CONTRIBUTI ISTAT

n. 12/2008

A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation

P.D. Falorsi e P. Righi

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

ISTAT – Produzione libraria e centro stampa

CONTRIBUTI ISTAT

n. 12/2008

A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation

P.D. Falorsi(*) e P. Righi(*)

(*) ISTAT – Servizio Progettazione e supporto metodologico nei processi di produzione

I lavori pubblicati riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto

Contributi e Documenti Istat 2008

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

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 that the sampling errors of domain estimates are lower than given thresholds. The sampling strategy that covers the multivariate-multidomain 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 design and estimation phase 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 *et al.* (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, *et al.*, 2003).
In fact, 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 interest 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 unfeasible 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.

In order to explain the problem, let us consider the population of 165 schools (Cochran, 1977, pag. 124) reported in table 1.1. Let us suppose that the parameters of interest are the totals of a $\boldsymbol{y}$ variable - related to school - separately for (*i*) *size of city* (5 categories: *I,II,III,IV,V*) and (*ii*) *Expenditure per pupil* (4 categories: *A,B,C,D*). Let us note the following: (*i*) the domains of interest are 9=5+4; (*ii*) the problem defines two distinct partitions of the population, indeed the size of city represents a partition of the target population in 5 non-overlapping domains, and the expenditure per pupil is an alternative partition in 4 domains. The standard *cross-classification design* defines 20=5x4 strata by crossing the categories of the domains of the two partitions; in each stratum should be selected at least one school (or two schools for estimating the sampling variance without bias) and, consequently, according to this design, the sample size should be of 20 (or 40 schools) at least. If the budgetary constraints limit the sample size to 10 schools the cross-classification design becomes unfeasible.

Table 1.1. Number of schools by Size of City and expenditure per pupil

Size of	*Expenditure per pupil*				*Totals*	*Sample size*
City	*A*	*B*	*C*	*D*		
I	15	21	17	9	62	4
II	10	8	13	7	38	2
III	6	9	5	8	28	2
IV	4	3	6	6	19	1
V	3	2	5	8	18	1
Totals	38	43	46	38	165	
Sample size	2	3	3	2		10

Cochran (1977, pag. 124)

The above background is typical of the business survey context. Indeed, the European *Council Regulation* n°58/97 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. As we may note by table 1.2, 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.

Table 1.2. 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.

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 main approaches: (*i*) Latin Squares or Latin Lattices schemes (Bryant *et al.*, 1960; Jessen, 1970); (*ii*) controlled rounding problems via linear programming (Causey *et al.*, 1985; Sitter and Skinner, 1994). The seminal paper of Bryant *et al.* (1960) suggests allocating the units in the sample by means of a two-way Latin Square table randomly selected and two estimators of the parameter of interest are proposed. The method has some drawbacks that limit the application in real survey contexts. For example, it implies that the expected sample counts in each stratum display independence between the rows and columns of the two-way table; furthermore it is not possible to implement the procedure when there is no population in one or more cross-classification strata. In order to solve these problems, Jessen (1970) proposes two approaches, both fairly complicated to implement and not always leading to a solution (Causey *et al.*, 1985). As concerns the methods based on linear programming are concerned, Causey *et al.* (1985) consider the controlled multi-way stratification as a rounding problem solved by means of transportation theory. The method may not have a solution in case of three or more stratification variables. Following the linear programming approach proposed by Rao and Nigam (1990, 1992), Sitter and Skinner (1994) suggest a method based on linear programming more flexible in different situations than the method proposed by Causey *et al.* (1985) and some further computational simplification of Sitter and Skinner method have been suggested by Lu and Sitter (2002). Nevertheless, the main weakness of the linear programming approach is the computational complexity. As a consequence, the drawbacks of both approaches have limited the use of multi-way stratification techniques as a standard solution for planning the survey sampling designs.

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, Tillé, 2004) and a *greg-type* estimation (Lehtonen, *et al.*, 2003). As shown in the study on empirical data, the main advantages of this solution is the computational feasibility and the efficiency, that is the sampling errors for multidomain-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 design and estimation phase and improving the efficiency of the direct domain estimators.

It is worthwhile to note that, 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. However, as briefly sketched in section 5, the indirect estimation is strengthened by the use of an improved direct estimator.

The paper is organised as follows. Section 2 states the problem and introduces the essential notation; moreover it describes the overall sampling strategy. Section 3 shows an iterative procedure that defines the inclusion probabilities and the corresponding planned domain sample sizes solving a non linear problem where the objective function is the minimization of the overall sample size guaranteeing the sampling errors of domain estimates to be lower than given thresholds. 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 two empirical studies conducted both on a real population of Italian enterprises and on a simulated population are shown in section 6. Finally some brief conclusions are 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 $\sum_{b=1}^{B} M_b = Q$ the overall number of domains.

In the table 1.1 example, b=1 when considering the partition of the population by size of the city and b=B=2 when considering the partition by expenditure per pupil, being $M_1 = 5$ and $M_2 = 4$. Continuing the example, U_{11} individuates the domain, of $N_{11} = 62$ schools, having the *I*-th city size; furthermore U_{24} individuates the domain, of $N_{24} = 38$ schools, having the *D-th* expenditure per pupil.

Let $y_{r,k}$ and ${}_{bd}\delta_k$ denote respectively the value of the $\mathcal{Y}_r$ (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 $M = R \times Q$ domains totals

$$ {}_{bd}t_r = \sum_{k \in U} y_{r,k}\ {}_{bd}\delta_k = \sum_{k \in U_{bd}} y_{r,k} \quad (r = 1,\ldots,R\,;\ b=1,\ldots,B;\ d=1,\ldots,\ M_b). \qquad (2.1.1) $$

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

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 π_k ($k \in U$). Let $s_{bd} = s \cap U_{bd}$ denote 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 vector of auxiliary variables known for each population unit 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 ${}_{bd}\hat{t}_{r,greg}$, are obtained with the *modified greg* estimator (Rao, 2003, page 20), given by (see section 2.5):

$$_{bd}\hat{t}_{r,greg} = \sum_{k \in s} {}_{bd}w_k \ y_{r,k} \tag{2.2.3}$$

where:

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

denote the sampling weights, $\mathbf{x}_k$ indicates a vector of 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), also known as *survey regression estimator* (Battese, Harter and Fuller, 1988), may be derived under the following *working* superpopulation model

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

where $\boldsymbol{\beta}_r$ denotes 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:

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

being

$$_{bd}\eta_{r,k} = \begin{cases} \varepsilon_{r,k} - \mathbf{z}'_k \ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon} & \text{for } k \in U_{bd} \\ -\mathbf{z}'_k \ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon} & \text{for } k \in U_{\overline{bd}} \end{cases} ,$$

with

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

where $U_{\overline{bd}}$ is 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 $V_p({}_{bd}\hat{t}_{r,greg} \mid \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}})$ are lower than prefixed level of precision thresholds, $_{bd}\overline{V}_r$:

$$V_p({}_{bd}\hat{t}_{r,greg} \mid \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \leq {}_{bd}\overline{V}_r \quad (b=1,\ldots,B;\ d=1,\ldots,M_b;\ r=1,\ldots,R)$$

The technical details are described in section 3.

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 (Royall and Herson, 1973; Valliant *et al.*, 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 $\mathcal{S}$ of all the subset s of the population U such that $\sum_{s\in\mathcal{S}} p(s)=1$, 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 $\boldsymbol{\pi}=E_p(\boldsymbol{\lambda})=\sum_{s\in\mathcal{S}} p(s)\boldsymbol{\lambda}$, 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 $s\in\mathcal{S}$ 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

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

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

$$\mathbf{z}'_k = \underbrace{(\overbrace{0,...,\pi_k,...,0}^{b=1},...,\overbrace{0,...,\pi_k,...,0}^{b=B})}_{Q} = \pi_k({}_{11}\delta_k,...,{}_{bd}\delta_k,...,{}_{BM_B}\delta_k)\,. \tag{2.3.2}$$

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), $\sum_{k\in U}(\mathbf{z}_k\lambda_k)/\pi_k = \sum_{k\in U}\mathbf{z}_k$, guarantees that the n_{bd} values are non random quantities.
Referring to the bd-th domain, the left hand-side of the balancing equation (2.2.2) is

$$\sum_{k\in U}(\pi_k\, {}_{bd}\delta_k\, \lambda_k)/\pi_k = \sum_{k\in U_{bd}} \lambda_k = \sum_{k\in s_{bd}} 1 = n_{bd}\,,$$

while the right hand-side is

$$\sum_{k\in U} \pi_k \, {}_{bd}\delta_k = \sum_{k\in U_{bd}} \pi_k = n_{bd} \, .$$

One relevant drawback of balanced sampling has always been implementing a general procedure giving a multivariate balanced random sample (see Valliant *et al.*, 2000). 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 (2.3.3) can be exactly satisfied. A free SAS (version 9) software code for the selection of balanced samples for large data sets 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

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

where $\tilde{y}_{r,k}$ denotes the prediction of $y_{r,k}$ under the assumed superpopulation 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 superpopulation model (2.2.4). The predictions $\tilde{y}_{r,k}$ are then obtained by

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

being

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

Let us observe that the linear model (2.2.4) allows us 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 generalised linear models (Lehtonen and Veijanen, 1998) or non parametric regression techniques (Montanari and Ranalli, 2003).

As noted by Rao (2003, pag. 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; furthermore it may also be viewed as a calibration estimator (Singh and Mian, 1995) with weights ${}_{bd}w_k$ minimizing a chi-squared distance $\sum_{k\in s} c_k \, (a_k \, {}_{bd}\delta_k - {}_{bd}w_k)^2 / a_k$ subject to the constraints $\sum_{k\in s} {}_{bd}w_k \, \mathbf{x}_k = {}_{bd}t_{\mathbf{x}}$.

The sum of the ${}_{bd}\hat{t}_{r,greg}$ estimates over all the domains of a partitions is benchmarked to the usual greg estimate of the total, $\sum_{d=1}^{M_b} {}_{bd}\hat{t}_{r,greg} = \hat{t}_{r,greg}$, being

$$\hat{t}_{r,greg} = \sum_{k\in s} y_{r,k}\, a_k [1 + (\sum_{k\in U}\mathbf{x}_k - \sum_{k\in s}\mathbf{x}_k a_k)' \, (\sum_{k\in s}(\mathbf{x}_k\mathbf{x}'_k\, a_k / c_k))^{-1}\, \mathbf{x}_k / c_k].$$

2.5. Sampling variances

In order to derive the expression of the variance (2.2.5), consider the result given in expression (7) of Deville and Tillé (2005), which takes into account the Horvitz-Thompson estimator $\hat{t}_{r,ht} = \sum_{k\in s} y_{r,k}\, a_k$ of the total $t_r = \sum_{k\in U} y_{r,k}$. This result states that, under balanced sampling, a good approximation of the sampling variance of the $\hat{t}_{r,ht}$ estimator is given by

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

where

$$\mathbf{B}_{\mathbf{z},y} = \left(\sum_{k\in U}\mathbf{z}_k\mathbf{z}'_k\left(\frac{1}{\pi_k} - 1\right)\right)^{-1} \sum_{k\in U}\mathbf{z}_k\, y_{r,k}\left(\frac{1}{\pi_k} - 1\right). \qquad (2.5.2)$$

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

$${}_{bd}\hat{t}_{r,greg} \cong {}_{bd}\hat{t}^{*}_{r,greg} = \sum_{k\in U_{bd}} \mathbf{x}'_k\, \boldsymbol{\beta}_r + \sum_{k\in s_{bd}} a_k\ \varepsilon_{r,k} = \sum_{k\in U_{bd}} \mathbf{x}'_k\, \boldsymbol{\beta}_r + \sum_{k\in s} a_k\ \varepsilon_{r,k}\ {}_{bd}\delta_k .$$

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

$$\begin{aligned} V_p({}_{bd}\hat{t}_{r,greg} \mid \hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) &\cong V_p({}_{bd}\hat{t}^{*}_{r,greg} \mid \hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) = \\ &= V_p(\sum_{k\in U_{bd}} \mathbf{x}'_k\, \boldsymbol{\beta}_r + \sum_{k\in s} a_k\ \varepsilon_{r,k}\ {}_{bd}\delta_k \mid \hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) = \\ &= V_p(\sum_{k\in s} a_k\ \varepsilon_{r,k}\ {}_{bd}\delta_k \mid \hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) = \\ &= V_p(\sum_{k\in s} a_k\ \varepsilon_{r,k}\ {}_{bd}\delta_k + (\mathbf{t}_{\mathbf{z},ht} - \hat{\mathbf{t}}_{\mathbf{z},ht})'\ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}) = \\ &= V_p(\sum_{k\in s} a_k\ (\varepsilon_{r,k}\ {}_{bd}\delta_k - \mathbf{z}'_k\ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}) = \\ &= V_p(\sum_{k\in s} a_k\ {}_{bd}\eta_{r,k}) \cong \frac{N}{N-Q}\sum_{k\in U}\left(\frac{1}{\pi_k} - 1\right) {}_{bd}\eta^2_{r,k} . \end{aligned}$$

The approximated sampling variance of ${}_{bd}\hat{t}_{rgreg}$ 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, ${}_{bd}t_r$. Let us suppose furthermore that the selected sample respects the *balancing equations,* $\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}$, being fixed the overall sample size *n*.

It is trivial to demonstrate 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,..., *B*; *d*=,..., M_b; *r*=1,..., *R*)

$$\pi_k = {}_{bd}\ddot{\pi}_{r,k} = n\left|{}_{bd}\eta_{r,k}\right| \Big/ \sum_{i \in U} \left|{}_{bd}\eta_{r,i}\right| . \qquad (2.5.3)$$

If the balanced sample is selected using the probabilities ${}_{bd}\ddot{\pi}_{r,k}$, the approximated variance $V_p({}_{bd}\hat{t}^{*}_{r,greg} \,|\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}})$ reaches its minimum value, ${}_{bd}V^{*}_{r|n}$, expressed by

$$V_p({}_{bd}\hat{t}_{r,greg} \,|\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) \geq {}_{bd}V^{*}_{r|n} = \frac{1}{n}\left(\sum_{k \in U} \left|{}_{bd}\eta_{r,k}\right|\right)^2 - \sum_{k \in U} {}_{bd}\eta^2_{r,k} . \qquad (2.5.4)$$

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 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 phase procedure: (*i*) in the first phase, denoted, as *optimization,* the preliminary inclusion probabilities, π'_k, are determined solving a minimum constrained problem; (*ii*) in the second phase, 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 phase of 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 phase. The empirical results may then be used for introducing proper adjustments in planning the next survey design. However, as illustrated in empirical analysis of section 6, the proposed strategy seems to be efficient and sufficiently robust with respect to small departures of ideal conditions.

3.1. 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

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

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…) which represent the solution of the following non linear problem $Min\left(\sum_h v_h\right)$ under the constraints $\sum_h A_{rh}/v_h \leq \overline{A}_r$, where A_{rh} and $\overline{A}_r$ (*r*=1,2,…) are known positive quantities. Bethel (1989) invokes the Khun-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, which is suitable for automated spreadsheets, but does not prove his algorithm always converges. A slight modification of the Chromy's algorithm – able to solve the problem (3.1.1) guaranteeing the inequalities $0<\pi'_k\leq 1$ $(k=1,...,N)$ 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.*

<u>Initialization</u>

Let τ $(\tau=0,1,2,..)$ denote the generic iteration. At initial iteration ($\tau=0$), set ${}^{\tau}\gamma_k=1$ (*k*=1,…,*N*).

<u>Calculus</u>

The generic iteration ($\tau=1,2,...$) develops the Chromy's algorithm and consists of a sequence of steps. The index *u* (*u*=0,1,…) - after a comma on the right of the iteration index, τ - denotes the generic step.

- At initial step ($u=0$), set ${}^{\tau,u}_{bd}\phi_r=1$ (for *b*=1,…,*B*; *d*=1,…, M_b ; *r*=1,…,*R*) 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,…), calculate the ${}^{\tau,u}\pi_k$ values using the following equation

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

Calculate, furthermore

$${}^{\tau,u}_{bd}V_r = \frac{N}{N-Q}\sum_{k\in U}\frac{1}{{}^{\tau,u}\pi_k}\;{}_{bd}\eta^2_{r,k}\;{}^{\tau}\gamma_k \text{ , and } {}^{\tau,u}_{bd}V'_r = {}^{\tau,u}_{bd}V_r + {}^{\tau}_{bd}V_{0r} . \qquad (3.1.3)$$

- If the following two conditions are respected for all b=1,…,B; d=1,…, M_b ; r=1,…,R:

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

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

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

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

$${}^{\tau+1}\gamma_k = \begin{cases} 1 \text{ if } {}^{\tau}\pi_k \le 1 \\ 0 \text{ if } {}^{\tau}\pi_k > 1 \end{cases} . \qquad (3.1.6)$$

The calculus is iterated at $\tau+1$ with ${}^{\tau+1}\gamma_k$ replacing ${}^{\tau}\gamma_k$.
The 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. 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,…,B; d=1,…,M_b). Sometimes a further data manipulation could be necessary in order to assure the condition $\sum_{d=1}^{M_b} n_{bd} = \sum_{l=1}^{M_{b'}} n_{b'l} = n$, for each $b \neq b'$. The probabilities π_k are then obtained as solution of the following *calibration problem*

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

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 IPF (Bishop *et al.*, 1975) or GIPF (Dykstra (1985); 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.

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:

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

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

$$AV({}_{bd}\hat{t}_{r,greg} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \cong E_m\, V_p({}_{bd}\hat{t}^{*}_{r,greg} \,|\, \hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) =$$

$$= E_m\left[\frac{N}{N-Q}\sum_{k \in U_{bd}}\left(\frac{1}{\pi_k}-1\right)(\varepsilon_k - \mathbf{z}'_k\ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 + \frac{N}{N-Q}\sum_{k \in U_{\overline{bd}}}\left(\frac{1}{\pi_k}-1\right)(\mathbf{z}'_k\ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\ )^2\right] =.$$

$$= E_m\left[\frac{N}{N-Q}\sum_{k \in U_{bd}}\left(\frac{1}{\pi_k}-1\right)\varepsilon^2_{r,k} + \frac{N}{N-Q}\sum_{k \in U}\left(\frac{1}{\pi_k}-1\right)(\mathbf{z}'_k\ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 + \right.$$

$$\left. -2\ \frac{N}{N-Q}\sum_{k \in U_{bd}}\left(\frac{1}{\pi_k}-1\right)\varepsilon_{r,k}\ \mathbf{z}'_k\ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\right] = \frac{N}{N-Q}\sum_{k \in U}\left(\frac{1}{\pi_k}-1\right){}^{a}_{bd}\eta^2_{r,k} \tag{4.2}$$

being

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

where: $(g_{k1},...,g_{kj},...,g_{kN}) = \mathbf{g}'_k = \mathbf{z}'_k \left(\mathbf{Z}'_U \mathbf{\Omega}_U^{-1} \mathbf{Z}_U\right)^{-1} \mathbf{Z}'_U \mathbf{\Omega}_U^{-1}$, $\mathbf{Z}_U = col\{\mathbf{z}'_k\}_{k=1}^N$ denotes the $(N \times Q)$ matrix of $\mathbf{z}'_k$,$\mathbf{\Omega}_U^{-1} = diag\{1/\pi_k - 1\}_{k=1}^N$. The (4.2) has been obtained under the following two results

$$E_m(\mathbf{z}'_k\, {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 = \sigma_r^2\, \mathbf{z}'_k \left(\mathbf{Z}'_U \mathbf{\Omega}_U^{-1} \mathbf{Z}_U\right)^{-1} \mathbf{Z}'_U\, \mathbf{\Omega}_U^{-1}\, {}_{bd}\mathbf{V}_r\, \mathbf{\Omega}_U^{-1} \mathbf{Z}_U \left(\mathbf{Z}'_U \mathbf{\Omega}_U^{-1} \mathbf{Z}_U\right)^{-1} \mathbf{z}_k =$$
$$= \sigma_r^2\, \mathbf{g}'_k\, {}_{bd}V_r\, \mathbf{g}_k = \sigma_r^2 \sum_{j \in U_{bd}} g_{kj}^2\, c_k\ ,$$
$$E_m\left(\varepsilon_{r,k}\, \mathbf{z}'_k\, {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\right) =$$
$$= \mathbf{z}'_k \left(\mathbf{Z}'_U \mathbf{\Omega}_U^{-1} \mathbf{Z}_U\right)^{-1} \mathbf{Z}'_U\, \mathbf{\Omega}_U^{-1}\, E_m(\varepsilon_{r,k}\, \mathbf{I}_N\, (\varepsilon_{r,1\ bd}\delta_1,...,\varepsilon_{r,k_{bd}}\delta_k,...,\varepsilon_{r,N\ bd}\delta_N)') =$$
$$= \sigma_r^2\, g_{kk}\, c_k\, {}_{bd}\delta_k\ ,$$

where ${}_{bd}\mathbf{V}_r = diag\{c_k\, {}_{bd}\delta_k\}_{k=1}^N$, and $\mathbf{I}_N = diag\{1\}_{k=1}^N$.
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 ${}_{bd}^{a}\eta_{r,k}^2$ (or their proper predictions ${}_{bd}^{a}\hat{\eta}_{r,k}^2$) are used as a replacement for the residuals ${}_{bd}\eta_{r,k}^2$, the problem (3.1.1) defines a sampling design which allows one to guarantee the following conditions that $AV({}_{bd}\hat{t}_{r,greg} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \leq {}_{bd}\overline{V}_r$ (b=1,…,B; d=1,…,M_b ; r=1,…,R).
An interesting result is the following. In the special case of a single partition, if the inclusion probabilities, π_k, and the etheroschedastic 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 Appendix 1.

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 $N_{bd} = \sum_{k \in s_{bd}} {}_{bd}w_k$.

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, pag. 116). In the herein studied context, for the domains of the $\ddot{b}-th$ partition, this model may be defined as

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

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 $\ddot{b}d-th$ domain, ${}_{\ddot{b}d}v_r$ ~ iid $(0, {}_{\ddot{b}}\sigma_{rv}^2)$

independent of the sampling error ${}_{\ddot{b}d}u_r$ ~ approximately *ind* $(0, {}_{\ddot{b}d}\sigma^2_{r\bar{t}})$, being ${}_{\ddot{b}d}\sigma^2_{r\bar{t}} = V_p({}_{\ddot{b}d}\hat{t}_{r,greg} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) / N^2_{\ddot{b}d}$. For known ${}_{\ddot{b}}\sigma^2_{rv}$ and ${}_{\ddot{b}d}\sigma^2_{r\bar{t}}$ values, the BLUP estimator of ${}_{\ddot{b}d}t_r$ is

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

being

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

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

The MSE of the BLUP estimator is

$$ MSE({}_{\ddot{b}d}\hat{t}_{r,blup}) = $$

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

Looking at expressions (5.5) and (5.3), it is possible to note that for a given values of the variance ${}_{\ddot{b}}\sigma^2_{rv}$, it is possible to control the $MSE({}_{\ddot{b}d}\hat{t}_{r,blup})$ in the sampling design phase, 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 $N/(N-Q)\sum_{k\in U}(1/\pi_k - 1)\,{}_{bd}\eta^2_{r,k} \le {}_{bd}\bar{V}_r$ (for $b \ne \ddot{b}$; d=1,…,M_b; $r = 1,...,R$) and $MSE({}_{\ddot{b}d}\hat{t}_{r,blup}) \le {}_{\ddot{b}d}V_r$ (d=1,…,$M_{\ddot{b}}$; $r = 1,...,R$).

Initialization

Let j ($j = 0,1,2,..$) denote the generic iteration. At initial iteration ($j = 0$), set ${}^{j}_{\ddot{b}d}\mathit{eff}_r = 1$ (d=1,…,$M_{\ddot{b}}$; r=*1,...,R*). By means of the algorithm described in section (3.1), find the ${}^{j}\pi'_k$ inclusion probabilities, solution of the problem (3.1.1), using the ${}^{j}_{\ddot{b}d}\eta^2_{r,k} = {}_{\ddot{b}d}\eta^2_{r,k}\;{}^{j}_{\ddot{b}d}\mathit{eff}_r$ (d=1,…,$M_{\ddot{b}}$; r=1,…,R; k=1,…,N) as replacement for the ${}_{\ddot{b}d}\eta^2_{r,k}$ values.

Iteration

The generic iteration ($j = 1,2,...$) 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,…,$M_{\ddot{b}}$; r=1,…,R).

- Calculate ${}^{j}_{\ddot{b}d}\gamma_r$ and ${}^{j}MSE({}_{\ddot{b}d}\hat{t}_{r,blup})$ (d=1,…,$M_{\ddot{b}}$; r=1,…,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 ${}^{j}_{\ddot{b}d}eff_r = {}^{j}MSE({}_{\ddot{b}d}\hat{t}_{r,blup}) / {}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}}$;
- Find the ${}^{j}\pi'_k$ inclusion probabilities, solution of the problem (3.1.1), using the ${}^{j}_{\ddot{b}d}\eta^2_{r,k} = {}_{\ddot{b}d}\eta^2_{r,k}\ {}^{j}_{\ddot{b}d}eff_r$ (d=1,…,$M_{\ddot{b}}$; r=1,…,R; k=1,…,N) as replacement for the ${}_{\ddot{b}d}\eta^2_{r,k}$ values.

Check

If the following condition is satisfied, for a small quantity v,

$$\sum_{k\in U} |{}^{j-1}\pi_k - {}^{j}\pi_k| \leq v\,, \tag{5.6}$$

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 condition (5.6) is respected.

6. Empirical Analysis

In order to verify the empirical properties of the proposed sampling strategies, two experiments have been implemented, the first one is on artificial data, the second experiment is based on a simulation on real enterprise data. In the experiment on artificial data the whole sampling strategy proposed in section 2 is implemented including the allocation phase described in section 3. The simulation on real enterprise data adopts a simplified allocation rule that may be easily implemented in real survey contexts but is different from that optimal described in section 3. Both experiments have showed good performances of the proposed strategy.

6.1. Artificial data

An artificial population U with 1,813 units has been created to evaluate different sampling strategies. Two categorical variables have been created: $\mathcal{U}_1$ with 5 categories (1,…,5) identifying the 5 domains U_{1d} of the first partition and $\mathcal{U}_2$ with 10 categories (1,…,10) identifying the 10 domains U_{2d} of the second partition. The values $u_{1,k}$ and $u_{2,k}$ of the variables $\mathcal{U}_1$ and $\mathcal{U}_2$ have been assigned to each unit in the population, generating a contingency table, assuming the categorical variables are independent. The contingency table (table 6.1.1) has skewed marginal distributions, where the marginal frequencies decrease as the levels of $\mathcal{U}_1$ or $\mathcal{U}_2$ increase. For each unit k, the auxiliary variable x_k has been created drawing from the normal distribution $N(30+u^3_{1,k};u^5_{2,k})$.

Table 6.1.1. Contingency table of artificial population and marginal sample sizes of the multi-way stratified balanced sampling

	Partition 2										Totals	Sample size
Partition 1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		
1	316	203	154	86	62	35	29	18	15	7	925	12
2	151	94	80	46	32	15	12	6	8	4	448	11
3	84	52	40	25	20	12	7	5	3	0	248	9
4	47	31	20	14	7	5	4	3	0	1	132	6
5	19	11	11	4	5	4	2	2	1	1	60	4
Totals	617	391	305	175	126	71	54	34	27	13	1,813	
Sample size	11	7	6	5	3	3	2	2	2	1		42

The variables of interest, $\mathcal{Y}_1$, has been generated with the following superpopulation model

$$y_{1,k} = 0.35x_k + \varepsilon_{1,k} \qquad (6.1.1)$$

with $E_m(\varepsilon_{1,k}) = 0$, $E_m(\varepsilon_{1,k}\,\varepsilon_{1,l}) = 0$, for $k \neq l$ and $V_m(\varepsilon_{1,k}) = 1.5x_k$.

Seven sampling designs, as reported in table 6.1.2, have been compared. The BAL design is the one described in the paper. The marginal sample size have been defined with the procedure described in section 3, assuring that the percent Coefficient of Variation (CV) of the first and second partition domain estimates are lower than 14% and 21% respectively: with the symbols of section 3, it is $\left(\sqrt{{}_{1d}\overline{V}_1} \big/ {}_{1d}t_1\right)100 = 14$ (d=1,…,5) and $\left(\sqrt{{}_{2d}\overline{V}_1} \big/ {}_{2d}t_1\right)100 = 21$ (d=1,…,10).

The overall sample size, n, is 42 units and the marginal distributions are shown in table 6.1.1. The marginal sample sizes of each partition are then adopted for the four one-way stratified design, OPT1, OPT2, STDOM1 and STDOM2, where the inclusion probabilities are described in table 6.1.2. The overall sample size n has been used to define the Simple Random Sampling Without Replacement and Probability Proportional to Size Sampling sampling designs.

Tab. 6.1.2. Sampling designs

	Sampling Design	Abbreviation	Inclusion Probability
	Multi-way stratification with balanced sampling	BAL	Defined in section 3
One –way stratification	Stratified by Partition 1 with Optimal inc. prob. for Partition 1	OPT1	$\pi_k = n_{1d}\left(\sqrt{x_k} / \sum_{k \in U_{1d}} \sqrt{x_k}\right)$
	Stratified by Partition 2 with Optimal inc. prob. for Partition 2	OPT2	$\pi_k = n_{2d}\left(\sqrt{x_k} / \sum_{k \in U_{2d}} \sqrt{x_k}\right)$
	Stratified by Partition 1 with SRSWOR* with in each stratum	STDOM1	$\pi_k = n_{1d} / N_{1d}$
	Stratified by Partition 2 with SRSWOR* with in each stratum	STDOM2	$\pi_k = n_{2d} / N_{2d}$
	SRSWOR*	SRS	$\pi_k = n / N$
	Probability Proportional to Size Sampling	PPS	$\pi_k = n\left(\sqrt{x_k} / \sum_{k \in U} \sqrt{x_k}\right)$

**SRSWOR: Simple Random Sampling Without Replacement*

Knowing all the population values, the sampling variances $V_p({}_{bd}\hat{t}_{1,greg})$ have been computed for each domain estimate for the modified greg estimator (2.2.3) based on the model (6.1.1). For the BAL design the variance has been calculated according to the expression (2.2.5), while the standard textbooks expression for the other designs has been adopted. Table 6.1.3 shows the mean values and maximum values of the percent CV by sampling design and partition, being the CV defined as $\sqrt{V_p({}_{bd}\hat{t}_{1,greg})} / {}_{bd}t_1$. Furthermore, the same mean and maximum CV values have been computed considering the overall set of 15 marginal domains.

Table 6.1.3 Mean and maximum values of the percent CV by sampling design and partition

Sampling Design	Mean			Max		
	Partition		Overall	Partition		Overall
	1	2		1	2	
BAL	13.6	20.0	16.8	14.1	21.5	21.5
OPT1	16.6	28.4	22.5	17.5	38.7	38.7
OPT2	18.7	24.0	21.4	21.5	27.7	27.7
STDOM1	17.7	35.4	26.5	19.0	67.2	67.2
STDOM2	21.1	25.3	23.2	26.7	30.0	30.0
SRS	21.5	33.0	27.3	27.7	55.4	55.4
PPS	18.3	26.6	22.5	20.8	36.3	36.3

The empirical results stress the efficiency of the proposed BAL sampling strategy with respect to the remaining strategies. In particular, we note the unexpected finding that BAL strategy has better performance in each partition with respect to the one-way optimal stratification design (OPT1 for partition 1 and OPT2 for partition 2). This may be explained by the fact that the BAL strategy exploits the auxiliary information, related to the domain membership, both at design and estimation phases; while the other strategy use these auxiliary variables just at the estimation phase.

Finally, in order to test the sensitivity of the BAL strategy, we examined the sampling variances of the domain estimates of the total of other variables (y_2, ..., y_{12}), reported in table 6.1.4, considering the sampling allocation based on y_1 and the modified greg estimator based on the assumption of the hetheroschedastic factor of the model (6.1.1).

Table 6.1.4. Description of the artificial variables of interest

Variable	Mean	Percent Coefficient of variation	Pearson Correlation with variables y_1	Pearson Correlation with variables x
$y_1 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5x$ *	20.27	111.00	1.00	0.85
$y_2 = 0.25x + \varepsilon;\ V(\varepsilon) = 1.5x$	14.46	122.96	0.99	0.75
$y_3 = 0.15x + \varepsilon;\ V(\varepsilon) = 1.5x$	8.65	167.83	0.52	0.60
$y_4 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5(x)^{3/2}$	20.13	204.34	0.42	0.49
$y_5 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5(x)^{1/3}$	20.33	101.57	0.85	0.99
$y_6 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5$	20.34	100.81	0.85	1.00
$y_7 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5\lvert y_1\rvert$	20.37	105.65	0.81	0.95
$y_8 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5\lvert y_2\rvert$	20.38	104.32	0.82	0.96
$y_9 = 0.35x + \varepsilon;\ V(\varepsilon) = 1.5\lvert y_3\rvert$	21.56	101.16	0.82	0.96
$y_{10} = 0.35y_1 + \varepsilon;\ V(\varepsilon) = 1.5\lvert y_1\rvert$	7.12	146.83	0.75	0.64
$y_{11} = 0.35y_2 + \varepsilon;\ V(\varepsilon) = 1.5\lvert y_2\rvert$	5.08	169.85	0.71	0.54
$y_{12} = 0.35y_3 + \varepsilon;\ V(\varepsilon) = 1.5\lvert y_3\rvert$	4.23	254.83	0.29	0.32
$x \sim N(30 + u_1^3;\ u_2^5)$	58.15	100.47	0.85	1.00

*$V(\varepsilon)$ *denotes the model variance*

The mean percent CV values are reported in table 6.1.5, 6.1.6 and 6.1.7 referred respectively to the domains of partition 1, 2 and to the overall set of domains.

Table 6.1.5. Mean percent CV by sampling design and variable of partition 1 domains

Variable	BAL	OPT1	OPT2	STDOM1	STDOM2	SRS	PPS
y_1	13,6	16,6	18,7	17,7	21,1	21,5	18,3
y_2	19,1	23,3	26,0	24,7	29,4	30,0	25,6
y_3	65,7	39,5	44,5	42,0	50,8	51,7	43,7
y_4	84,1	17,2	19,4	18,2	22,2	22,6	19,0
y_5	7,0	16,8	18,9	17,8	21,4	21,8	18,5
y_6	3,6	16,8	18,9	17,8	21,4	21,8	18,5
y_7	16,6	16,7	18,8	17,8	21,4	21,8	18,5
y_8	14,1	16,7	18,8	17,8	21,4	21,8	18,5
y_9	14,7	15,9	17,8	16,8	20,3	20,7	17,5
y_{10}	49,4	47,4	53,1	50,3	60,2	61,3	52,1
y_{11}	59,4	65,9	73,8	70,1	83,6	85,1	72,5
y_{12}	121,1	82,0	92,6	87,0	106,3	108,2	91,0

Table 6.1.6. Mean percent CV by sampling design and variable of partition 2 domains

Variable	BAL	OPT1	OPT2	STDOM1	STDOM2	SRS	PPS
y_1	20,0	28,4	24,0	35,4	25,3	33,0	26,6
y_2	28,2	40,3	34,0	50,4	35,8	46,9	37,7
y_3	90,7	66,8	56,5	83,5	59,4	78,1	62,6
y_4	123,9	29,4	24,7	37,0	26,0	34,6	27,5
y_5	9,3	27,8	23,6	34,3	24,8	32,1	26,0
y_6	4,9	27,7	23,6	34,3	24,8	32,1	26,0
y_7	22,6	27,9	23,7	34,5	24,9	32,3	26,1
y_8	19,2	27,9	23,7	34,5	24,9	32,3	26,1
y_9	20,4	26,6	22,5	33,0	23,7	30,8	24,9
y_{10}	72,1	83,2	70,0	104,8	73,6	97,6	77,9
y_{11}	88,6	119,4	99,9	151,3	105,2	140,8	111,6
y_{12}	178,7	154,6	128,3	198,6	135,0	185,3	144,9

Table 6.1.7. Mean percent CV by sampling design and variable of the overall set of domains

Variable	BAL	OPT1	OPT2	STDOM1	STDOM2	SRS	PPS
y_1	16,8	22,5	21,4	26,5	23,2	27,3	22,5
y_2	23,6	31,8	30,0	37,5	32,6	38,4	31,6
y_3	78,2	53,2	50,5	62,7	55,1	64,9	53,2
y_4	104,0	23,3	22,0	27,6	24,1	28,6	23,3
y_5	8,2	22,3	21,2	26,1	23,1	27,0	22,3
y_6	4,2	22,3	21,2	26,0	23,1	26,9	22,3
y_7	19,6	22,3	21,3	26,2	23,2	27,1	22,3
y_8	16,6	22,3	21,2	26,1	23,1	27,0	22,3
y_9	17,6	21,2	20,2	24,9	22,0	25,8	21,2
y_{10}	60,8	65,3	61,5	77,6	66,9	79,5	65,0
y_{11}	74,0	92,7	86,8	110,7	94,4	113,0	92,0
y_{12}	149,9	118,3	110,5	142,8	120,6	146,7	118,0

Examining the three tables we underline that the BAL strategy performs poorly for the variables 3, 4, 12 with respect to the other designs. In particular the BAL approach seems to be strongly influenced by the high variability of the variables of interest not considered in the sample design phase, while the other strategies seem to be less influenced by this problem. These findings point out that (*i*) the variables of interest with high variability must be included in the design phase; and (*ii*) the proposed strategy must be adopted after a careful analysis of the hetheroschedastic factors of the variables of interest.

6.2. Real business data

The experiment examines a situation characterizing many real survey contexts in which the overall sample size n is fixed and the marginal sample sizes n_{bd} are determined by a quite simple rule which turns out to be a compromise between the Allocation Proportional to Population size (APP) and the allocation uniform for each domain of a given partition:

$$n_{bd} = \alpha_b \; n(N_{bd} / N) + (1 - \alpha_b) n / M_b \qquad (6.2.1)$$

being α_b $(0 \leq \alpha_b \leq 1)$ a fixed constant.

The analysis has been carried out on the 1999 population of the enterprises from 1 to 99 employers belonging to the *Computer and related economic activities* (2-digits of the NACE rev.1 classification). In order to simplify the empirical analysis, some units with outlier values have been deleted. At the end of the cleaning procedure, 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. According to the EU *Council Regulation* n°58/97 on Structural Business Statistics the estimation domains are defined as different partition subsets of the population. In particular, 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. In order to determine the values of the parameters α_1 and α_2 of the expression (6.2.1), two simple one-way stratification designs have been taken into account: a sampling design stratified by Partition 1 with SRSWOR in each stratum (STDOM1) and a sampling design stratified by Partition 2 with SRSWOR in each stratum (STDOM2). The parameter α_1 and the related marginal sample sizes n_{1d} (*d*=1,…,20) guarantee that with the STDOM1 sample design the percent CV of the Horvitz-Thompson (HT) estimates of totals of the auxiliary variable *number of employers* be lower than than 34.5% for all domain of the partition 1; the α_2 value assures that with the STDOM2 sample design the percent CV of the HT estimates of totals of the auxiliary variable be lower than than 8.7% for all domain of the partition 2. We note that the above allocation rules are straightforward to implement in any real survey contexts. In the following we refer to the domains with the planned sample size greater than the APP sample size as *oversized* domains. These domains need to have a sample size larger than the APP sample size to bound the sampling errors; roughly speaking these domains may be classified as *small domains*. In the following the analysis is based on the set of small domains.

Given the marginal sample sizes, five sampling designs have been considered in the experiment, as reported in table 6.2.1.

Table 6.2.1. 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*

The inclusion probabilities of STDOM1 and STDOM2 are those described in the table 6.1.2. Two balanced sample designs are examined: 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 $\sum_{k\in s_{bd}} \pi_k\ {}_{bd}\delta_k \big/ \pi_k = n_{bd}$ and $\sum_{k\in s_{bd}} {}_{bd}\delta_k \big/ \pi_k = N_{bd}$ $(b=1,...,B; d=1,...,M_b)$. The probabilities π_k of both designs have been obtained as solution of the calibration problem (3.2.1) where the marginal sample sizes are computed by equation (6.2.1) and the initial probabilities π'_k are set uniformly equal to $\pi'_k = n/N$. These probabilities are no more *optimal* in the sense described in section 3; however they have been 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. The coordinated design CPAR selects a single sample for each marginal population with Pareto Sampling (Särndal and Lundström, 2005), assuring the maximum overlap of the two samples; the marginal sample sizes (6.2.1) are satisfied only as expectation over repeated sampling; 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 estimator* (*HT*), *modified greg estimator* (*greg*) and *synthetic estimator* (*syn*), expressed as ${}_{bd}\hat{t}_{r,syn} = \sum_{k\in U_{bd}} \tilde{y}_{r,k}$. As far as the estimators using auxiliary information are concerned, two simple homoschedastic linear models have been implemented: the model (6.2.2) 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 (6.2.3) uses the 44 domain membership indicator variables. The linear model (6.2.2) is expressed by

$$E_m(y_k) = \beta_h + \beta_j \qquad \text{for } k \in U_h \cap U_j\ , \tag{6.2.2}$$

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 (6.2.3) is

$$E_m(y_k) = \beta_{1d} + \beta_{2d} \qquad \text{for } k \in U_{1d} \cap U_{2d}\ , \tag{6.2.3}$$

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

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 (6.2.2) is somewhat more reliable, since the estimates of the regression parameters are based on large sample sizes; while in model (6.2.3) 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. Obviously, more flexible model formulations could be possible as described for instance in Lehtonen *et al.* (2005).

Using the model (6.2.3) the synthetic and the modified 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 2 and *est* assumes the categories *HT, syn*, and *greg* above indicated.

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

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

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

denoting with: *F* a specific subset of the marginal domains; *card*(*F*) the cardinality of *F*; ${}_{bd}\hat{t}^{i}_{r,est}(dis)$ 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 6.2.2. 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 (6.2.2)); the forth block presents the results of *syn* or *greg* estimators based on the 44 domain membership indicator variables (model (6.2.3)). 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.

Table 6.2.2. Average Absolute Relative Bias (ARB) and Relative Mean Square Error (RMSE) of small domain sampling strategies

Sampling Design	Value Added						Labour Cost					
	DOM1		DOM2		Overall		DOM1		DOM2		Overall	
	ARB	RMSE	ARB	RMSE	ARB	RMSE	ARB	RMSE	ARB	RMSE	ARB	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

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 domain 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 modified *greg* estimator in third block of the table 3 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 $\overline{\text{RMSE}}$.

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-multidomain case, is useful when the overall sample size is bounded. In these instances, the standard solution of 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 an indirect small area estimators.

The empirical analysis, implemented both on artificial and on real enterprise data has generally showed a good performance of the strategy which is proposed in this paper.

Computational feasibility is one of the main advantages of the solution which is proposed, since it allows to easily implement an overall small area strategy which jointly considers the design and the estimation phase, and allows to improve the efficiency of the direct domain estimators.

The results of the proposed strategy are 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 are different from the optimal ones. Furthermore, the robustness of the approach is confirmed by using different working superpopulation model in the estimation phase thus pointing out the adaptability of the approach to the complex survey contexts.

REFERENCES

Battese, G. E., Harter, R. M., Fuller W. A. (1988) An Error Component Model for Prediction of Counting Crop Areas Using Survey Data and Satellite Data, *Journal American Statistical Society*, **83**: 28-36.

Bethel, J. (1989) Sample Allocation in Multivariate Surveys, *Survey Methodology*, **15**: 47-57.

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

Bryant E. C., Hartley H. O., Jessen R. J. (1960) Design and Estimation in Two-Way Stratification, *Journal of the American Statistical Association*, **55**: 105-124.

Causey B. D., Cox L. H., Ernst L. R. (1985) Applications Transportation Theory to Statistical Problem, *Journal American Statistical Society*, **80**: 903-909.

Chauvet G., Tillé Y. (2006) A Fast Algorithm of Balanced Sampling, *Journal of Computational Statistics*.

Chromy J. (1987) Design Optimization with Multiple Objectives, *Proceedings of the Survey Research Methods Section*. American Statistical Association, 194-199.

Deville J.-C., 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., Tillé Y. (2004) Efficient Balanced Sampling: the Cube Method, *Biometrika*, **91**: 893-912.

Dykstra R. (1985) An Iterative Procedure for Obtaining I-Projections onto the Intersection of Convex Sets, *Annals of Probability*, **13**, 975-984.

Dykstra R., Wollan P. (1987) Finding I-Projections Subject to a Finite Set of Linear Inequality Constraints, *Applied Statistics*, **36**, **3**77-383.

Ernst L. R., Paben S. P. (2002) Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications, *Journal of Official Statistics*, **18**: 185-202.

Falorsi, P. D., Orsini, D., Righi, P., (2006) Balanced and Coordinated Sampling Designs for Small Domain Estimation, *Statistics in Transition*, **7**, 1173-1198.

Jessen R. J. (1970) Probability Sampling with Marginal Constraints, *Journal American Statistical Society*, **65**: 776-795.

Lehtonen R., Veijanen A. (1998) Logistic Generalized Regression Estimators, *Survey Methodology*, **24**: 51-55.

Lehtonen R., Särndal C. E, Veijanen A. (2003) The effect of Model Choice in Estimation for Domains, Including Small Domains, *Survey Methodology*, **1**: 33-44.

Lehtonen R., Särndal C. E., Veijanen A. (2005) Does the Model Matter? Comparing Model-Assisted and Model-Dependent Estimators of Class Frequencies for Domains, *Statistics In Transition*, **7**: 649-673.

Lu W., Sitter R. R. (2002) Multi-way Stratification by Linear Programming Made Practical, *Survey Methodology*, **2**: 199-207.

Montanari G. E., Ranalli M.G. (2003) Nonparametric Methods in Survey Sampling, in: *New Developments in Classification and Data Analysis*, Vinci M., Monari P., Mignani S., Montanari A. (eds), Springer , Berlin.

Ohlsson E. (1995) Coordination of Samples using Permanent Random Numbers, in: *Business Survey Methods*, Cox B. G., Binder D. A., Chinnappa B. N., Chirstianson A., Colledge M. J., Kott, P.S. (eds), Wiley, New York, 153-169.

Rao J. N. K. (2003) *Small Area Estimation*, Wiley, New York.

Rao J. N. K., Nigam A. K. (1990) Optimal Controlled Sampling Design, *Biometrika*, **77**: 807-814.

Rao J. N. K., Nigam A. K. (1992) Optimal Controlled Sampling: a Unifying Approach, *International Statistical Review*, **60**: 89-98.

Rosén B. (1997a) Asymptotic Theory for Order Sampling, *Journal of Statistical Planning and Inference*, **62**, 135-158.

Rosén B. (1997b) On Sampling with Probability Proportional to Size, *Journal of Statistical Planning and Inference*, **62**, 159-191.

Royall R., Herson J. (1973) Robust Estimation in Finite Population, *Journal American Statistical Society*, **68**: 880-889.

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

Singh M. P., Gambino J., Mantel H. J. (1994) Issues and Strategies for Small Area Data, *Survey Methodology*, **20**: 3-22.

Singh M. P., Mian I. U. H. (1995) Generalized Sample Size Dependent Estimators for Small Areas, *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, Washington , DC, 687-701.

Sitter R. R., Skinner C. J. (1994) Multi-way Stratification by Linear Programming, *Survey Methodology*, **20**: 65-73.

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

Appendix 1

In the special case of a single partition (b=B=1 and $M_b = Q$), under the following three conditions

(i) $N/(N-Q) \cong 1$, (ii), $c_k \approx {}_{bd}c \;\; \forall k \in U_{bd}$, and (iii) $\pi_k \approx n_{bd} / N_{bd} = f_{bd}$, $\forall k \in U_{bd}$ (A.1.1)

then

$$\frac{N}{N-Q}\sum_{k\in U}\left(\frac{1}{\pi_k}-1\right){}_{bd}^{\;a}\eta_{r,k}^2 = AV({}_{bd}\hat{t}_{r,greg}\,|\,\hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) < AV({}_{bd}\hat{t}_{r,greg}) = \sum_{k\in U}\left(\frac{1}{\pi_k}-1\right)c_k\,\sigma_r^2 \text{ .(A.1.2)}$$

In order to demonstrate (A.1.2), rank the N units by the order of domain. The matrices $\left(\mathbf{Z}'_U\mathbf{\Omega}_U^{-1}\mathbf{Z}_U\right)^{-1}$ and $\mathbf{Z}'_U\mathbf{\Omega}_U^{-1}$ become

$$\left(\mathbf{Z}'_U\mathbf{\Omega}_U^{-1}\mathbf{Z}_U\right)^{-1} = diag\left\{1\Big/\sum_{j\in U_{bd}}\pi_j(1-\pi_j)\right\}_{d=1}^{M_b},$$

$$\mathbf{Z}'_U\mathbf{\Omega}_U^{-1} = \begin{bmatrix} (1-\pi_1) & \ldots & (1-\pi_l) & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ \ldots & & & & & & & & & & \ldots \\ 0 & \ldots & 0 & (1-\pi_{l+1}) & \ldots & (1-\pi_k) & \ldots & (1-\pi_m) & 0 & \ldots & 0 \\ \ldots & & & & & & & & & & \ldots \\ 0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & (1-\pi_{m+1}) & \ldots & (1-\pi_N) \end{bmatrix}.$$

Then, if the unit k belongs to domain d, it is

$$\mathbf{g}'_k = \left(0,\ldots,0,\frac{\pi_k(1-\pi_{l+1})}{\sum\limits_{j\in U_{bd}}\pi_j(1-\pi_j)},\ldots,\frac{\pi_k(1-\pi_k)}{\sum\limits_{j\in U_{bd}}\pi_j(1-\pi_j)},\ldots,\frac{\pi_k(1-\pi_m)}{\sum\limits_{j\in U_{bd}}\pi_j(1-\pi_j)},0,\ldots,0\right),$$

where the units l+1, …,m, belong to domain d. The ${}_{bd}^{\;a}\eta_{r,k}^2$ quantities may be expressed as

$${}_{bd}^{\;a}\eta_{r,k}^2 = \begin{cases} c_k\,(1-2g_{kk}) + \sum\limits_{j\in U_{bd}} g_{kj}^2\,c_j & \text{if } k \in U_{bd} \\ 0 & \text{otherwise} \end{cases}.$$

Under the above results, it is trivial to note that, under conditions (i) of (A.1.1), the inequality (A.1.2) is realized if $c_k\,(1-2g_{kk}) + \sum\limits_{j\in U_d} g_{kj}^2\,c_j < c_k$.

Under the condition (ii) of (A.1.1), the above is $(1-2g_{kk}) + \sum\limits_{j\in U_d} g_{kj}^2 < 1$. By substituting

$$-2\frac{\pi_k(1-\pi_k)}{\sum\limits_{j\in U_{bd}}\pi_j(1-\pi_j)}+\sum_{j\in U_{bd}}\left[\frac{\pi_k(1-\pi_j)}{\sum\limits_{j\in U_{bd}}\pi_j(1-\pi_j)}\right]^2<0 \Rightarrow -2\pi_k(1-\pi_k)+\frac{\sum\limits_{j\in U_{bd}}\left[\pi_k(1-\pi_j)\right]^2}{\sum\limits_{j\in U_{bd}}\pi_j(1-\pi_j)}<0. \quad \text{(A.1.3)}$$

Under the hypothesis (iii) of (A.1.1), the (A.1.3) turns out to be

$$-2f_{bd}(1-f_{bd})+\frac{1}{\sum\limits_{j\in U_d}f_{bd}(1-f_{bd})}\sum_{j\in U_d}\left[f_{bd}(1-f_{bd})\right]^2<0$$

$$-2f_{bd}(1-f_{bd})+\frac{N_{bd}\left[f_{bd}(1-f_{bd})\right]^2}{N_{bd}\,f_{bd}(1-f_{bd})}<0$$

$$-2f_d(1-f_d)+f_d(1-f_d)<0.$$

The above shows that the selection of balanced sample reduces the anticipated variance.

Contributi ISTAT(*)

1/2004 – Marcello D'Orazio, Marco Di Zio e Mauro Scanu – *Statistical Matching and the Likelihood Principle: Uncertainty and Logical Constraints*

2/2004 – Giovanna Brancato – *Metodologie e stime dell'errore di risposta. Una sperimentazione di reintervista telefonica*

3/2004 – Franco Mostacci, Giuseppina Natale e Elisabetta Pugliese – *Gli indici dei prezzi al consumo per sub popolazioni*

4/2004 – Leonello Tronti – *Una proposta di metodo: osservazioni e raccomandazioni sulla definizione e la classificazione di alcune variabili attinenti al mercato del lavoro*

5/2004 – Ugo Guarnera – *Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi: il softaware* Quis

6/2004 – Patrizia Giaquinto, Marco Landriscina e Daniela Pagliuca – *La nuova funzione di analisi dei modelli implementata in Genesees v. 3.0*

7/2004 – Roberto Di Giuseppe, Patrizia Giaquinto e Daniela Pagliuca – *MAUSS (Multivariate Allocation of Units in Sampling Surveys): un software generalizzato per risolvere il problema dell' allocazione campionaria nelle indagini Istat*

8/2004 – Ennio Fortunato e Liana Verzicco – *Problemi di rilevazione e integrazione della condizione professionale nelle indagini sociali dell'Istat*

9/2004 – Claudio Pauselli e Claudia Rinaldelli – *La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute*

10/2004 – Eugenio Arcidiacono, Marina Briolini, Paolo Giuberti, Marco Ricci, Giovanni Sacchini e Giorgia Telloli – *Procedimenti giudiziari, reati, indagati e vittime in Emilia-Romagna nel 2002: un'analisi territoriale sulla base dei procedimenti iscritti nel sistema informativo Re.Ge.*

11/2004 – Enrico Grande e Orietta Luzi – *Regression trees in the context of imputation of item non-response: an experimental application on business data*

12/2004 – Luisa Frova e Marilena Pappagallo – *Procedura di now-cast dei dati di mortalità per causa*

13/2004 – Giorgio DellaRocca, Marco Di Zio, Orietta Luzi, Emanuela Scavalli e Giorgia Simeoni – *IDEA (Indices for Data Editing Assessment): sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*

14/2004 – Monica Pace, Silvia Bruzzone, Luisa Frova e Marilena Pappagallo – *Review of the existing information about death certification practices, certificate structures and training tools for certification of causes of death in Europe*

15/2004 – Elisa Berntsen – *Modello Unico di Dichiarazione ambientale: una fonte amministrativa per l'Archivio delle Unità Locali di Asia*

16/2004 – Salvatore F. Allegra e Alessandro La Rocca – *Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la definizione di un indice composto*

17/2004 – Francesca R. Pogelli – *Un'applicazione del modello "Country Product Dummy" per un'analisi territoriale dei prezzi*

18/2004 – Antonia Manzari – *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*

19/2004 – Claudio Pauselli – *Intensità di povertà relativa: stima dell'errore di campionamento e sua valutazione temporale*

20/2004 – Maria Dimitri, Ersilia Di Pietro, Alessandra Nuccitelli e Evelina Paluzzi – *Sperimentazione di una metodologia per il controllo della qualità di dati anagrafici*

21/2004 – Tiziana Pichiorri, Anna M. Sgamba e Valerio Papale – *Un modello di ottimizzazione per l'imputazione delle mancate risposte statistiche nell'indagine sui trasporti marittimi dell'Istat*

22/2004 – Diego Bellisai, Piero D. Falorsi, Annalisa Lucarelli, Maria A. Pennucci e Leonello G. Tronti – *Indagine pilota sulle retribuzioni di fatto nel pubblico impiego*

23/2004 – Lidia Brondi – *La riorganizzazione del sistema idrico: quadro normativo, delimitazione degli ambiti territoriali ottimali e analisi statistica delle loro caratteristiche strutturali*

24/2004 – Roberto Gismondi e Laura De Sandro – *Provisional Estimation of the Italian Monthly Retail Trade Index*

25/2004 – Annamaria Urbano, Claudia Brunini e Alessandra Chessa – *I minori in stato di abbandono: analisi del fenomeno e studio di una nuova prospettiva d'indagine*

26/2004 – Paola Anzini e Anna Ciammola – *La destagionalizzazione degli indici della produzione industriale: un confronto tra approccio diretto e indiretto*

27/2004 – Alessandro La Rocca – *Analisi della struttura settoriale dell'occupazione regionale: 8° Censimento dell'industria e dei servizi 2001 7° Censimento dell'industria e dei servizi 1991*

28/2004 – Vincenzo Spinelli e Massimiliano Tancioni – *I Trattamenti Monetari non Pensionistici: approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10*

29/2004 – Paolo Consolini – *L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie*

1/2005 – Fabrizio M. Arosio – *La stampa periodica e l'informazione on-line: risultati dell'indagine pilota sui quotidiani on-line*

2/2005 – Marco Di Zio, Ugo Guarnera e Orietta Luzi – *Improving the effectiveness of a probabilistic editing strategy for business data*

3/2005 – Diego Moretti e Claudia Rinaldelli – *EU-SILC complex indicators: the implementation of variance estimation*

4/2005 – Fabio Bacchini, Roberto Iannaccone e Edoardo Otranto – *L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione*

5/2005 – Marco Broccoli – *Analisi della criminalità a livello comunale: metodologie innovative*

6/2005 – Claudia De Vitiis, Loredana Di Consiglio e Stefano Falorsi – *Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro*

7/2005 – Edoardo Otranto e Roberto Iannaccone – *Continuous Time Models to Extract a Signal in Presence of Irregular Surveys*

(*) ultimi cinque anni

8/2005 – Cosima Mero e Adriano Pareto – *Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie*

9/2005 – Filippo Oropallo – *Enterprise microsimulation models and data challenges*

10/2005 – Marcello D' Orazio, Marco Di Zio e Mauro Scanu – *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*

11/2005 – Stefania Macchia, Manuela Murgia, Loredana Mazza, Giorgia Simeoni, Francesca Di Patrizio, Valentino Parisi, Roberto Petrillo e Paola Ungaro – *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*

12/2005 – Piero D. Falorsi, Monica Scannapieco, Antonia Boggia e Antonio Pavone – *Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione*

13/2005 – Ciro Baldi, Francesca Ceccato, Silvia Pacini e Donatella Tuzi – *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*

14/2005 – Stefano De Francisci, Giuseppe Sindoni e Leonardo Tininini – *Da Winci/MD: un sistema per data warehouse statistici sul Web*

15/2005 – Gerardo Gallo e Evelina Palazzi – *I cittadini italiani naturalizzati: l'analisi dei dati censuari del 2001, con un confronto tra immigrati di prima e seconda generazione*

16/2005 – Saverio Gazzelloni, Mario Albisinni, Lorenzo Bagatta, Claudio Ceccarelli, Luciana Quattrociocchi, Rita Ranaldi e Antonio Toma – *La nuova rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*

17/2005 – Maria Carla Congia – *Il lavoro degli extracomunitari nelle imprese italiane e la regolarizzazione del 2002. Prime evidenze empiriche dai dati INPS*

18/2005 – Giovanni Bottazzi, Patrizia Cella, Giuseppe Garofalo, Paolo Misso, Mariano Porcu e Marianna Tosi – *Indagine pilota sulla nuova imprenditorialità nella Regione Sardegna. Relazione Conclusiva*

19/2005 – Fabrizio Martire e Donatella Zindato – *Le famiglie straniere: analisi dei dati censuari del 2001 sui cittadini stranieri residenti*

20/2005 – Ennio Fortunato – *Il Sistema di Indicatori Territoriali: percorso di progetto, prospettive di sviluppo e integrazione con i processi di produzione statistica*

21/2005 – Antonella Baldassarini e Danilo Birardi – *I conti economici trimestrali: un approccio alla stima dell'input di lavoro*

22/2005 – Francesco Rizzo, Dario Camol e Laura Vignola – *Uso di XML e WEB Services per l'integrazione di sistemi informativi statistici attraverso lo standard SDMX*

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

1/2007 – Paolo Roberti, Maria Grazia Calza, Filippo Oropallo e Stefania Rossetti – *Knowledge Databases to Support Policy Impact Analysis: the EuroKy-PIA Project*

2/2007 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, e Marina Sorrentino – *Production of job vacancy statistics: coverage*

3/2007 – Carlo Lucarelli e Giampiero Ricci – *Working times and working schedules: the framework emerging from the new Italian lfs in a gender perspective*

4/2007 – Monica Scannapieco, Diego Zardetto e Giulio Barcaroli – *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*

5/2007 – Giulio Barcaroli e Tiziana Pellicciotti – *Strumenti per la documentazione e diffusione dei microdati d'indagine: il Microdata Management Toolkit*

6/2007 – AA.VV. – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 1ª giornata*

7/2007 – Raffaella Cianchetta, Carlo De Gregorio, Giovanni Seri e Giulio Barcaroli – *Rilevazione sulle Pubblicazioni Scientifiche Istat*

8/2007 – Emilia Arcaleni, e Barbara Baldazzi – *Vivere non insieme: approcci conoscitivi al Living Apart Together*

9/2007 – Corrado Peperoni e Francesca Tuzi – *Trattamenti monetari non pensionistici metodologia sperimentale per la stima degli assegni al nucleo familiare*

10/2007 – AA.VV – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 2ª giornata*

11/2007 – Leonello Tronti – *Il prototipo (numero 0) dell'Annuario di statistiche del Mercato del Lavoro (AML)*

12/2007 – Daniele Frongia, Raffaello Martinelli, Fernanda Panizon, Bruno Querini e Andrea Stanco – *Il nuovo Sistema informatico Altri Servizi. Progetto di reingegnerizzazione dei processi produttivi delle indagini trimestrali di fatturato degli altri servizi*

1/2008 – Carlo De Gregorio, Stefania Fatello, Rosanna Lo Conte, Stefano Mosca, Francesca Rossetti – *Sampling design and treatment of products in Istat centralised CPI surveys*

2/2008 – Mario Abissini, Elisa Marzilli e Federica Pintaldi – *Test cognitivo e utilizzo del questionario tradotto: sperimentazioni dell'indagine sulle forze di lavoro*

3/2008 – Franco Mostacci – *Gli aggiustamenti di qualità negli indici dei prezzi al consumo in Italia: metodi, casi di studio e indicatori impliciti*

4/2008 – Carlo Vaccari e Daniele Frongia – *Introduzione al Web 2.0 per la Statistica*

5/2008 – Antonio Cortese – *La conta degli stranieri: una bella sfida per il censimento demografico del 2011*

6/2008 – Carlo De Gregorio, Carmina Munzi e Paola Zavagnini – *Problemi di stima, effetti stagionali e politiche di prezzo in alcuni servizi di alloggio complementari: alcune evidenze dalle rilevazioni centralizzate dei prezzi al consumo*

7/2008 – AA.VV. – *Seminario: metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali*

8/2008 – Monica Montella – *La nuova matrice dei margini di trasporto*

9/2008 – Antonia Boggia, Marco Fortini, Matteo Mazziotta, Alessandro Pallara, Antonio Pavone, Federico Polidoro, Rosabel Ricci, Anna Maria Sgamba e Angela Seeber – *L'indagine conoscitiva della rete di rilevazione dei prezzi al consumo*

10/2008 – Marco Ballin e Giulio Barcaroli – *Optimal stratification of sampling frames in a multivariate and multidomain sample design*

11/2008 – Grazia Di Bella e Stefania Macchia – *Experimenting Data Capturing Techniques for Water Statistics*

12/2008 – Piero Demetrio Falorsi e Paolo Righi – *A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation*