

A Balanced Sampling Approach for Preliminary Estimates in Panel Surveys

Un Approccio mediante il Campionamento Bilanciato per le Stime Anticipate nelle Indagini Panel

Paolo Righi

Tiziana Tuoto

ISTAT – Via Magenta 4 – 00185 Roma – {parighi, tuoto}@istat.it

Riassunto: La tempestività nella produzione del dato statistico sta assumendo sempre maggiore importanza nell'ambito della statistica ufficiale nazionale ed internazionale. Il problema, in letteratura, è, in genere, affrontato proponendo tecniche di rilevazione che riducono i tempi di raccolta dei dati e tecniche che interessano la strategia campionaria. Il lavoro propone un approccio applicabile nelle indagini in cui è presente una componente di tipo panel, basato sulla selezione di un sottocampione di tipo bilanciato del campione utilizzato per le stime definitive, da osservare nei tempi necessari per produrre le stime anticipate, che permette di migliorare l'accuratezza delle stime. Si illustra una sperimentazione sui dati dell'indagine per le vendite al dettaglio dell'Istat che confronta le stime basate su tale sottocampione con quelle ottenute dalla procedura attualmente usata.

Keywords: Preliminary estimates, two-phase sampling, panel sampling, balanced sampling.

1. Introduction

Timeless of statistical data is becoming a pressing target at both national and international level. Recently, several projects in the EU, one performed by Italian Statistical Institute (Istat), are studying how to deal with the problem of reducing the time to deliver statistical information. Generally, the literature suggests two approaches: the first one considers the overall production process, underling, in particular, the aspects of acceleration of data-collection phase; the second approach implements an *ad hoc* sampling strategies to obtain preliminary estimates with respect to the ordinary date of deliver. This second approach can require some changing of productive process, but they can be restricted to a small part of the sample. In particular, given an overall sample for the final estimate we propose the selection of a sub-sample by means of balanced sampling to obtain preliminary estimates (section 2). For the estimation phase we use a traditional estimator adopted in two phase sampling. Using the data of real survey (section 3), the procedure based on preliminary balanced sampling is compared via a simulation study with the current procedure based on an early self-selected sample of the overall sample (section 4). The results show a general improving of the accuracy of the estimates with respect to the current procedure.

2. Balanced sampling for preliminary estimates for panel surveys data

Given a sample s_{at} used for the final estimates, with first order inclusion probability of

unit k equal to $\pi_{at,k}$, where t denotes the current period of observation, we define a sub-sample of units, s_t (with $s_t \subset s_{at}$), with inclusion probability equal to $\pi_{t,k|at}$, for which the variables of interest are collected within a deadline to produce preliminary estimate. This scheme of selection resembles a two-phase sampling design, where the sub-sample for the preliminary estimates is the second phase sample. Assuming that the parameter of interest is the total $Y_t = \sum_{k \in P_t} y_{t,k}$ with $y_{t,k}$ the value of the variable of interest on unit k at time t , an unbiased estimator is given by

$$\tilde{Y}_t = \sum_{k \in s_t} y_{t,k} / \pi_{at,k} \pi_{t,k|at} = \sum_{k \in s_t} \tilde{y}_{t,k} / \pi_{t,k|at} . \quad (1)$$

The variance of estimator (1) is up to the variance of $\left| \sum_{k \in s_t} \tilde{y}_{t,k} / \pi_{t,k|at} - \sum_{k \in s_{at}} \tilde{y}_{t,k} \right|$ (Särndal *et al.*, 1992, pp. 348-349) and it can be reduced if the difference is equal (or approximately equal) to zero. To obtain this aim, given a set of inclusion probabilities $\{\pi_{t,k|at}\}$ and the variable $y_{t-l,k}$ (known in panel surveys), where l is a lag, and assuming that the relationship $y_{t,k} = b y_{t-l,k}$ is (or approximately) true, we search a balanced sub-sample (Deville and Tillé, 2004) such that

$$\tilde{Y}'_{t-l} = \tilde{Y}'_{at-l} \quad (\text{balancing equation}), \quad (2)$$

with $\tilde{Y}'_{t-l} = \sum_{k \in s_t} y_{t-l,k} / \pi_{at,k} \pi_{t,k|at}$ and $\tilde{Y}'_{at-l} = \sum_{k \in s_{at}} y_{t-l,k} / \pi_{at,k}$. If the linear function between $y_{t,k}$ and $y_{t-l,k}$ is exactly true and the equation (2) is satisfied, in the design based approach the difference term is equal to zero. Balancing equation can consider more than one variable. For instance, if the underlined model is $y_{t,k} = \mathbf{z}'_{t,k} \mathbf{b}$ with $\mathbf{z}'_k = (y_{t-l,k}, \mathbf{x}'_k)$ where $\mathbf{x}_k$ is a vector of auxiliary variables, these must be considered in the balancing equation (2). Recently, Deville and Tillé (2004) proposed a method, called *cube method*, that allows the selection of balanced samples (or approximately balanced) for a large set of auxiliary variables. In this paper the properties of balancing sample are studied, in a simulation approach, to estimate functions of totals such as

$$I_t = \left(\sum_c \alpha_{t-j,c} \gamma_c \sum_{k \in P_{(t-j,t),c}} y_{t,k} / \sum_{k \in P_{(t-j,t),c}} y_{t-j,k} \right) / \sum_c \gamma_c , \quad (3)$$

where c is a sub-population index, $\alpha_{t-j,c}$ and γ_c are given weights, $P_{(t-j,t),c}$ is the sub-population c of units present at time $t-j$ and t .

3. Description of survey data and simulation study

The balanced sampling is applied to the data of the 2004 Italian Monthly Retail Trade

Index survey (Gismondi and De Sandro, 2004) based on a monthly measurement of all goods sold by a stratified probability sample of retail stores of different types and sizes. The sample is composed by a panel and a non-panel component, drawn every year and observed for 12 months. In 2004 the sample size is 7,667 units. The survey provides preliminary estimates within 30 days from the reference time. The preliminary retail trade indices are referred to the mostly food, mostly non-food and they are expressed by the formula (3). In particular, c indicates 5 class of number of employers, t represents the month m ($m=1, \dots, 12$), while j is equal to 12. Therefore, variables $y_{t,k}$ and $y_{t-j,k}$ are respectively the sales of a given month and of the same month of the last year. The weight $\alpha_{t-j,c}$ is the deflector coefficient and γ_c is a class weight, both referred to 2000. The current estimation procedure is based on a sub-sample of the whole survey sample of size 2,349 units. Generally, sub-sample units are the largest within each class c . The 80% belongs to the panel sample and 20% belongs to the non-panel sample. Since October 2004 preliminary estimates are calculated considering all the respondents available within 30 days, even though they are not belonged to the preliminary sub-sample. All non-respondents within 30 days are imputed for the calculation of the preliminary estimates. In order to carry out the simulation study, the final sample of 2004 was replaced by an artificial sample created by an imputation method (ONS, 2001) for which complete monthly data are available. A probability of *Early Response* (ER) within 30 days has been introduced in the simulation study. The probabilities have been estimated, applying a non-parametric classification tree technique. Three different kinds of ER probabilities are considered: firstly, they are equal to those estimated by the classification tree (a); secondly, they are equal to 1 for the units belonging to the balanced sample, 0 otherwise (b); finally, they are equal to 1 for the units belonging to the balanced sample, equal to those estimated by the classification tree otherwise (c). In order to define the balanced design, the set of $\{\pi_{t,k|at}\}$ has been calculated proportionally to the turnover values of 2002 (the last available), obtained from the Istat business register (ASIA). We select an yearly unique preliminary sub-sample, used in each month, composed by two balanced sub-samples with an overall size of 1,929 units as recommended by Eurostat (Gismondi and De Sandro, 2004). The first balanced sub-sample contains 1,543 panel units (the 80% of the 1,929). Balancing equation considers the following variables: the totals of sales in the year 2003 and the ASIA turnovers in 2002. The equations are (approximately) satisfied for food and non-food domains. The second sub-sample with 386 non-panel units (20% of 1,929) is based on the ASIA turnovers and it is also balanced to satisfy the total for food and non-food domains of estimate. The preliminary estimates based on balanced sample are calculated imputing non-respondents within 30 days just if they belong to the preliminary balanced sample. The respondents not-belonging to the preliminary balanced sample are included in the estimate calculation with sample weights equal to 1. The sample weights of the units belonging to the preliminary balanced sample are corrected to take into account the respondents not belonging to the preliminary sample.

4. Results

The simulation study, based on the Monte Carlo method, consists of 500 iterations, in each of them a preliminary balanced sample is selected and the realisation of a sample

of respondents within 30 days is generated. The indices for food, non-food and overall sales are estimated and the *Relative Bias Percentage* (RBP) and the *Mean Absolute Percentage Errors* (MAPE) are calculated to evaluate the measurement errors:

$$RBP_d = \frac{1}{500} \sum_i \frac{1}{12} \sum_m \left(\frac{\hat{I}_{m,d,i} - I_{m,d,i}}{I_{m,d,i}} \right) \cdot 100; MAPE_d = \frac{1}{500} \sum_i \frac{1}{12} \sum_m \left| \frac{\hat{I}_{m,d,i} - I_{m,d,i}}{I_{m,d,i}} \right| \cdot 100,$$

where, for the month m , the domain of estimate d ($d = \text{overall, mostly food, mostly non-food}$), at the iteration i ($i = 1, \dots, 500$), $I_{m,d,i}$ and $\hat{I}_{m,d,i}$ are the index estimates respectively by means of all units of final sample and by means of the early respondents. Table 1 shows the measurement errors relative to the current procedure of estimate and to the procedure based on preliminary balanced sampling. Considering the ER probability of type (a), where data-collection phase is not stressed to obtain ER for balanced sub-sample, performances of current procedure seem better at least for MAPE indicators. With the ER probability (b), where in data-collection phase the balanced sub-sample is carefully considered (response probability equal to 1), the procedure proposed shows better values for RBP and MAPE. Finally, with the ER probability (c), the two procedures are similar for MAPE measures. In this case, it needs to underline that the MAPE indices are very small in both the procedures, because the expected respondent sample size (2,639 units) is very large with respect to the final sample (about 1/3 of the size). For all the three ER probabilities, quite often the approach with balanced sampling provides RBP values smaller than those of the current procedure.

Table 1: *RBP and MAPE of the estimates given by the current and balanced sampling approach*

	Mechanism of early response					
	(a)		(b)		(c)	
	Estimation procedure					
Measurement error	Current	Balanced	Current	Balanced	Current	Balanced
$RBP_{overall}$	0.79	0.77	0.13	0.09	0.10	0.01
RBP_{food}	1.31	1.21	-0.02	0.08	-0.06	0.01
$RBP_{non-food}$	0.42	0.46	0.26	0.10	0.24	0.02
$MAPE_{overall}$	1.58	1.70	1.20	1.00	0.86	0.87
$MAPE_{food}$	1.71	1.89	1.98	1.20	1.45	0.96
$MAPE_{non-food}$	2.24	2.32	1.31	1.45	0.99	1.29
Expected sample size	1,426	1,426	1,929	1,929	2,638	2,638

References

- Deville J.-C., Tillé Y. (2004) Efficient Balanced Sampling: the Cube Method, *Biometrika*, 91, 893-912.
- Gismondi R., De Sandro L (2004) Provisional Estimation of the Italian Monthly Retail Trade Index, *Contributi-Istat*, 24/2004.
- ONS (2001) Artificial Methods for Completing the Population Data Sets, EURAREA, WP1.2 D1.2.2.
- Särndal C.E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.