

Integrated use of Different Administrative Data Sources for the Sampling Strategy of the Survey on Structural Business Statistics

*Uso Integrato di Fonti di Dati Amministrativi per la Strategia di
Campionamento dell'Indagine sulle Statistiche Strutturali sulle Imprese*

Piero Demetrio Falorsi

Istat, via Magenta, 2 - Roma, falorsi@istat.it

Salvatore Filiberti

Istat, via Tuscolana, 1782 - Roma, filibert@istat.it

Alessandro Pallara

Istat, via Giuseppe Verdi, 18 - Napoli, pallara@istat.it

Riassunto: Il crescente uso di dati amministrativi presso gli istituti nazionali di statistica comporta lo sviluppo di nuove metodologie e strategie d'indagine al fine di sfruttare al meglio queste fonti informative. Tuttavia il potenziale descrittivo costituito dai dati rilevati attraverso le indagini dirette rimane di importanza fondamentale. In questo lavoro si propone una nuova strategia di campionamento per le indagini strutturali sulle imprese, basata sull'uso combinato di fonti amministrative e statistiche, che consente di migliorare alcune caratteristiche delle stime finali delle variabili economiche riducendo, allo stesso tempo, il carico statistico sulle unità di rilevazione.

Keywords: Administrative Data, Collocated Sampling, Anticipated Variance.

1. Introduction

In recent years data from administrative sources have been using increasingly at NSIs for improving structural business statistics (SBS). In this paper a new sample strategy based on extensive use of administrative data is suggested. The proposed strategy faces the shortcomings related to the estimation of economic variables through joint use of several data sources. In order to define the estimation methodology the predictive approach is followed for exploiting all available information and thus increasing the accuracy of the estimates. The paper is structured as follows. In section 2 the administrative and statistical data sources are introduced and the sampling strategy is shortly described. Some empirical results are presented in section 3. Finally, section 4 proposes some concluding remarks.

2. Data sources and methodology

Business statistical Register (BR) of the 4.2 million Italian active enterprises, Balance Sheets (BS) for about 570,000 joint-stock companies and Social Security (SS) data for broadly 1.2 million enterprises with employees represent essential sources of

administrative information for SBS. Statistical data are provided by the Enterprises Economic Accounts Survey (EEAS) (Istat, 2005) about several variables, for three types of domain (DOM1: *Class* of economic activity; DOM2: *Group* of economic activity by size-class of employment; DOM3: *Division* of economic activity by geographical region). Each domain defines a partition of the population of interest. EEAS stratified sampling design is based on the BR structural variables, under the assumption that they are strongly correlated to the target variables. Appropriate combinations of these strata identify the *planned domains*. Sample units allocation into strata is based on the generalization of the Neyman method (Bethel, 1989). All large enterprises (roughly 10,000) with at least 100 employees belong to a take-all stratum. Approximately 120,000 small and medium enterprises are included in the sample. Final estimates are obtained through calibration estimators (Deville and Särndal, 1992). We are interested in the estimation of totals defined by:

$$vY_{tdU} = \sum_{k \in U} v y_{k \quad td} \delta_k = \sum_{k \in tdU} v y_k = \sum_{k=1}^{tdN} v y_k \quad (t = 1, \dots, T; d = 1, \dots, D_t)$$

where tdU , for the t -th partition (type of domain) on U (population of interest of size N), is a specific domain of interest of size tdN , v is the generic variable of interest ($v \in V$), $v y_k$ is the value of v with reference to the k -th unit of U , $td \delta_k$ is a binary variable that is 1 if the unit k belongs to domain tdU and is 0 otherwise. Let a be a generic administrative data source ($a = 1, \dots, A$) such that: *i*) it provides information about sub-population U_a (of size N_a) ($U_a \subseteq U$); *ii*) $\forall k \in U_a$, a vector ${}_a \mathbf{x}_k$ of auxiliary variables is available; *iii*) a subset V_a of the variables of interest ($V_a \subseteq V$), is provided by the a -th administrative data source; v_a ($v_a \in V_a$) is the generic variable of interest, $v_a y_k$ is its value for the unit k in sub-population U_a , ($v_a \in V_a$ and $k \in U_a$); *iv*) the $v_a y_k$ ($k \in U_a$) values represent a specific sub-set of the variables included in the vector ${}_a \mathbf{x}_k$, then the totals $v_a Y_{U_a}$ are known ($v_a Y_{U_a} = \sum_{k \in U_a} v_a y_k$); v) for the unit k it is defined a column vector $\boldsymbol{\gamma}_k = ({}_1 \gamma_k, \dots, {}_a \gamma_k, \dots, {}_A \gamma_k)'$ of A binary variables in which ${}_a \gamma_k = 1$ if $k \in U_a$, or ${}_a \gamma_k = 0$ otherwise. Define ${}_{(g)}U$ as the sub-population of the units having the same information provided by administrative sources (a *configuration*). Let ${}_{(g)}\boldsymbol{\gamma} = ({}_{(g),1} \gamma, \dots, {}_{(g),a} \gamma, \dots, {}_{(g),A} \gamma)'$ be the vector that defines the configuration g , ($g = 1, \dots, G$), where ${}_{(g),a} \gamma$ is a binary variable (0; 1). Then ${}_{(g)}U$, ($g = 1, \dots, G$), define a partition over U . A super-population model for the k units in ${}_{(g)}U$ is:

$$v y_k = ({}_{(g),v} \boldsymbol{\beta})' {}_{(g)} \mathbf{x}_k + {}_{(g),v} e_k \quad k \in {}_{(g)}U, \quad (1)$$

where ${}_{(g)} \mathbf{x}_k$ is a vector of ${}_{(g)}P$ auxiliary variables; ${}_{(g),v} \boldsymbol{\beta}$ is a vector of ${}_{(g)}P$ regression parameters; ${}_{(g),v} e_k$ are the random residuals with model expectations $E_m(\cdot)$: $E_m({}_{(g),v} e_k) = 0$, $E_m({}_{(g),v} e_k^2) = c_k {}_{(g),v} \sigma^2$ and $E_m({}_{(g),v} e_k, {}_{(g),v} e_{k'}) = 0$ (for $k \neq k'$), where c_k is the known eteroschedasticity factor and ${}_{(g),v} \sigma^2$ is an unknown scalar

($\forall k \in (g)U$) of model (1). The $(g)\mathbf{x}_k$ vector is defined as follows:
 $(g)\mathbf{x}_k = \left((g), 1' \mathbf{x}_k, \dots, (g), a' \mathbf{x}_k, \dots, (g), A' \mathbf{x}_k \right)'$ for $k \in (g)U$. A *synthetic sampling design* based on information from administrative data sources can be then described as follows:
 (i) sample sizes are planned for each sub-population $(g)U$, that represents a sampling design stratum, but are not planned for the subpopulations $_{td}U$; (ii) the $_{v}Y_{td}U$ totals are estimated by synthetic regression estimator $_{syn,v}\tilde{Y}_{td}U$ (Särndal *et al.*, 1992); (iii) the sampling design is aimed at minimizing the *Anticipated Variance* (AV) of these estimates (Brewer, 1999); (iv) from $(g)U$ a sample $(g)^S$ of $(g)^n$ units is selected by a *collocated sampling* (Ohlsson, 1995) with $\pi_k \propto \sqrt{c_k}$. The synthetic estimator for the total of the domain td is:

$$_{syn,v}\tilde{Y}_{td}U = \sum_{g=1}^G \sum_{k \in (g,td)U} (g),_v \tilde{\mathbf{b}}' (g)\mathbf{x}_k + \sum_{g=1}^G \sum_{k \in (g)^S} r_k (g),_v \tilde{e}_{k\ td} \delta_k = \sum_{g=1}^G \sum_{k \in (g)^S} _v y_{k\ td}$$

where $(g),_v \tilde{\mathbf{b}}'$ are the estimated regression parameters, $(g),_v \tilde{e}_k$ are the estimated residuals, $_{td}w_k$ are the final sampling weights, that are function of known values that may identify specific forms of the estimator (r_k and q_k). Summing up the G subpopulations and the D_t domains of the partition t , we obtain:

$$_{v_a}\tilde{Y}_U = _{v_a}Y_{U_a} + _{v_a}\tilde{Y}_{U_{\bar{a}}} = \sum_{k \in _a U} _{v_a}y_k + \sum_{k \in s \cap _{\bar{a}} U} _{v_a}y_k w_k.$$

The synthetic estimator, for each of the sampled units, produces as many sampling final weights as are the domains of interest in the g -th model group in which the unit k is included. When the number of domains of interest is relatively small, it may be used a *calibrated sampling strategy* (cfr. Casciano *et al.*, 2005). This strategy is based on the selection of T independent samples from each partition related to the domains of interest through permanent random numbers technique (Ohlsson, 1995). An important property of the Calibrated Direct Domain (CDD) estimator is that it produces a unique final weight $_z w_k$ for the unit k and for each domain of interest. This guarantees internal coherence of estimates, easy computational procedure and interpretability of results.

3. Empirical results

An empirical study is carried out using data coming from BR, BS and SS. The population of interest is represented by the enterprises belonging to the “Computer and related activities” sector (NACE rev.1.1 classification of economic activities). The data base used includes $N=10,392$ enterprises and the sample selected has $n=922$ units. Model groups ($G=6$) have been obtained by clustering, with a non parametric regression technique (CART), the population units on the basis of their characteristics, such as size class and geographical location. The following sampling strategies have been studied: the simple random sampling (*SRS*); the *synthetic* simple random sampling (*SSRS*) [$c_k=1$] and [c_k =employees, ADD in Table 1]; the *calibrated* sampling (*Cal*) [c_k =ADD]. For each partition of the population introduced in section 2, the Coefficient of Variation (CV) of

the variable *personnel costs* have been obtained, for each approach, as the ratio between square root of the AV and the total of the variable of interest, based on equal sample size, and then compared. The results are presented in Table 1, which shows the median CV for all domains in the partition as well total CV for each sampling strategy. The results show that the *Cal* is generally more efficient than SRS.

Table 1: *CV of a SRS compared to the proposed sampling strategies for personnel cost estimation [(1)=SSRS; (2)=Cal].*

<i>type of Domain</i>	D_t	CV_{SRS}	$CV_{(1)}$ $c_k = 1$	$CV_{(1)}$ $c_k = ADD$	$CV_{(2)}$ $c_k = ADD$
<i>DOM1</i>	6	0,0752	0,0617	0,0469	0,0221
<i>DOM2</i>	24	0,1013	0,1030	0,0901	0,0552
<i>DOM3</i>	20	0,1214	0,0856	0,0626	0,0510
Total		0,039	0,0146	0,0112	0,0107

4. Conclusions

Simulation results show that both sampling strategies, synthetic and calibrated, provide estimates of SBS that are more accurate than those obtained through a traditional strategy. Moreover, the first one has also the computational advantage to produce one final sampling weight. The synthetic estimator may be an useful tool for producing sampling estimates when there are non-planned domains of interest or cells with no observed sample since the selection of sampling units is needed only at model group level. Using a compromise sampling strategy, calibrated methodology may then be adopted for those partitions of the population of interest that are subject to international Regulations on SBS, while synthetic approach may be used for non priority domains.

References

- Bethel, J. (1989), "Sample Allocation in Multivariate Surveys", *Survey Methodology*, 15: 47-57.
- Brewer, K.R.W. (1999), "Cosmetic Calibration with Unequal Probability Sampling", *Survey Methodology*, 25: 205-212.
- Casciano, M.C., Falorsi, P.D., Filiberti, S., Orsini, D., Pallara, A., Russo, A. (2005), "Definizione della strategia di indagine della rilevazione sui conti economici delle imprese mediante l'uso congiunto di dati statistici e amministrativi", in: Falorsi, P.D., Pallara, A., Russo, A. (eds.), *L'integrazione di dati di fonti diverse. Tecniche e applicazioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative*, Franco Angeli, Milano, pp. 233-282.
- Deville, J.-C., Särndal, C. E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87: 376-382.
- Istat (2005), *Conti economici delle imprese – Anno 2000*, Collana informazioni, 6.
- Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers", in Cox, B.G. et al. (eds.): *Business Survey Methods*, Wiley, New York, pp. 153-169.
- Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.