

Sampling Strategies for the Estimation of Measures of Competitiveness for Small Subpopulations of Enterprises

Strategie campionarie per la stima di misure di competitività per piccole sottopopolazioni di imprese

Paolo Righi

ISTAT, Via Magenta 4 00185 Roma, parighi@istat.it

Riassunto: Il presente lavoro analizza il processo di stima campionaria di indicatori di produttività per piccole sottopopolazioni di imprese definite su differenti partizioni della popolazione di riferimento. Il problema è affrontato ponendo particolare attenzione alla fase di pianificazione del disegno campionario proponendo l'adozione di tecniche di campionamento bilanciato e coordinato. Il principale vantaggio che determina l'adozione di tali tecniche è quello di poter realizzare una strategia campionaria, implementabile dal punto di vista computazionale, che tenga conto della fase di pianificazione del disegno e dello stimatore permettendo di migliorare l'efficienza delle stime. Le proprietà dei metodi proposti sono esaminate attraverso una sperimentazione basata su una popolazione reale di imprese.

Keywords: small domain estimate, balanced sampling, sample coordination

1. Introduction

Competitiveness and economic development is, in general, focused on the macroeconomic, political, legal and social circumstances that support a successful economy. However, these broader conditions are necessary but not sufficient to a healthy economy. They provide the opportunities to attain development and wealth but the last ones are actually created at a microeconomic level: therefore, an economy cannot be competitive unless operating companies are competitive (Porter, 2003).

The competitiveness of the enterprises of an economic system may be measured by means of the productivity, commonly defined as a ratio of a volume measure of output to a volume measure of input use (OECD, 2001), since it is one of the main factors of competitiveness.

Usually, these measures may be justly estimated through a sample survey that observes a sample of enterprises working in an economy system such as a nation. In this context, the paper deals with some sampling methodological aspects on the estimate of the productivity for subpopulations (or domains) of interest; for instance, considering the subpopulations of enterprises at geographic sub-area level or taking into account other kinds of subpopulations such as the enterprises operating in the specific economic activity sectors. In such a way, it is possible to study different competitive behaviours contributing in understanding the state of the competitiveness of the economic system as a whole.

Hence, the objectives of the survey sample are often to produce estimates for two or more partitions of the enterprises' population and, consequently, there will be a lot of domains of estimate.

When the overall sample size is relatively small with respect to the number of domains we have to deal with small area estimation problems because the sample size for some domains cannot guarantee reliable estimates by means of the direct estimators.

The small area problem is commonly thought of as one to be dealt with via estimation. However, if the domain membership indicator variables are available for each unit in the population there are opportunities to be exploited at the survey design stage to face this problem. 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 each domain is treated at design stage as a planned domain, for which it is possible to produce direct estimates with a prefixed level of precision. For small domains, it will often be needed to use indirect estimators (Rao, 2003), such as synthetic estimators, for producing reliable estimates; nevertheless, synthetic estimators are very sensitive to the assumptions of the model. Even a small departure from these assumptions can make the design bias high, inflating the mean square error of the estimates. In order to overcome this problem, it is decisive to base the indirect domain estimation on the sample units belonging to the small domain (Lehtonen *et al.*, 2003). A standard solution to obtain planned sample sizes for the domains of interest is to use a stratified sample with the strata defined by the cross-classification of variables that define the different partitions, hereinafter referred to as *cross-classification design*. In many practical situations, this design is unfeasible since it needs the selection of at least a number of sampling units as large as the product between the number of categories of the stratification variables that can produce a large overall sample size. To reduce the number of sample units, an easy strategy is to drop one or more stratifying variables or to group some of the categories so that some planned domains become unplanned with small or null sample size.

In the literature, many methods to keep under control the sample size in all the categories for all the stratifying variables, without using cross-classification design, have been proposed. These methods are, generally, referred to as *multi-way stratification techniques* (for a detailed review see Falorsi *et al.*, 2006) and they are based on a linear programming approach computationally complex that has limited the use of them as a standard solution in planning the sampling design.

This paper illustrates that two well known sampling techniques may be useful in controlling the sample size in all the domains of interest without suffering from the disadvantages of the above methods.

The first technique faces the problem by means of the selection of a *balanced sample* based on domain membership indicator variables (Deville, Tillé, 2004). The second technique deals with the problem by means of procedures developed to guarantee the *sample coordination* (Ohlsson, 1995).

The paper is organised as follows: Section 2 introduces the essential notation, Section 3 is dedicated to the balanced sampling, Section 4 focuses on coordination techniques, Section 5 deals with estimation topics and Section 6 shows the main results of a simulation study conducted on a real population of Italian enterprises. In particular, some productivity measures have been estimated. Some brief conclusions are finally underlined in Section 7.

2. Notation and parameters of interest

In order to simplify the notation, the paper will be restricted to the case of two stratifying variables. However, the generalisation to the multi-way case is straightforward.

Let us define with U the population of interest of size N . Suppose that two auxiliary variables $\mathbf{V}_1$ and $\mathbf{V}_2$ are known for each unit k of the population ($k=1, \dots, N$) and that respectively assume R and C modalities. In particular, the generic value or class of values assumed by $\mathbf{V}_1$ and $\mathbf{V}_2$ is denoted, respectively with ${}^r v_1$ and ${}^c v_2$. Let us specify with $U_{rc} = U_{r.} \cap U_{.c}$ the subpopulation of size

$N_{rc} \geq 0$, being $U_{r.}$ and $U_{.c}$ the subpopulations of size $N_{r.} = \sum_c N_{rc}$ and $N_{.c} = \sum_r N_{rc}$ for which $V_1 = {}^r v_1$ and $V_2 = {}^c v_2$. Given two target variables, let us define respectively with y_{jk} and y_{lk} the observed values on the unit k . There are $R + C + 1$ parameters of interest defined by,

$$R = \sum_{k \in U} y_{lk} / \sum_{k \in U} y_{jk} ; \quad R_{r.} = \sum_{k \in U_{r.}} y_{lk} / \sum_{k \in U_{r.}} y_{jk} ; \quad R_{.c} = \sum_{k \in U_{.c}} y_{lk} / \sum_{k \in U_{.c}} y_{jk} .$$

In order to obtain the sample estimates of the parameters of interest, the standard cross-classification designs select a stratified sample s of size n from the population U with probability $p(s)$, where the strata are obtained combining the categories of the variables $\mathbf{V}_1$ and $\mathbf{V}_2$. Let s_{rc} ($r = 1, \dots, R, c = 1, \dots, C$) be the subsample of s of size n_{rc} , satisfying $\min(b, N_{rc}) \leq n_{rc} \leq N_{rc}$, selected from population U_{rc} , being b the minimum sample size fixed in advance, where b is equal to 1 for producing unbiased estimates of the parameters of interest or equal to 2 for being able to compute unbiased variance estimates. Furthermore, $\mathbf{n} = \{n_{rc}\}$ is the set of the strata frequencies, while $\mathbf{n}_1 = (n_{1.}, \dots, n_{r.}, \dots, n_{R.})'$ and $\mathbf{n}_2 = (n_{.1}, \dots, n_{.c}, \dots, n_{.C})'$ are the marginal distributions of $\mathbf{n}$ with respect to $\mathbf{V}_1$ and $\mathbf{V}_2$, being $n_{r.} = \sum_c n_{rc}$ and $n_{.c} = \sum_r n_{rc}$. We will call *marginal stratification* with respect to $\mathbf{V}_1$ the partition of U in $U_{1.}, \dots, U_{r.}, \dots, U_{R.}$, so that $n_{r.}$ is the sample size of subsample $s_{r.} = \bigcup_{c=1}^C s_{rc}$ of the generic marginal strata r ; while we will call *marginal stratification* with respect to $\mathbf{V}_2$ the partition of U in $U_{.1}, \dots, U_{.c}, \dots, U_{.C}$, so that $n_{.c}$ is the sample size of subsample $s_{.c} = \bigcup_{r=1}^R s_{rc}$ of the generic marginal strata c .

3. Methods based on balanced sampling

Multi-way stratification designs can be dealt with by means of balanced sampling.

The definition of a balanced sample depends on the assumed inferential framework. In the model based approach, a sample is defined as balanced on a set of auxiliary variables if there is the equality between the sample and the known population means of the auxiliary variables (Valliant *et al.*, 2000). Following the design based approach, considered in this paper, a sample is balanced when the Horvitz-Thompson estimators of the totals of the auxiliary variables are equal to their known population totals (Deville and Tillé, 2004). Balanced sampling, in this last case, represents a generalisation of probability sampling designs exploiting auxiliary information known for each unit of the population. Standard designs like simple design with fixed size, stratified sampling and unequal probability sampling represent some specific applications of balanced designs. The multi-way stratification designs belong to this class of designs as well.

3.1. General introduction to balanced sampling

A sampling design is balanced with respect to a set of auxiliary variables available in the planning phase, if and only if, given a vector of inclusion probabilities $\boldsymbol{\pi}' = (\pi_1, \dots, \pi_k, \dots, \pi_N)'$, it satisfies the balancing equations given by

$$\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \lambda_k = \sum_{k \in U} \mathbf{x}_k, \quad (3.1)$$

being $\lambda_k = 1$ if unit k belongs to the sample and $\lambda_k = 0$ otherwise and where $\mathbf{x}_k = (x_{1k}, \dots, x_{hk}, \dots, x_{Hk})'$ is the vector of the observed values of the auxiliary variables for the unit k . Then, denoting with S set of all the subset S , and with λ the random variable that takes the values $\lambda = (\lambda_1, \dots, \lambda_k, \dots, \lambda_N)'$, the sampling designs considered assign a probability $p(s)$ to each sample s such that $E(s) = \sum_{s \in S} p(s) \lambda = \pi$.

Deville and Tillé (2004) proposed an algorithm that allows the selection of balanced samples (or approximately balanced) for a large set of auxiliary variables and with respect to different vectors of inclusion probabilities.

The algorithm, called *cube method*, is based on, at most, two phases. In the first phase, namely the *flight phase*, the cube method searches for a sampling design satisfying equations (3.1). The sample selection is performed by means of rounding off randomly to 1 or 0 almost all the elements of the vector π . The procedure is implemented iteratively and in each step the equations (3.1) are satisfied.

At the end of the flight phase the elements of π equal to 1 or 0 indicate respectively that the corresponding unit is included or not in the sample. However, the balancing equations could be exactly satisfied with some fractional element of π (the number of these elements are at most equal to the number of balancing variables), and therefore the flight phase does not determine if the corresponding units have to be drawn or not in the sample. In these cases, the cube method searches for an approximately balanced sample solution applying the *landing phase*; this solution relaxes some constraints expressed in (3.1) minimising the variance of the Horvitz-Thompson estimators of the respective variables. The landing phase is set up as a rounding problem and it is solved via linear programming techniques. In the multi-way stratification design, as described below, the cube method is able to select a balanced sample via the flight phase only (Deville and Tillé, 2000). In Chauvet and Tillé (2006), a fast algorithm for executing the flight phase implemented by means of a SAS-IML macro is presented. The procedure is computationally feasible for large data sets with a lot of balancing variables as it has been tested by the authors.

3.2. Balanced sampling applied for marginal stratification

Let us suppose that a vector of inclusion probabilities π consistent with the marginal distribution $\mathbf{n}_1$ and $\mathbf{n}_2$ is available, that is

$$n_{r.} = \sum_{k \in U_r} \pi_k \quad (\forall r = 1, \dots, R) \quad n_{.c} = \sum_{k \in U_c} \pi_k \quad (\forall c = 1, \dots, C). \quad (3.2)$$

If π is not available, the well known Iterative Proportional Fitting (IPF, see Bishop *et al.*, 1975) or the Generalised Iterative Proportional Fitting (GIPF, Dykstra, 1985) procedure can be used to define it. Starting from an initial vector $^*\pi = (^*\pi_1, \dots, ^*\pi_k, \dots, ^*\pi_N)'$ of inclusion probabilities such that $\sum_{k \in U} ^*\pi_k = n$, IPF or GIPF adjust the initial probabilities computing the final vector π satisfying the marginal sample sizes constraints (3.2). The main useful difference between IPF e GIPF is that the second procedure avoids to define inclusion probabilities greater than 1. Therefore, it may be better to adopt GIPF to satisfy the equations (3.2). However, GIPF and IPF obtain the same results when the

latter procedure does not yield a solution with elements of π greater than 1. A third more empirical procedure to calculate π , satisfying the constraints (3.2), is to use IPF procedure iteratively in the following way: starting from initial vector of inclusion probabilities $^*\pi$, at the end of the procedure we equal to 1 the π_k 's greater than value 1. Then, we reply the IPF with the same starting inclusion probabilities removing the units with inclusion probabilities equal to 1 and updating the marginal constraints so that if the generic unit k belongs to the subpopulation U_{rc} , the new marginal sample sizes become $n_{r.} - 1$ and $n_{.c} - 1$.

The selection by means of a balanced design, taking into account planned small sample size for two marginal distributions, is up to the choice of the auxiliary variables. Then, we use the set of variables related to the two marginal stratifications variables $\mathbf{V}_1$ and $\mathbf{V}_2$ such that for unit k have the following values:

$$^1x_{rk} = \begin{cases} \pi_k & \text{if } k \in U_r. \\ 0 & \text{if } k \notin U_r. \end{cases} \quad ^2x_{ck} = \begin{cases} \pi_k & \text{if } k \in U_{.c} \\ 0 & \text{if } k \notin U_{.c} \end{cases}.$$

The flight phase of cube method selects a random sample which satisfies the following constraints:

$$\sum_{k \in U} \frac{^1x_{rk}}{\pi_k} \lambda_k = \sum_{k \in U_r} \pi_k = n_{r.}; \quad \sum_{k \in U} \frac{^2x_{ck}}{\pi_k} \lambda_k = \sum_{k \in U_{.c}} \pi_k = n_{.c} \quad (3.3)$$

Equations (3.3) imply that the sample sizes of each subsample $s_{r.}$ and $s_{.c}$ must be respectively equal to the fixed marginal frequencies $n_{r.}$ and $n_{.c}$. Deville and Tillé (2000), shows that equations (3.3) can be exactly satisfied by means of the cube method via the flight phase.

Finally we underline some interesting aspects of balanced sampling:

- the cube method is quite general and can be easily applied to more than two not-nested partitions in the domains of interest;
- we may define additional balancing equation involving other variables correlated with the study variable allowing to improve the precision of estimators of totals.
- as far as concerning the estimation phase, the balanced sampling can be used jointly with an estimator exploiting auxiliary information. This sampling strategy allows to bound the variability of the sampling weights (for instance, calibrated weights), that causes problems especially in small domains of estimate. For instance, if in the calibration phase we use the known population count at domain level, it should be useful to use the same auxiliary information as constraints in the design phase, in addition to the auxiliary variables allowing to respect the marginal sample sizes.

4. Methods based on coordinated sampling

The problem under study may be considered as a particular case of coordination of the samples. Generally, coordinated sampling is adopted in order to guarantee an expected predefined overlap rate between two or more surveys. Minimum and maximum overlap is indicated respectively as *negative* and *positive* coordination. In our case, the sampling units are selected from two separate sampling designs on the same target population using respectively $\mathbf{V}_1$ and $\mathbf{V}_2$ as stratification variables. The two

samples are maximum positively coordinated and the resulting final sample is composed by the union of the two samples.

In this context the techniques can be classified into:

- methods based on the *permanent random numbers* (Ohlsson, 1995);
- methods implementing linear programming algorithm (Ernst and Paben, 2002).

Within the permanent random number (PRN) techniques, each unit in the frame list is assigned a random number drawn independently from the uniform distribution on the interval $[0,1]$. The PRN techniques allow to implement different sampling designs such as: simple random sampling without replacement, Bernoulli, Poisson and other probability proportional to size sampling procedures. Then, two maximum positively coordinated samples are selected from the two sampling designs using the assigned PRNs. The second type of methods, that will not be dealt within this paper, are procedures maximizing the overlap among samples and are based on sequential application of transportation problems. The algorithms seem to be rather complex if compared with permanent random numbers based procedures.

4.1. General introduction to coordinated sampling using permanent random numbers

In the PRN context, the sampling units are selected from two separate sampling designs, say $p_1(\cdot)$ and $p_2(\cdot)$, on the same target population using respectively $\mathbf{V}_1$ and $\mathbf{V}_2$ as stratification variables.

Let iS and iS denote respectively the generic sample and the set of all the subset iS generated by the design $p_i(\cdot)$ ($i=1,2$). The inclusion probabilities of unit k in the marginal design $p_i(\cdot)$ are indicated with $i\pi_k = \sum_{iS \in iS} p_i(iS) i\lambda_k$ being $i\lambda_k$ a dummy variable that equals to 1 if $k \in iS$ and equal to 0 otherwise.

Two positively coordinated samples, say $1S$ and $2S$ are selected respectively according to $p_1(\cdot)$ and $p_2(\cdot)$ designs and the resulting final sample s is composed by the union of the two samples. Let us denote with z_k the PRN assigned to the unit k ($k=1, \dots, N$) and for each design above considered let us introduce the following quantities: the sample sizes, $\mathbf{n}_1 = (n_{1.}, \dots, n_{1r}, \dots, n_{1R})'$ and $\mathbf{n}_2 = (n_{2.}, \dots, n_{2c}, \dots, n_{2C})'$, being $\sum_{r=1}^R n_{1r} = \sum_{c=1}^C n_{2c} = n$, $\sum_{k \in U_{1r}} 1\pi_k = n_{1r}$ and $\sum_{k \in U_{2c}} 2\pi_k = n_{2c}$. The final sample s , being $s = 1S \cup 2S$, is selected with the joint design $p(s)$ and the corresponding inclusion probabilities are given by $\pi_k = \Pr(k \in s) = \sum_{s \in S} p(s) \lambda_k$, with $\lambda_k = 1$ if either $1\lambda_k$ or $2\lambda_k$ equals to 1.

With these techniques, the realised sizes of the final sample are random outcomes with expected values larger than the constraints n , $\mathbf{n}_1$ and $\mathbf{n}_2$. Empirical results on Italian business surveys has shown that the overall realised sample size is generally larger than the prefixed n between 10% and 30%. In order to partially overcome this problem Falorsi *et al.* (2006) show an algorithm which, starting from initial $1\pi_k^*$ and $2\pi_k^*$ probabilities, calculates modified inclusion probabilities, $1\pi_k$ and $2\pi_k$. These guarantee that the expected marginal sample sizes in the joint sample design are roughly equal to the constraints $\mathbf{n}_1 = (n_{1.}, \dots, n_{1r}, \dots, n_{1R})'$ and $\mathbf{n}_2 = (n_{2.}, \dots, n_{2c}, \dots, n_{2C})'$. In the following section we will illustrate how coordination may be realised with some important designs.

4.2. Coordination with some sampling designs

Design I: Poisson sampling

With this design we have

$${}_1\pi_k = n_r g_k / \sum_{k \in U_r} g_k \quad (r=1, \dots, R); \quad {}_2\pi_k = n_c g_k / \sum_{k \in U_c} g_k \quad (c=1, \dots, C), \quad (4.1)$$

where g_k is a non-negative measure of size for the unit k . The unit k is included in the sample ${}_i s$ ($i=1$ or 2) if $z_k \leq {}_i\pi_k$. The inclusion probabilities of the final sample are, thus, $\pi_k = \max({}_1\pi_k, {}_2\pi_k)$.

Design II: probability proportional to size without replacement

This design may be only approximated with PRN techniques. One approximated solutions proposed in literature is the *order sampling* (Rosen 1997a, 1997b). The design may be summarised as follows: the expected sample sizes are larger than the constraints n , $\mathbf{n}_1$ and $\mathbf{n}_2$; there is no formal proof of the values assumed by the inclusion probabilities π_k , but in practical situations they may be well approximated by $\pi_k = \max({}_1\pi_k, {}_2\pi_k)$.

The order sampling represents an interesting way to avoid the perceived drawback given by random sample size of Poisson sampling and still come very close to achieving the desired inclusion probabilities (4.1). A particular case of order sampling is *Pareto sampling*, carried out as follows:

- (i) for every $k \in U$, compute $\xi_{ik} = z_k(1 - {}_i\pi_k)/(1 - z_k) {}_i\pi_k$ ($i=1,2$);
- (ii) for the sample design $p_i(\cdot)$ ($i=1,2$), the frame units are ordered by stratum and in each stratum by ascending order of ξ_{ik} . In stratum U_r of the design $p_1(\cdot)$, the sample ${}_1 s$ is composed of the first n_r units in the ordered list and in stratum U_c of the design $p_2(\cdot)$, the sample ${}_2 s$ is composed of the first n_c units in the ordered list.

Another case of order sampling is the *sequential Poisson sampling*. The sampling selection is carried out with steps (i) and (ii) above illustrated but the ξ_{ik} values are defined by $\xi_{ik} = \xi_k = z_k / N g_k$. Both the above order sampling schemas realise the marginal fixed sample sizes $\mathbf{n}_i$ ($i=1$ or 2) for a given marginal design ${}_i s$ ($i=1$, or 2), but the inclusion probabilities for $p_1(\cdot)$ and $p_2(\cdot)$ are slightly different from the planned probabilities ${}_1\pi_k$ and ${}_2\pi_k$ expressed by (4.1).

5. Estimation of domain totals

Some considerations on the estimation are presented in this section.

In order to simplify the notation of the estimators, we denote the populations U , U_r and U_c generically as U_d ($d=1, \dots, D=R+C+1$). Furthermore with analogous symbolism, we indicate with N_d , s_d and Y_{jd} ($j=1,2,3$) respectively the population size of U_d , the sample and the total of the variable interest Y_j on the population U_d . Moreover, we suppose that the variables Y_1 is known for each unit of U , while the values of Y_2 and Y_3 are known for the sample units. This different

informative context produces two types of *direct* estimator of parameter of interest considered in the following

$${}_{dir}\tilde{T}_d = \sum_{k \in s_d} y_{3k} a_k / Y_{1d} , \quad (5.1a)$$

$${}_{dir}\tilde{R}_d = \sum_{k \in s_d} y_{3k} a_k / \sum_{k \in s_d} y_{2k} a_k , \quad (5.1b)$$

where a_k are the *base* weights. If the probabilities π_k are known (as in the case of balanced sampling and in Poisson sampling), the *base* weight is $a_k = 1/\pi_k$. Otherwise, if the π_k values are unknown (as in the case of Pareto sampling and sequential Poisson sampling), we may approximate each of them as $\pi_k = \max(\pi_{k,1}, \pi_{k,2})$, but other values may be proposed (Falorsi *et al.*, 2006). Domain estimates could be improved using auxiliary information and a superpopulation model $E_m(y_{jk}) = f(\mathbf{x}_k; \boldsymbol{\beta})$ that for each unit k links the model expected value, $E_m(y_{jk})$, to the auxiliary vector $\mathbf{x}_k$, where $f(\cdot; \boldsymbol{\beta})$ is a functional form depending on the vector $\boldsymbol{\beta}$ of unknown parameters. Having obtained the sample estimates $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, based on $\{y_{jk}, \mathbf{x}_k, a_k; k \in s\}$, the predicted value $\tilde{y}_{jk} = f(\mathbf{x}_k; \tilde{\boldsymbol{\beta}})$ for each unit $k \in U$ are computed, assuming that $\mathbf{x}_k$ is known for each unit k in the population. Considering the ratio between the totals of the variables Y_2 and Y_3 two different estimator types are analysed:

$${}_{syn}\tilde{R}_d = \frac{\sum_{U_d} \tilde{y}_{3k}}{\sum_{U_d} \tilde{y}_{2k}} , \quad (5.2)$$

$${}_{greg}\tilde{R}_d = \frac{\sum_{U_d} \tilde{y}_{3k} + \frac{N_d}{\sum_{s_d} a_k} \sum_{s_d} a_k (y_{3k} - \tilde{y}_{3k})}{\sum_{U_d} \tilde{y}_{2k} + \frac{N_d}{\sum_{s_d} a_k} \sum_{s_d} a_k (y_{2k} - \tilde{y}_{2k})} . \quad (5.3)$$

The estimators (5.2) are computed as the ratio of *synthetic* estimators of the totals Y_{3d} and Y_{2d} , while in (5.3) the *generalised regression* estimators for the two totals are considered (Sämdal *et al.*, 1992, p. 399). The expressions of ${}_{syn}\tilde{T}_d$ and ${}_{greg}\tilde{T}_d$ are quite straightforward, substituting the denominators values given by the *synthetic* or *generalised regression* estimators with the known totals.

Other estimators of ratio parameter could be computed on other total estimators like the composite estimators, obtained by the linear combination between synthetic and generalised regression estimators (Rao, 2003). However, the aim of the paper is to examine the design effect on the sampling strategy and for brevity sake we just consider the estimators given by (5.2) and (5.3).

The predictions $\{\tilde{y}_{jk}; k \in U\}$ differ from a model specification to another, depending on the functional form and the choice of the auxiliary variables. In particular, the linear model allows to define the above estimators knowing only the domain totals of the auxiliary information and the $\mathbf{x}_k$ values

for the sampling units. However, having access to $\mathbf{x}_k$ values for every $k \in U$, it is possible to construct estimators with more efficient predictions $\tilde{y}_{jk}$ obtained by generalised linear models (Lehtonen and Veijanen, 1998).

Finally, let us point out that the synthetic estimator relies on the truth of the model and is usually biased; nevertheless, as noted in Lehtonen *et al.* (2003), if the model incorporates a specific domain effect (fixed or random) the bias is strongly reduced and therefore it is quite important to control the sample size at the domain level.

6. Monte Carlo experiment to evaluate the sampling strategies

6.1. Background information

The simulation is carried out on the 1999 population of the enterprises from 1 to 99 employees belonging to the *Computer and related economic activities* (2-digits of the NACE rev.1 classification), in which the units with outlier values have been deleted obtaining a set of $N=10,392$ units. According to the EU Council Regulation n°58/97 on Structural Business Statistics the estimation domains are defined by means of two partitions: (1) *geographical region* with 20 marginal domains, hereinafter referred to as *DOM1*; (2) *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, hereinafter referred to as *DOM2*. The number of the cross-classification strata is 480 but only 360 strata have one or more population units. Three variables are considered in the simulation; the *number of employers*, *labour cost* and *value added*. The variable values are available for each unit in the population by an administrative data source. The first variable is supposed to be known at the sampling stage, while the other two variables represent the variables of interest. Starting from these variables, two measures of the *productivity* have been built: the *Labour Productivity based on value added* (LP) defined by the ratio between value added and number of employers¹; the ratio of *Labour cost* to *Value added* (LV). These two variables have one main difference: in LP the denominator is a known value, while LV is a ratio of two estimated values.

The experiment is based on an overall sample size of 360 units, planning the sample size of each domain of DOM1 and DOM2 such that the coefficient of variation of the domain estimates for the variable *number of employers* is bounded to be lower than 34.5% for the DOM1 domains and to 8.7% for the DOM2 domains. This sample allocation represents a compromise between the Allocation Proportional to Population (APP) size and an allocation uniform for each domain of each partition. In the following we refer 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* or *small subpopulations*. The analysis of the empirical results will be focused on these domains.

In the simulation we have considered seven sampling designs as reported in table 1.

In particular, in the BALPOP design the balancing equations $\sum_{k \in s_d} 1/\pi_k = N_d$ ($d=1, \dots, 44$), is added to the equations (3.3). Starting from initial probabilities $\pi_k = n/N$, the inclusion probabilities for BALPOP and BAL have been attained applying the IPF procedure iteratively (four times) as described in section 3.2. In the coordinated designs CPAR, CPOI and CSPOI, the initial

¹From the perspective of production analysis, labour input is most appropriately measured as the total number of hours worked. Therefore, this is just a rough measure of labour productivity. For detailed review of productivity measure issues, see OECD (2001).

probabilities ${}^1\pi_k = n_r./N_r.$ for $k \in U_r.$ and ${}^2\pi_k = n_c./N_c$ for $k \in U_c$ have been modified and the final inclusion probabilities ${}^1\pi_k$ and ${}^2\pi_k$ have been obtained in eight iterations by means of the algorithm described in Falorsi *et al.* (2006).

Five hundred Monte Carlo samples have been selected for each sampling design. The estimates of the domain totals have been computed, for each sample, by the *direct*, *generalised regression* and *synthetic* estimators.

Table 1: *Sampling Design used in the simulation study*

Sampling Design	Abbreviation
Stratified on DOM1 with SRSWOR [*] in each stratum	STDOM1
Stratified on DOM2 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
Coordinated Poisson sampling	CPOI
Coordinated Sequential Poisson sampling	CSPOI

^{*}SRSWOR: Simple Random Sampling Without Replacement

As far as concerning the estimators using auxiliary information, a linear models have been implemented using 10 auxiliary variable, six of them are the *economic activity group* membership indicators, and the remaining four are the *size class* membership indicators. In the following, each sampling strategy is indicated in short by the couple (*dis*, *est*), where *dis* indicates one of the 7 designs referred to in table 1 and *est* assumes the categories *dir*, *syn*, *greg* respectively for direct, synthetic and generalised regression estimators.

We have computed two quality measures: the average *Absolute Relative Bias* ($\overline{ARB}$) and the average *Relative Mean Square Error* ($\overline{RMSE}$). In the case of estimators like (5.1b), (5.2) and (5.3) they are expressed by

$$\overline{ARB}_F(dis, est) = \frac{1}{card(F)} \sum_{d \in F} \left| \frac{1}{500} \sum_{i=1}^{500} [\tilde{R}_d^i(dis) - R_d] / R_d \right| \times 100,$$

$$\overline{RMSE}_F(dis, est) = \frac{1}{card(F)} \sum_{d \in F} \left\{ \frac{1}{500} \sum_{i=1}^{500} [\tilde{R}_d^i(dis) - R_d]^2 / R_d^2 \right\} \times 100$$

denoting with: F a specific subset of the marginal domains; $card(F)$ the cardinality of F ; ${}_{est}\tilde{R}_d^i(dis)$ the i -th Monte Carlo sample estimate ($i=1, \dots, 500$) of the ratio R_d 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). Furthermore, when the total at the denominator of the ratio is known, let us substitute ${}_{est}\tilde{R}_d^i(dis)$ with ${}_{est}\tilde{T}_d^i(dis)$.

6.2. Simulation results²

² Section 6.2 has been written together with Danilo Orsini, who has implemented the software for the empirical study.

The Monte Carlo simulation study highlights that the multi-way stratification techniques proposed in this paper are able to take under control bias and variability with respect to the benchmark strategies called STDOM1 and STDOM2 collapsing one of the two stratification variables. The results in table 2 are referred to the quality measures on the small subpopulation set. The first part is devoted to the direct estimator, the second part is dedicated to *greg* estimator. We exclude in the table the *syn* estimators because in the simulation results it shows high bias for all the designs. Finally, we underline the findings only to the quality measures of the estimates computed on the overall set of small subpopulation. In particular, we point out that:

- the balanced and coordinated designs seem to perform better than the benchmark strategies STDOM1 and STDOM2 for the two variables and with respect to the *dir* and the *greg* estimators;
- considering the proposed designs, BALPOP has the smallest bias and variability for variables LV, while in the case of LP variables BAL presents good performances in terms of variability;
- usually balanced designs are preferred to coordinated designs, especially for the LV estimates, even if, for the LP estimates, CSPOI does not present a large difference at least with the *greg* estimators;
- among CPOI, CPAR and CSPOI, the latter shows the best results and in some case it is quite close to the BAL results.
- adopting the STDOM1 occurs that the productivity parameters for about 11% of the small subpopulation cannot be estimated because of null sample size. In STDOM2 designs that happens for about 2.5% of the overall set of small domain.

Table 2: *Comparison of small domains sampling strategies: Average Absolute Relative Bias (ARB) and Relative Mean Square Error (RMSE)*

	Direct estimator				Generalised regression estimator			
	LP		LV		LP		LV	
	ARB	RMSE	ARB	RMSE	ARB	RMSE	ARB	RMSE
STDOM1	5.41	102.74	6.81	29.41	4.78	54.14	8.56	41.62
STDOM2	1.75	55.23	100.93	86.73	6.10	51.08	8.07	53.66
BALPOP	1.06	32.58	2.13	15.28	1.83	22.48	2.36	12.55
BAL	1.19	34.03	2.91	17.08	2.10	22.28	3.54	16.72
CPAR	2.18	44.60	2.65	24.98	2.17	23.72	3.12	31.30
CPOI	2.36	52.67	5.81	49.40	3.48	26.34	9.80	60.50
CSPOI	2.24	46.15	2.26	18.06	1.88	23.82	2.67	17.55

7. Conclusions

In the present work, balanced and coordinated sampling designs for controlling the sample size in all the subpopulation of interest for different partitions of the population, have been examined. In particular, these sampling methods may be applied to deal with the estimation of small subpopulation in the phase of the sampling design allowing to have planned size for each subpopulation and to define an overall strategy considering jointly the design and estimation phase. The properties of the sampling techniques have been analysed by means of a simulation study conducted on real business population data for the estimates of some measures of productivity that can be defined on the available information. The findings show the optimal behaviour of balanced sampling with direct estimator, while with estimators involving auxiliary information coordinated samplings are slightly worse. Moreover, the experiment highlights that the proposed techniques do not suffer from the computational complexity.

Acknowledgments

I wish to thank Piero Demetrio Falorsi for many helpful discussions.

REFERENCES

- Bishop Y., Fienberg S., Holland P. (1975) *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA.
- Chauvet G., Tillé Y. (2006) A Fast Algorithm of Balanced Sampling, to appear in *Journal of Computational Statistics*.
- Derville J.-C., Tillé, Y. (2000) Selection a several unequal probability samples from the same population, *Journal of Statistical Planning and Inference*, **86**: 89-101.
- Derville 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.
- 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., Falorsi S., Righi P., Solari F. (2006) Sampling Design for Small Domains Estimation through multi-way stratification techniques, in *Metodi Statistici per l'Integrazione di Dati da Fonti Diverse*, Liseo B., Montanari G. M., Torelli N. (eds), Franco Angeli, Milano.
- 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.
- OECD (2001) *Measuring Productivity – OECD Manual*, OECD, Paris (www.SourceOECD.org).
- 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.
- Porter M. E. (2003) Building the Microeconomic Foundations of Prosperity: Findings from the Business Competitiveness Index in: *The Global Competitiveness Report 2003-2004*, Porter M. E., Sala-i-Martin X., Lopez- Carlos A., Schwab K. (eds) Oxford University Press Inc, New York.
- Rao J. N. K. (2004) *Small Area Estimation*, Wiley, New York.
- 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.
- 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.
- Valliant R., Dorfman A. H., Royall R. M. (2000) *Finite Population Sampling and Inference: A Prediction Approach*, Wiley, New York.