

The Use of Administrative Sources in Planning Sampling Strategy of Structural Business Surveys

*Utilizzo di Fonti Amministrative nella Progettazione della Strategia di
Campionamento delle Rilevazioni Strutturali sulle Imprese*

Salvatore Filiberti

ISTAT – Via Tuscolana, 1788 – Roma, e-mail: filibert@istat.it

Piero Demetrio Falorsi

ISTAT – Via Magenta, 4 – Roma, e-mail: falorsi@istat.it

Riassunto: Obiettivo principale del presente lavoro è quello di delineare i principali criteri che sottostanno alla proposta di nuova strategia di campionamento della rilevazione sui *Conti Economici delle Imprese* condotta dall'Istituto Nazionale di Statistica. Dopo aver definito in modo rigoroso il contesto informativo, caratterizzato dalla disponibilità di dati di indagine e di dati di fonte amministrativa, viene proposta una strategia originale fondata sull'uso congiunto di tali dati e di due differenti approcci all'inferenza: l'approccio predittivo e quello tipico del campionamento da popolazioni finite. Il risultato consiste nell'ottenere stime degli aggregati economici di interesse più accurate rispetto quelle prodotte nell'ambito della strategia campionaria tradizionale.

Keywords: Estimating Equations, Administrative Data, Multi-purpose and Multi-domain surveys.

Introduction

In recent years the Italian National Statistical Institute (ISTAT) has increasingly used administrative data in order to face some important changes in the frame of the Structural Business Statistics (SBS). First of all, the dramatic increase in the demand of SBS in accordance with EU Council Regulation N° 58/97 asking for more detailed and timely information. Secondly, the traditional Enterprises' Economic Accounts Survey (EEAS) is excessively cumbersome in terms of necessary resources as well as of statistical burden for respondent units. Finally, a major problem is represented by non response phenomenon.

Many analyses have been carried out in order to check the opportunities offered by the use of administrative data for statistical purposes. However, the joint use of statistical and administrative data entails some difficulties, particularly with reference to definition of variables. In fact, if the statistical definition of the variable of interest is different from that adopted in the administrative source, then final estimates may be biased. On the other hand, when the definition of variable provided by administrative source is consistent with that requested by EU Regulation, the use of administrative data for statistical purposes has several and non negligible advantages. The most important benefits are: (i) the statistical burden for enterprises and survey costs are reduced; (ii) the accuracy of estimates is improved by using this auxiliary information (Corsini et al.,

2000; Filiberti et al., 1999); (iii) information on phenomena hardly covered by statistical surveys can be collected; (iv) cross-checking can be carried out; (v) the statistical Business Register can quickly be updated.

In this paper we focus on the definition of a new sample strategy based on a massive use of administrative data, aiming to increase the accuracy of parameters of interest with respect to traditional strategy, based on survey data only. The new strategy proposes, in the context of EEAS, a joint use of two different approaches to inference: the predictive approach and the classical approach adopted in sampling for finite populations.

1. Short Description of Traditional EEAS Sampling Strategy

EEAS is a multi-purpose and multi-domain survey which produces statistics on several variables (mainly economic and employment variables) for three types of domains, each defining a partition of the population of interest (see Table 1).

Table 1: *Types of domains of EEAS*

Type of domain		Number of Domains
Code	Description	
DOM1	Class of economic activity (4-digit NACE Rev.1*)	461
DOM2	Group of economic activity (3-digit NACE Rev.1) by size-class of employment	1.047
DOM3	Division of economic activity (2-digit NACE Rev.1) by region	984

*NACE Rev.1 = Statistical Classification of Economic Activities in the European Communities

EEAS is based on a stratified sampling design (ISTAT, 1998). All *large* enterprises (roughly 10,000) with more than 99 persons employed belong to a *take-all stratum*. The small and medium enterprises are stratified by cross-classification of the modalities of the variables DOM1, DOM2 and DOM3. A fixed number of enterprises is selected in each stratum without replacement and with equal probabilities. The number of units to be selected in each stratum is defined as a solution of a linear integer problem. In particular, the minimum sample size is determined in order to ensure that the variance of sampling estimates of the variables of interest in each domain do not exceed a given threshold. In this way, roughly 120,000 small and medium enterprises are included in the sample over a population of about 4.2 millions enterprises.

2. Administrative Sources Available for EEAS

The main administrative sources available for EEAS are the following: *i*) Social Security (*SS*) data; *ii*) Balance Sheet (*BS*) data; *iii*) Fiscal (*F*) data.

SS data are provided by the National Social Security Institute (INPS) and refer to enterprises with employees. They are used to get information on employment, wages and salaries, and personnel costs. A major problem in the use of INPS data for statistical purposes refers to the unit of analysis (the unit who pays contribution to the Social Security System) which does not easily match with that (enterprise) identified in the frame of EU Regulation. Consequently, some bias may occur particularly for those economic sectors mainly represented by individual enterprises, i.e. enterprises with only one person employed.

BS data are provided by the Chambers of Commerce with reference to roughly 650,000 limited enterprises in the Industry and Services economic sectors covering about one half of the total employment. Since 1998 as reference year this source is used for the production of SBS preliminary estimates of the main economic variables (turnover, value added at factor cost, production value, total purchases of goods and services, etc.). *BS* data are also helpful for unit and item non-response treatment in the frame of SBS final estimates.

F data are provided by the Ministry of Economy and Finance and are planned to be used soon in order to estimate economic variables.

3. Defining the New EEAS Sample Strategy

In this paragraph we will give a short description of the proposed new strategy of EEAS based on a massive use of administrative sources. Let us consider one *administrative data* source a , which covers a subset of the population of interest and collects, with good accuracy, a *subset* V_a of the variables of interest. We exclude the phenomenon of non-response. The extension of our model to the real more complicated situation is straightforward. Let us further define: U the population of interest; U_a ($U_a \in U$) the subpopulations of enterprises covered by the administrative source a ; $U_{\bar{a}}$ ($U_{\bar{a}} \in U$) the subpopulations of enterprises complementary to U_a , being $U_{\bar{a}} \cup U_a \equiv U$, $U_{\bar{a}} \cap U_a \equiv \emptyset$; y_k the value of variable of interest y of the unit k ; $\mathbf{y}_{a,k}$ the column vector containing, for the unit $k \in U_a$, the values of all the variables of interest belonging to V_a . Then, if the variable of interest y belongs to the subset V_a , its total Y may be expressed as:

$$Y = \sum_{k \in U} y_k = \sum_{k \in U_a} y_k + \sum_{k \in U_{\bar{a}}} y_k = Y_a + Y_{\bar{a}} \quad (1)$$

where the y_k value, for $k \in U_a$, may be collected directly from administrative source a .

If the variable of interest y does not belong to the subset V_a , its total Y may be expressed as:

$$Y = \sum_{k \in U} y_k = \boldsymbol{\beta}' \mathbf{Y}_a + \sum_{k \in U_{\bar{a}}} \varepsilon_k + Y_{\bar{a}} \quad (2)$$

in which $\mathbf{Y}_a = \sum_{k \in U_a} \mathbf{y}_{a,k}$; $\boldsymbol{\beta}$ and e_k are respectively a vector of *unknown* fixed

regressors and the random residuals of a linear model. This model, for $k \in U_a$, links the variable of interest y_k , to the vector of predictors $\mathbf{y}_{a,k}$:

$$y_k = \boldsymbol{\beta}' \mathbf{y}_{a,k} + e_k, \quad (\text{for } k \in U_a) \quad (3)$$

where the expectations of residual terms are $E(e_k) = 0$, $E(e_k^2) = c_k \sigma^2$, $E(e_k^2, e_{k'}) = 0$ (for $k' \neq k$), being c_k a known scalar.

Expressions (1) and (2) may guide in the definition of the *new* sampling strategy. The most important aspects may be summarised as follows: (i) when $y \in V_a$ and $k \in U_a$,

the value y_k may be collected directly from the administrative source a ; (ii) the sample is used only for estimating β and $Y_{\bar{a}}$. Thus, if the variable of interest y belongs to the subset V_a , the estimate of the total Y may be obtained as:

$$\tilde{Y} = Y_a + \tilde{Y}_{\bar{a}} = Y_a + \sum_{k \in s \cap U_{\bar{a}}} y_k / \pi_k \quad (4)$$

in which s denotes the selected sample and π_k the inclusion probability of unit k . Conversely, if the variable of interest y does not belong to the subset V_a , the estimate of the total Y may be obtained estimating the component Y_a with a predictive approach as:

$$\tilde{Y} = \sum_{k \in s \cap U_a} y_k + \tilde{\beta}' \sum_{k \in \bar{s} \cap U_a} \mathbf{y}_{a,k} + \tilde{Y}_{\bar{a}} \quad (5)$$

where $\bar{s} \cap U_a$ denote the subset of U_a not included in the sample s ; $\tilde{\beta}$ is the sampling

estimate of the vector β , given by $\tilde{\beta}' = \left(\sum_{k \in s \cap U_a} \mathbf{y}_{a,k} \mathbf{y}_{a,k}' / c_k \right)^{-1} \sum_{k \in U_a} \mathbf{y}_{a,k} y_k / c_k$. Alternative

expressions of the estimate of β are illustrated in Fuller (2002). Sampling variances of expression (5) may be calculated with re-sampling methods or using the estimating equation approach (Kovacevic and Binder, 1997).

Conclusions

A new sampling strategy for EEAS aimed to combine survey data with administrative data is suggested. This original method is the result of two different approaches to inference: the predictive approach and the finite population sampling traditional approach. Since a large amount of the totals of main variables of interest is observed with no sampling error by means of the administrative sources, accuracy of estimates may considerably increase, thus improving the quality of SBS.

References

- Corsini V., Di Francescantonio T., Filiberti S., Sanzo, R. (2000), Utilizzo integrato di fonti amministrative e fonti statistiche per la produzione di stime preliminari di alcuni principali aggregati economici previsti dal regolamento della UE n°58/97 sulle statistiche strutturali, mimeo.
- Kovacevic M.S., Binder D. A. , (1997), Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equation Approach, *Journal of Official Statistics*, 13, n.1, pp. 41-58.
- Filiberti S., Di Francescantonio, T., (1999), Progetto di utilizzo dei dati INPS per la produzione di statistiche correnti su occupazione dipendente, retribuzione, costo del lavoro e ore lavorate, mimeo.
- Fuller W.A. (2002), Regression Estimation for Survey Samples, *Survey Methodology*, 28, n.1, pp.5-23.
- ISTAT, (1998), Conti economici delle imprese, *Collana informazioni*.