Joint Use of Balanced Sampling and Calibration for Multivariate and Multi-Domain Sample Designs

L’uso congiunto del campionamento bilanciato e della calibrazione per la pianificazione di disegni campionari multivariati e multi-dominio

Claudia De Vitiis, Paolo Righi, Tiziana Tuoto
Italian National Statistical Institute
e-mail: devitiis@istat.it

Riassunto: Molte indagini campionarie hanno come obiettivo la produzione di stime per una pluralità di parametri e per diversi domini pianificati, definiti da partizioni non annidate della popolazione. Un approccio standard è quello di progettare un disegno stratificato in cui gli strati sono ottenuti combinando le diverse partizioni della popolazione; ciò può introdurre delle inefficienze in termini di dimensione complessiva del campione. Nel lavoro si affronta la pianificazione dei domini secondo un disegno stratificato a più vie, superando la stratificazione incrociata ma garantendo che l’errore campionario delle stime di interesse sia inferiore a valori soglia prefissati. Il problema è qui ricondotto a un caso particolare di campionamento bilanciato, ed è proposta un’espressione della varianza che tiene conto sia del modello di regressione dello stimatore GREG, sia della tecnica di selezione basata sul bilanciamento.

Keywords: planning sampling size, controlled selection, multiway stratification

1. Introduction

When planning a sample strategy for a survey aiming at producing estimates for several domains, defined as non-nested partitions of the overall population, a widespread issue is to define the sample size so that the sampling errors of domain estimates of several parameters are lower than given thresholds.

The present work illustrates a sampling strategy dealing with multivariate-multidomain surveys when the overall sample size must satisfy budget constraints; in these cases the standard solution of a stratification given by cross-classification of the domain variables is not feasible because the number of strata can be larger than the overall sample size. Moreover, even if the overall sample size allows covering all the strata, the resulting allocation could lead to an inefficient design.

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. Several more sophisticated solutions have been proposed 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 main weaknesses of these approaches derive from the
computational complexity and moreover a solution is not always reached. These drawbacks have limited the use of multi-way stratification techniques as a standard solution for planning the survey sampling.

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 overall sampling strategy exploits the auxiliary information both in the sample selection and in the estimator definition. In particular, the strategy considers an algorithm for selecting a balanced sample (Deville, Tillé, 2004), using as auxiliary information that related to the domain definition, in order to control the sample size in each domain; in the estimation phase, the auxiliary information is exploited through a GREG-type estimator which allows one to obtain more efficient estimates. In the sample allocation procedure, a variance expression is utilized taking jointly into account both the regression model motivating GREG estimator and the proposed balanced sampling design. Furthermore, with respect to other solutions proposed to face with multi-domain surveys, a relevant advantage of the approach is due to the computational feasibility, since it allows to easily implement an overall strategy which jointly considers the design and the estimation phase and allows to improve the efficiency of the domain estimators.

The paper is organised as follows. Section 2 describes the overall sampling strategy. Section 3 shows the iterative algorithm that defines the inclusion probabilities and the corresponding planned domain sample sizes. In sections 4 some application fields are pointed up and further developments are underlined.

2. The sampling strategy

In order to define the problem formally, let us denote with $U$ a population of $N$ elements and with $b$ a specific partition of $U$ ($b=1,\ldots,B$) in which $b$-th partition defines $M_b$ different non overlapping domains $U_{bd}$ ($d=1,\ldots,M_b$) 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. Let $y_{r,k}$ and $bd\delta_k$ denote respectively the value of the $r$-th ($r=1,\ldots,R$) 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

$$bd t_r = \sum_{kd} y_{r,k} bd\delta_k = \sum_{kd/bd} y_{r,k} (r=1,\ldots,R; b=1,\ldots,B; d=1,\ldots,M_b). \quad (1)$$

The expression (1) defines a multivariate-multidomain problem.

In order to estimate the $bd t_r$ parameters, a sample $s$ of fixed size $n$ is selected from population $U$ with inclusion probabilities $\pi_k$ ($\pi_k > 0, k \in U$). The sample is selected by a multi-way stratification technique developed under the balanced sampling framework.
guaranteeing, in general, that the selected sample respects the following balancing equations
\( \hat{t}_{z,ht} = t_z \), where \( \hat{t}_{z,ht} = \sum_{k \in U} z_k \lambda_k a_k \) denote the Horvitz-Thompson estimates of \( t_z = \sum_{k \in U} z_k \), being \( z_k \) a vector of auxiliary variables known for each population unit and \( a_k = 1/\pi_k \) and \( \lambda_k = 1 \) if \( k \in s \) and \( \lambda_k = 0 \) otherwise. By defining the vector \( z_k \) as \((Q-B)\) zeros and \( B \) entries equal to \( \pi_k \) for the domains which the unit \( k \) belongs to, i.e.

\[
\mathbf{z}_k^* = (0, \ldots, \pi_k, \ldots, 0, \ldots, 0) = \pi_k (\delta_k, 1, \ldots, \delta_k, 1, \ldots, \delta_k, 1),
\]

the balancing equations assure that for each selected sample \( s \), the size \( n_{bd} \) of the subsample \( s_{bd} = s \cap U_{bd} \) is a non-random quantity and \( n_{bd} = \sum_{k \in U_{bd}} \pi_k \).

One relevant drawback of balanced sampling has always been implementing a general procedure giving a multivariate balanced random sample. 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. Furthermore, Deville and Tillé (2000) show that with specification (2) of the \( z_k \) vectors, the balancing equations can be exactly satisfied. A free SAS macro for the selection of balanced samples for large data sets may be downloaded in the website:

The estimates of \( \hat{b}_r \) can be obtained through the usual GREG estimator (Särndal et al. 1992) or with the modified GREG estimator described by Rao (2003, page 20). In this paper we will consider the latter, denoted with \( \hat{b}_r \), given by

\[
\hat{b}_r y_{r,k} = \sum_{k \in s} b_d w_k y_{r,k},
\]

where:

\[
b_d w_k = a_k b_d \delta_k + (b_d f_s - b_d \hat{t}_{s,ht})' \left( \sum_{k \in s} a_k x_k x_k' / c_k \right)^{-1} a_k x_k / c_k
\]
denotes the sampling weight, \( x_k \) indicates a vector of auxiliary variables, \( c_k \) is a known constant chosen in an appropriate way, \( b_d \hat{t}_s = \sum_{k \in U_{bd}} x_k \) and \( b_d \hat{t}_{s,ht} = \sum_{k \in U_{bd}} x_k a_k \). The estimator (3) is derived under the following working superpopulation model

\[
y_{r,k} = x_k' \beta_r + \varepsilon_{r,k}
\]
where $\beta$ 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 assumed to be: $E_{m}(r,\varepsilon_{k}) = 0$; $V_{m}(\varepsilon_{r,k}) = \sigma^{2}_{r}$; $E_{m}(\varepsilon_{r,k},\varepsilon_{r,i}) = 0$ if $k \neq i$.

Let us observe that the linear model (4) allows us to define the estimator only knowing the domain totals of the auxiliary information and the $x_{k}$ values for the sampling units. However, knowing the $x_{k}$ values for every $k \in U$, it is possible to build an estimator with more efficient predictions $\gamma_{r,k}$ obtained by generalised linear models (Lehtonen and Veijanen, 1998) or non parametric regression techniques (Montanari and Ranalli, 2003).

As noted by Rao (2003), the modified GREG estimator (3) is approximately design unbiased as the overall sample size increases, even if the domain sample size $n_{bd}$ is small.

It is useful to note that the sum of the $bd \hat{t}_{r,greg}$ estimates over all the domains of a partition is equal to the usual GREG estimate of the overall total $t_{r} = \sum_{d=1}^{M} bd \hat{t}_{r}$. This feature allows one to produce the estimates through a mixed strategy: the GREG estimator providing a unique weight for obtaining the survey estimates at the whole population level and the modified GREG estimator for the domain level estimates.

The approximated sampling variance of the modified GREG estimator under balanced sampling is, given a suitable asymptotic framework,

$$V_{p}(bd \hat{t}_{r,greg} | \hat{t}_{z,ht} = t_{z}) = \frac{N}{N - Q} \sum_{k \in \bar{U}} \left(1 - \frac{1}{\pi_{k}}\right) bd \hat{\eta}_{r,k}^{2} ,$$

being

$$bd \hat{\eta}_{r,k} = \begin{cases} 
\varepsilon_{r,k} - z'_{k} \ BD_{z,e} & \text{for } k \in U_{bd} \\
- z'_{k} \ BD_{z,e} & \text{for } k \in U_{\bar{U}} 
\end{cases} ,$$

with

$$bd \ BD_{z,e} = \left( \sum_{k \in U} z_{k} z'_{k} \left( \frac{1}{\pi_{k}} - 1 \right) \right)^{-1} \sum_{k \in U} z_{k} \varepsilon_{r,k} \ bd \delta_{k} \left( \frac{1}{\pi_{k}} - 1 \right) ,$$

where $U_{\bar{U}}$ is the subset of $U$ complementary to $U_{bd}$.

To prove (5), let us consider 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)

$$bd \hat{t}_{r,greg} \cong bd \hat{t}_{r,greg}^{*} = \sum_{k \in U_{bd}} x'_{k} \beta_{r} + \sum_{k \in s_{bd}} a_{k} \varepsilon_{r,k} = \sum_{k \in U_{bd}} x'_{k} \beta_{r} + \sum_{k \in s} a_{k} \varepsilon_{r,k} \ bd \delta_{k} .$$
Hence, generalizing the result given in expression (7) of Deville and Tillé (2004), which takes into account the Horvitz-Thompson estimator, we have

\[ V_p(b_d \hat{t}_{r,greg} \mid \hat{t}_{z,ht} = t_z) \approx V_p \left( \sum_{k \in U_{bd}} x_k^t \beta_r + \sum_{k \in \varepsilon} a_k \epsilon_{r,k} \delta_k \mid \hat{t}_{z,ht} = t_z \right) = \]

\[ = V_p \left( \sum_{k \in \varepsilon} a_k \epsilon_{r,k} \delta_k + (t_z - \hat{t}_{z,ht}) \gamma_{b_d} B_{z,c} \right) = \]

\[ = V_p \left( \sum_{k \in \varepsilon} a_k \beta_{k,bd} \eta_{r,k} \right) = \frac{N}{N - Q_{k \in U}} \left( \frac{1}{\pi_k} - 1 \right) b_d \eta_{r,k}^2. \]

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

The inclusion probabilities \( \pi_k \) 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 \( V_p(b_d \hat{t}_{r,greg} \mid \hat{t}_{z,ht} = t_z) \) are lower than prefixed precision thresholds \( b_d \Psi_r \):

\[ V_p(b_d \hat{t}_{r,greg} \mid \hat{t}_{z,ht} = t_z) \leq b_d \Psi_r \quad (b = 1, \ldots, B; \quad d = 1, \ldots, M_b; \quad r = 1, \ldots, R) \]

The algorithm is briefly described in the next section.

### 3. Algorithm to determine the sample sizes

The inclusion probabilities \( \pi_k \) and the corresponding domain sample sizes \( n_{bd} = \sum_{k \in U_{bd}} \pi_k \) are defined through an iterative algorithm that solves a non linear problem where the objective function to be minimized is the overall sample size, under the constraints on the sampling errors of domain estimates. The procedure is subdivided into two phases: (i) in the first one, denoted as **optimization**, the preliminary inclusion probabilities \( \pi_k' \) are determined solving a minimum constrained problem; (ii) in the second one, denoted as **calibration**, the final inclusion probabilities \( \pi_k \) are obtained through a slight modification of the \( \pi_k' \); the calibration problem is implemented to assure that the domain sample sizes \( n_{bd} \) are integers.

The \( \pi_k \) values may be expressed as implicit functions of the unknown residuals \( b_d \eta_{r,k}^2 \); in real survey context, the calculus of the inclusion probabilities \( \pi_k \) may be done using some predictions \( b_d \eta_{r,k}^2 \). This is a general problem concerning the sample design phase, when the variances are generally unknown quantities that have to
be suitably estimated. In Tillé and Favre (2005) is given a criterion for obtaining a prediction $\hat{\eta}_{r,k}$ of the $\eta_{r,k}$ values, useful in repeated sampling contexts.

**Step 1- Optimization**

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

$$
\begin{align*}
\text{Min} \left( \sum_{k \in U} \pi_k' \right) \\
\frac{N}{N - Q} \sum_{k \in U} \left( \frac{1}{\pi_k} - 1 \right)_{bd} \eta_{r,k}^2 \leq \eta_F (b = 1, \ldots, B; d = 1, \ldots, M_b; r = 1, \ldots, R) \\
0 < \pi_k' \leq 1 \quad (k = 1, \ldots, N)
\end{align*}
$$

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

Bethel (1989) invoked the Khun-Tucker theorem to show that a solution exists and described a simple algorithm but computationally complex. Then numerical solution is obtained with a slight modification of the Chromy’s algorithm (Chromy, 1987), described in Falorsi et al. (1998). Usually, this algorithm is able to solve the problem guaranteeing that the inequalities $0 < \pi_k' \leq 1 \ (k = 1, \ldots, N)$ are respected, even though a proof of the convergence does not exist.

**Step 2- Calibration**

The quantities $n_{bd}$ are firstly defined by rounding the results of the $Q$ sums $\sum_{k \in U_{bd}} \pi_k' \ (b = 1, \ldots, B; \ d = 1, \ldots, M_b)$, replacing $n_{bd} = 1$ when the rounding gives $n_{bd} = 0$. Then the overall sample size $n$ is obtained by summing the $n_{bd}$'s ($b = 1, \ldots, B; \ d = 1, \ldots, M_b$). The probabilities $\pi_k$ are then obtained as solution of the standard calibration problem where the objective function is $\text{Min} \sum_{k \in U} G(\pi_k; \pi_k')$, under the constraints $\sum_{k \in U} \pi_k = n$ and $\sum_{k \in U_{bd}} \pi_k = n_{bd}$. The $G(\pi_k; \pi_k')$ is the distance function between $\pi_k$ and $\pi_k'$. Note that the problem above may be solved by the well known iterative methods.
4. Application fields and further developments

The methodology proposed in this paper deals with sample frameworks typical of both business and social-demographic surveys, when the parameters of interest refers to estimation domains defined by several different partition subsets of the population. For instance, in the business survey context, the Italian National Statistical Institute (ISTAT) carries out the *Structural Business Statistics Survey*: the parameters of interest refer to estimation domains defined by three different partition subsets of the population of enterprises: economic activity classes, economic activity groups by number of employees classes, economic activity divisions by Regions. This leads to an overall number of estimation domains roughly equal to 1,820 and the number of non-empty strata of the cross-classification design results larger than 37,000. The proposed approach allows to introduce a more flexible strategy removing the rigidity of the cross-classification design, obtaining a large reduction of the overall sample size needed to meet the contraints on the sampling errors of the estimates, as shown by some empirical applications.

In the social context, an example of an useful application of the proposed method regards the *Italian Graduates’ Career Survey*. This survey disseminates estimates on the employment status of graduates three years after the degree at level of several types of domains: gender, course, group of course, university center, defining an overall number of domains greater than 600. The usual sample design considers the cross-classification of all the variables defining the domains, obtaining more than 2,200 strata. Allocating sample over such an amount of strata leads to an inefficient sampling design. Actually, through the standard cross-classification design, the sample size needed to guarantee that all the requested estimates meet the constraints in terms of standard error is very large. The proposed approach allows controlling the sample size in a more efficient way, as verified in an empirical study.

A relevant extension of the proposed method regards the small area estimation issue. 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. 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. As shown in Falorsi *et al.* (2006), the sampling strategy here proposed may be easily extended to a mixed small area estimation strategy employing the use of both direct and indirect estimators through a modification of the $b_d \eta_{r,k}$ values.

Finally, we stress that some empirical analyses have shown that the proposed strategy is 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 complex survey contexts.
References