {}_{\ddot{b}d}\hat{t}_{r,\,\text{greg}}/N_{\ddot{b}d}$, of the $\ddot{b}d$ domain may be defined as

$$ {}_{\ddot{b}d}\hat{\bar{t}}_{r,\,\text{greg}} = {}_{\ddot{b}d}\mathbf{a}'\boldsymbol{\varphi}_r + {}_{\ddot{b}d}h\ {}_{\ddot{b}d}v_r + {}_{\ddot{b}d}u_r $$
$$ (d = 1,\ \ldots,\ M_{\ddot{b}};\ r = 1,\ \ldots,\ R) \qquad (5.1) $$

where ${}_{\ddot{b}d}\mathbf{a}$ is a $p \times 1$ vector of area level covariates, $\boldsymbol{\varphi}_r$ is an unknown $p \times 1$ vector of regression coefficients, ${}_{\ddot{b}d}h$ is a known quantity related to the $\ddot{b}d^{\text{th}}$ domain, ${}_{\ddot{b}d}v_r \sim \text{iid}(0, {}_{\ddot{b}}\sigma^2_{r\nu})$ independent of the sampling error ${}_{\ddot{b}d}u_r \sim$ approximately $\text{ind}(0, {}_{\ddot{b}d}\sigma^2_{r\bar{t}})$, being ${}_{\ddot{b}d}\sigma^2_{r\bar{t}} = V_p({}_{\ddot{b}d}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}})/N^2_{\ddot{b}d}$. For known ${}_{\ddot{b}}\sigma^2_{r\nu}$ and ${}_{\ddot{b}d}\sigma^2_{r\bar{t}}$ values, the BLUP estimator of ${}_{\ddot{b}d}t_r$ is

$$ {}_{\ddot{b}d}\hat{t}_{r,\,\text{blup}} = N_{\ddot{b}d}\left({}_{\ddot{b}d}\gamma_r\ {}_{\ddot{b}d}\hat{\bar{t}}_{r,\,\text{greg}} + (1 - {}_{\ddot{b}d}\gamma_r)\,{}_{\ddot{b}d}\mathbf{a}'\hat{\boldsymbol{\varphi}}_r\right) \qquad (5.2) $$

being

$$ {}_{\ddot{b}d}\gamma_r = {}_{\ddot{b}}\sigma^2_{r\nu}\ {}_{\ddot{b}d}h^2 / ({}_{\ddot{b}d}\sigma^2_{r\bar{t}} + {}_{\ddot{b}}\,\sigma^2_{r\nu}\ {}_{\ddot{b}d}h^2) \qquad (5.3) $$

and

$$ \hat{\boldsymbol{\varphi}} = \left[\sum_{d=1}^{M_{\ddot{b}}} {}_{\ddot{b}d}\mathbf{a}\ {}_{\ddot{b}d}\mathbf{a}' \Big/ ({}_{\ddot{b}d}\sigma^2_{r\bar{t}} + {}_{\ddot{b}}\sigma^2_{r\nu}\ {}_{\ddot{b}d}h^2)\right]^{-1} $$
$$ \left[\sum_{l=1}^{M_{\ddot{b}}} {}_{\ddot{b}l}\mathbf{a}\ {}_{\ddot{b}l}\hat{\bar{t}}_{r,\text{greg}} \Big/ ({}_{\ddot{b}d}\sigma^2_{r\bar{t}} + {}_{\ddot{b}}\,\sigma^2_{r\nu}\ {}_{\ddot{b}d}h^2)\right] \qquad (5.4) $$

The MSE of the BLUP estimator is

$$ \text{MSE}({}_{\ddot{b}d}\hat{t}_{r,\,\text{blup}}) = N^2_{\ddot{b}d}\Bigg[\ {}_{\ddot{b}d}\gamma_r\ {}_{\ddot{b}d}\sigma^2_{r\bar{t}} + (1 - {}_{\ddot{b}d}\gamma_r)^2 $$
$$ {}_{\ddot{b}d}\mathbf{a}'\left(\sum_{d=1}^{M_{\ddot{b}}} {}_{\ddot{b}d}\mathbf{a}\ {}_{\ddot{b}d}\mathbf{a}' \Big/ ({}_{\ddot{b}d}\sigma^2_{r\bar{t}} + {}_{\ddot{b}}\sigma^2_{r\nu}\ {}_{\ddot{b}d}h^2)\right)^{-1} {}_{\ddot{b}d}\mathbf{a}\Bigg]. \qquad (5.5) $$

Looking at expressions (5.3) and (5.5), it is possible to note that for a given values of the variance ${}_{\ddot{b}}\sigma^2_{r\nu}$, it is possible to control the $\text{MSE}({}_{\ddot{b}d}\hat{t}_{r,\,\text{blup}})$ in planning the sampling design, by defining a proper value of the variance ${}_{\ddot{b}d}\sigma^2_{r\bar{t}}$. The following iterative procedure finds the π'_k inclusion probabilities which guarantee the minimum sample size and assure the respects of the following constraints: $V_p({}_{bd}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \le {}_{bd}\bar{V}_r$ (for $b \ne \ddot{b}; d = 1,\ \ldots,\ M_b$; $r = 1,\ \ldots,\ R$) and $\text{MSE}({}_{\ddot{b}d}\hat{t}_{r,\,\text{blup}}) \le {}_{\ddot{b}d}\bar{V}_r$ $(d = 1,\ \ldots,\ M_{\ddot{b}}$; $r = 1, \ldots, R)$.

Initialization: at initial iteration $(j = 0)$ find the ${}^{j}\pi'_k$ inclusion probabilities, solution of the problem (3.1.1), using the constraints $V_p({}_{\ddot{b}d}\hat{t}_{r,\,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \le {}_{bd}\bar{V}_r$ (for $b = 1, \ldots, B$; $d = 1,\ \ldots,\ M_b$; $r = 1, \ldots, R$).

Iteration: the generic iteration $(j = 1, 2, \ldots)$ is articulated as follows.

- Calculate ${}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}} = [N/(N^2_{\ddot{b}d}(N - Q))]\sum_{k \in U}[(1/{}^{j-1}\pi_k) - 1]\,{}_{\ddot{b}d}\eta^2_{r,k}$ $(d = 1,\ \ldots,\ M_{\ddot{b}};\ r = 1,\ \ldots,\ R)$.
- Calculate ${}^{j}_{\ddot{b}d}\gamma_r$ and ${}^{j}\text{MSE}({}_{\ddot{b}d}\hat{t}_{r,\,\text{blup}})$ $(d = 1,\ \ldots,\ M_{\ddot{b}}$; $r = 1,\ \ldots,\ R)$ respectively by means of equation (5.3) and (5.5) by using the sampling variances ${}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}}$ instead of ${}_{\ddot{b}d}\sigma^2_{r\bar{t}}$.
- Calculate ${}^{j}_{\ddot{b}d}\text{eff}_r = {}^{j}\text{MSE}({}_{\ddot{b}d}\hat{t}_{r,\,\text{blup}})/({}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}}\,N^2_{\ddot{b}d})$.
- Find the ${}^{j}\pi'_k$ inclusion probabilities, solution of the problem (3.1.1), using the ${}^{j}_{\ddot{b}d}\eta^2_{r,k} = {}_{\ddot{b}d}\eta^2_{r,k}\ {}^{j}_{\ddot{b}d}\text{eff}_r$ $(d = 1,\ \ldots,\ M_{\ddot{b}};\ r = 1,\ \ldots,\ R;\ k = 1,\ \ldots,\ N)$ as replacement for the ${}_{\ddot{b}d}\eta^2_{r,\,k}$ values.

Check: if the following condition is satisfied, for a small quantity ν, $\sum_{k \in U} |{}^{j-1}\pi_k - {}^{j}\pi_k| \le \nu$, then the algorithm stops and the inclusion probabilities π'_k are those calculated at iteration j. Otherwise, the iteration is calculated over and over again until the above condition is respected.

6. Empirical analysis

In order to verify the empirical properties of the proposed sampling strategy, two experiments have been implemented. Both experiments have showed good performances of the proposed strategy. The first experiment, on artificial data, is described in Falorsi and Righi (2008); the whole sampling strategy proposed in section 2 is implemented including the sampling allocation described in sections 3.1 and 3.2. The second experiment, based on a simulation on real enterprise data, is described herein in the following.

The analysis has been carried out on the 1999 population of the enterprises from 1 to 99 employees belonging to the *Computer and related economic activities* (2-digits of the NACE rev.1 classification. The data base used for the simulation study has $N = 10{,}392$ enterprises. The *value added* and *labour cost* are the variables of interest chosen in the simulation. The variable values are available for each unit in the population by an administrative data source. We consider two partitions: (DOM1) *geographical region* with 20 marginal domains; (DOM2) *Economic activity group* (3-digits of the NACE rev.1 classification with 6 different groups) by *Size class* (defined in terms of number of persons employed: $1 = 1-4$; $2 = 5-9$; $3 = 10-19$; $4 = 20-99$) with 24 marginal domains. Therefore, the overall number of marginal domains is 44, while the number of the cross-classification strata is 480 but only 360 strata have one or more population units.

In this study n is set equal to to 360. Five sampling designs have been considered, as reported in table 6.1. The first two benchmarking designs are two simple one-way stratification designs with simple random sampling without replacement in each stratum. The first design herein referred as STDOM1 is stratified by partition 1 and the second one, STDOM2, is stratified by partition 2. The marginal sample sizes for STDOM1 have been defined by (3.3.1). The parameter α_1 and the related marginal sample sizes n_{1d} $(d = 1, ..., 20)$ guarantee the percent Coefficient of Variation (CV) of the Horvitz-Thompson estimates of totals of the *auxiliary variable number of employers* be lower than than 34.5% for all domain of the partition 1. Analougsly, the parameter value α_2 has been defined by means of (3.3.1), assuring that, with the STDOM2 sample design, the percent CV of the Horvitz-Thompson estimates of totals of the auxiliary variable are lower than 8.7% for all the domains of the partition 2. In the following we refer to the domains with the planned sample size greater than the sample size deriving from an allocation rule with $\alpha_b = 1$ $(b = 1, 2)$ as *small domains*. These domains need to be oversampled to bound the sampling errors (Marker 2001).

We note that the above allocation rules are straightforward to implement in any real survey contexts. Two balanced sample designs are examined respecting the marginal sample sizes defined by STDOM1 for the first partition and by STDOM2 for the second one: the BAL design consider the balancing equations (2.2.2) with the specification (2.3.2) of the $\mathbf{z}_k$ vector; the BALPOP samples satisfy (or approximately satisfy) the following balancing equations $\sum_{k \in s} \pi_{k\ bd} \delta_k / \pi_k = n_{bd}$ and $\sum_{k \in s\ bd} \delta_k / \pi_k = N_{bd}$ $(b = 1, ..., B;\ d = 1, ..., M_b)$. The probabilities π_k of both designs have been obtained with the simplified procedure described in section 3.3. Furthermore, the comparison has been completed considering a coordinated design (referred as CPAR) selecting a single sample for each marginal population with Pareto Sampling (Särndal and Lundström 2005) and assuring the maximum overlap of the two samples. The marginal sample sizes, respectively defined by the STDOM1 and STDOM2 designs, are satisfied only as expectation over repeated sampling in the CPAR design; the inclusion probabilities are computed with the iterative procedure described in Falorsi *et al.* (2006). Five hundred Monte Carlo samples have been selected for each sampling design.

For each sample, the estimates of the domain totals have been computed by the *Horvitz-Thompson* (HT) estimator, *modified* GREG (greg) estimator and *synthetic* (syn) estimator, expressed as ${}_{bd}\hat{t}_{r,\text{syn}} = \sum_{k \in U_{bd}} \tilde{y}_{r,k}$. As far as the estimators using auxiliary information are concerned, two simple homoschedastic linear models have been implemented: the model (6.1) uses 10 auxiliary variables, six of them are the *economic activity group* membership indicators, and the remaining four are the *size class* membership indicators; the model (6.2) uses the 44 domain membership indicator variables. The linear model (6.1) is expressed by

$$E_m(y_k) = \beta_h + \beta_j \text{ for } k \in U_h \cap U_j, \qquad (6.1)$$

where U_h is the population of enterprises of h^{th} $(h = 1, ..., 6)$ *economic activity group* and U_j is the population of enterprises of j^{th} $(j = 1, ..., 4)$ *size class* of the number of employers and β_h and β_j are the fixed effects of the h^{th} economic activity group and of the j^{th} size class.
The linear model (6.2) is

$$E_m(y_k) = \beta_{1d} + \beta_{2d} \text{ for } k \in U_{1d} \cap U_{2d}, \qquad (6.2)$$

where β_{1d} and β_{2d} are the separate domain-specific effects.

Table 6.1
Sampling design used in the simulation study

Sampling Design	Abbreviation
Stratified by Partition 1 with SRSWOR* in each stratum	STDOM1
Stratified by Partition 2 with SRSWOR* in each stratum	STDOM2
Balanced sampling on the marginal sample sizes and on population sizes	BALPOP
Balanced sampling on the marginal sample sizes	BAL
Coordinated Pareto sampling	CPAR

*SRSWOR: Simple Random Sampling Without Replacement

We point out that the main aim of the experiment is to compare different sampling designs using the same estimator. In this context, the choice of the *best model* does not represent a central issue; hence, we have considered two quite general feasible models that can be implemented in all situations of planned domains. The model (6.1) is somewhat more reliable, since the estimates of the regression parameters are based on large sample sizes; while in model (6.2) it is possible to evaluate the effect of planning the domain sample sizes, although the estimates of each regression parameter are based on small sample sizes. Using the model (6.2) the syn and the greg estimators give identical results. In the following each sampling strategy is indicated in short by the couple (dis, est), where dis indicates one of the 5 sample designs referred in table 6.1 and est assumes the categories HT, syn, and greg above indicated.

In the following the analysis is based on the set of small domains. Two quality measures have been computed: the average *Absolute mean Relative Bias* $(\overline{\text{ARB}})$ and the average *Relative Mean Square Error* $(\overline{\text{RMSE}})$ expressed by

$$\overline{\text{ARB}}_F(\text{dis,est}) = \frac{1}{\text{card}(F)}\sum_{bd\in F}\left|\frac{1}{500}\sum_{i=1}^{500}\left[{}_{bd}\hat{t}^{i}_{r,\text{est}}(\text{dis}) - {}_{bd}t_r\right]\Big/{}_{bd}t_r\right|\times 100,$$

$$\overline{\text{RMSE}}_F(\text{dis,est}) = \frac{1}{\text{card}(F)}\sum_{bd\in F}\left\{\frac{1}{500}\sum_{i=1}^{500}\left[{}_{bd}\hat{t}^{i}_{r,\text{est}}(\text{dis}) - {}_{bd}t_r\right]^2\Big/{}_{bd}t_r^2\right\}\times 100$$

denoting with: F a specific subset of the marginal domains; $\text{card}(F)$ the cardinality of F; ${}_{bd}\hat{t}^{i}_{r,\text{est}}(\text{dis})$ the i^{th} Monte Carlo sample estimate ($i = 1, ..., 500$) of the total ${}_{bd}t_r$ in the strategy (dis, est). In particular, F represents alternatively the subset of small domains of DOM1, DOM2 or the overall set of small domains (of both DOM1 and DOM2).

The Monte Carlo simulation study highlights that the multi-way stratification techniques proposed in this paper are able to take bias and variability under control with respect to two benchmark strategies (STDOM1 and STDOM2) collapsing one of the two stratification variables.

The main results of the experiment referred to the small domains set are shown in table 6.2. The table is organised in four blocks: the first one illustrates the quality measures of the HT estimator; the second and third block are dedicated respectively to the syn and greg estimators based on 10 auxiliary variables (model (6.1)); the forth block presents the results of syn or greg estimators based on the 44 domain membership indicator variables (model (6.2)). We restrict the comments only on the *value added* variable, but similar consideration could be expressed for the *labour cost* variable. In general, the comments are referred to the overall set of small domains.

Examining firstly the HT estimator, we observe the following.

- The two benchmark designs (STDOM1 and STDOM2) have an $\overline{\text{RMSE}}$ value for the unplanned domains equal to 148.28% and 107.49% respectively. These values cause the large $\overline{\text{RMSE}}$ values computed for the overall set of small domains and respectively equal to 102.74% and 55.23%.
- The STDOM2 shows better results than those attained by STDOM1. This finding is explained by the fact that the STDOM2 stratification criterion is correlated with the variables of interest and takes under control a larger number of small domains than the STDOM1 stratification.
- As far as the overall set of small domains, the BALPOP is the more efficient design, both in terms of $\overline{\text{ARB}}$ (1.06%) and $\overline{\text{RMSE}}$ (32.58%), even if BAL is only slightly worse.
- The strategy adopting the coordinated sampling shows worse values with respect to balanced sampling but it performs better in terms of $\overline{\text{RMSE}}$ than benchmark strategies.

Considering the synthetic estimator based on 10 auxiliary variables, some issues may be pointed out.

- All designs are characterized by a large bias. The STDOM1 has an $\overline{\text{ARB}}$ equal to 13.99% (although it has an unacceptable $\overline{\text{RMSE}}$ that is equal to 65.16%). The rest of the designs have the $\overline{\text{ARB}}$ values higher than 18%. This evidence gives a warning against the use of synthetic estimator.
- The STDOM2 design has the lowest $\overline{\text{RMSE}}$ (26.16%), because of a strong reduction of the DOM1 variance. However, the $\overline{\text{ARB}}$ value (20.34%) is the largest than all designs.
- The behaviour of balanced and coordinated designs in terms of bias and variance are more or less equal. The BAL has the lowest $\overline{\text{ARB}}$ (18.33%) and $\overline{\text{RMSE}}$ (31.61%) values.

The experimental results of the greg estimator suggest some considerations.

- All the designs show strong improvements of the quality measures. In general, the $\overline{\text{ARB}}$ measure has a remarkable reduction with respect to the same indicator computed on the synthetic estimator. Only the STDOM1 still presents a high $\overline{\text{ARB}}$ value (7.40%).
- In the STDOM2, the reduction of the bias is more than compensated from the increase of the variability. This produces an $\overline{\text{RMSE}}$ equal to 34.05%.
- Both the balanced and the coordinated designs have good performances, though the balanced designs are slightly better being the $\overline{\text{RMSE}}$ roughly equal to the 23%.

Finally in the fourth block we note that the syn or greg estimator based on 44 auxiliary variables show analogous results to those of the greg estimator based on 10 auxiliary variables. The balanced designs are the best with slight preference for the BALPOP sampling.

As general findings, the balanced designs seem to guarantee a good strategy to take under control bias and variance of the overall set of the small domains.

The conclusion is that for all blocks, BALPOP generally shows the best overall performance with respect to bias and accuracy. The strategy based on the BALPOP sample design coupled with the greg estimator with the ten auxiliary variables (block 3) is a safe choice for both value added and labour cost. The BAL design performs well too. Moreover, the results show that the synthetic estimator of block 2 must be considered carefully because the bias can be unexpectedly large and the squared bias would be the dominating part of the $\overline{\text{RMSE}}$.

Table 6.2
Average Absolute Relative Bias $(\overline{\text{ARB}})$ and Relative Mean Square Error $(\overline{\text{RMSE}})$ of small domain sampling strategies

Sampling Design	Value Added						Labour Cost					
	DOM1		DOM2		Overall		DOM1		DOM2		Overall	
	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$
Horvitz-Thompson estimator (block 1)												
STDOM1	1.79	43.19	8.18	148.28	5.41	102.74	1.72	42.82	6.86	155.87	4.63	106.88
STDOM2	3.42	107.49	0.47	15.26	1.75	55.23	3.32	105.66	0.46	12.66	1.70	52.96
BALPOP	0.77	24.86	1.29	38.49	1.06	32.58	0.74	23.60	1.20	34.26	1.00	29.64
BAL	0.84	25.43	1.45	40.61	1.19	34.03	0.79	24.22	1.57	35.80	1.23	30.78
CPAR	1.35	32.52	2.18	53.85	2.18	44.60	1.44	31.68	2.62	51.44	2.11	42.88
Synthetic estimator with 10 auxiliary variables (block 2)												
STDOM1	14.22	18.88	13.81	100.55	13.99	65.16	12.29	18.40	9.25	95.03	10.57	61.83
STDOM2	24.82	33.96	14.48	15.96	20.34	26.16	13.13	14.79	12.46	23.11	12.75	19.51
BALPOP	13.68	17.51	24.98	43.98	20.09	32.51	11.89	15.60	12.35	33.08	12.15	25.50
BAL	14.92	18.46	21.82	41.66	18.83	31.61	13.37	16.91	10.41	32.64	11.69	25.82
CPAR	13.68	17.83	23.45	44.63	19.22	33.02	11.82	16.13	11.69	34.93	11.75	26.78
Modified GREG estimator with 10 auxiliary variables (block 3)												
STDOM1	2.35	30.13	11.26	119.95	7.40	81.03	1.86	29.28	11.79	119.23	7.49	80.25
STDOM2	3.98	58.62	0.95	15.26	2.26	34.05	2.90	52.66	0.93	12.66	1.78	29.99
BALPOP	1.11	19.41	2.20	25.80	1.73	23.03	1.01	16.42	1.99	21.73	1.57	19.43
BAL	1.63	19.41	1.76	26.11	1.70	23.21	1.21	16.72	2.08	21.96	1.70	19.69
CPAR	1.04	21.27	1.63	29.30	1.37	25.82	1.03	18.27	1.11	24.60	1.08	21.86
Synthetic or Modified GREG estimator with 44 auxiliary variables (block 4)												
STDOM1	3.39	31.30	27.48	63.22	17.04	49.39	2.76	30.80	28.67	63.05	17.44	49.08
STDOM2	17.24	102.24	1.37	20.65	8.25	56.00	23.00	102.64	1.42	19.10	10.77	55.30
BALPOP	1.07	20.71	1.97	26.98	1.58	24.26	1.08	17.62	1.93	24.07	1.56	21.27
BAL	1.47	20.36	2.13	28.46	1.84	24.95	1.41	17.66	2.02	25.10	1.75	21.88
CPAR	1.79	23.38	2.22	32.39	2.03	28.48	1.65	20.73	2.08	30.39	1.90	26.21

7. Conclusions

This work illustrates an efficient sampling strategy useful for obtaining planned sample size for domains belonging to different partitions of the population and in order to guarantee that sampling errors of domain estimates are lower than given thresholds. The sampling strategy, that covers the multivariate-multi-domain case, is useful when the overall sample size is bounded. In these instances the standard solution, using a stratified sample with the strata given by the cross-classification of variables defining the different partitions, is not feasible since the number of strata is larger than the overall sample size.

The sampling strategy which is proposed is based on the use of the balanced sampling selection technique and on a GREG-type estimator. The proposal may be easily extended to a strategy employing the use of both direct and indirect small area estimators.

The easy feasibility is one of the main advantages of the proposed solution since it is implemented by algorithms that are either based on free software tools or suitable for automated spreadsheets. But some other interesting aspects seem to appear.

The empirical analysis of real enterprise data shows good performances of the proposed strategy, which seems to be robust even when departing from ideal conditions (*i.e.*, the estimates appear to be of high quality even when the inclusion probabilities of the sample differ from the optimal ones). These results encourage additional work to give a systematic account of conditions under which the proposed method will have good performance.

Furthermore, the proposed strategy does seems to work well for large datasets, in terms of computer time, and therefore it seems to be suitable for large scale surveys.

Finally, the approach represents an original overall small area sampling strategy, which jointly considers the sampling design and the estimator. The paper deeply analyzes the design issues, but more research is needed to study more carefully the estimation issues. In particular, future research should be focused on the improvement of the model-based or model-assisted estimators due to the presence of sample units in each estimation domain, allowing the use of models with specific small area effects and giving more accurate estimates of the parameters of interest at small area level. These aspects seem to be an appealing strength to be investigated.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 47-57.

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.

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

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

Deville, J.-C., and Tillé, Y. (2000). Selection a several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, 86, 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.

Dykstra, R., and Wollan, P. (1987). Finding I-projections subject to a finite set of linear inequality constraints. *Applied Statistics*, 36, 377-383.

Falorsi, P.D., Orsini, D. and Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. *Statistics in Transition*, 7, 1173-1198.

Falorsi, P.D., and Righi, P. (2008). An efficient multi-way design strategy for domain estimation. *Contributi ISTAT* (downloadable on http://www.istat.it/dati/pubbsci/contributi/).

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal American Statistical Society*, 65, 776-795.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51-55.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for Domains, including small domains. *Survey Methodology*, 1, 33-44.

Lu, W., and Sitter, R.R. (2002). Multi-way stratification by linear programming made practical. *Survey Methodology*, 2, 199-207.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.

Montanari, G.E., and Ranalli, M.G. (2003). Nonparametric methods in survey sampling. In *New Developments in Classification and Data Analysis*, (Eds., M. Vinci, P. Monari, S. Mignani, A. Montanari), Springer, Berlin.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: Springer-Verlag.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.

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

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