

# Frame Error Impact on Structural Business Statistics Surveys

## *Impatto dell'Errore di Lista nelle Indagini sulle Statistiche Strutturali sulle Imprese*

Salvatore Filiberti

Istat, via Tuscolana 1782 Roma, [filibert@istat.it](mailto:filibert@istat.it)

Caterina Viviano

Istat, via Tuscolana 1782 Roma, [viviano@istat.it](mailto:viviano@istat.it)

**Riassunto:** Gli errori di lista costituiscono una specifica tipologia di errore non campionario rilevante nell'ambito delle indagini statistiche sulla struttura delle imprese visto che i registri statistici, basati su fonti amministrative e statistiche, necessari per la raccolta dei dati sulle unità economiche, non sono perfettamente aggiornati al momento del lancio dell'indagine campionaria. Questo sfasamento temporale tra periodo di riferimento dell'indagine e aggiornamento della lista di unità statistiche presenti nella popolazione di riferimento può generare distorsione oltre a ridurre l'accuratezza delle stime finali. La metodologia qui proposta, basata sull'uso integrato di dati ausiliari amministrativi e statistici, consente di valutare l'effetto provocato dagli errori di lista sulle stime finali. Si riportano inoltre i primi risultati dell'applicazione ai dati dell'indagine campionaria ISTAT sulle piccole e medie imprese.

**Keywords:** Frame Error, Administrative Data, Structural Business Statistics, Business Register.

## 1. Introduction

Structural Business Statistics (SBS) surveys are often the only data source used to get relevant economic information about enterprises. In particular, Small and Medium Enterprises sample survey (SMEs) is carried out by using the auxiliary information coming from the Business Register (BR) ASIA. This frame represents the population of interest for SMEs in the sample selection phase and in the final estimation procedure. Quality of statistics produced is then related to the quality of the BR. Anyway, frame and non response errors may occur as main sources of non sampling errors. Their impact on final estimates is not easy to measure. Frame errors are identified as errors in coverage and may lead to bias and underestimation of variance. Non response errors, in particular *unit (total) non-responses*, have a similar effect on final estimates too. The aim of this paper is to identify and measure the impact of frame errors on the sample-based estimations using auxiliary administrative variables. Characteristics of administrative and statistical (SMEs) sources, of the estimation methodology and of simulation study are described in section 2. Empirical results are showed in section 3.

## 2. Data sources and methodology

Variable of interest for the analysis is “turnover”, as defined in the Commission Regulation EC (1998). The total amount of this variable is asked to the sampled SMEs enterprises by postal questionnaire (ISTAT, 2005-1). Turnover is strictly correlated ( $r=0.97$ ) to the variable “business fiscal turnover” (BFT) a benchmark that is provided, on a yearly basis, by the Fiscal Register, one of the main administrative sources that updates the BR (Garofalo, 2000). Checks on BFT are carried out in order to assure an acceptable level of quality. Coherence with respect to the SBS final data is also guaranteed. We are interested in the identification and assessment of the impact of frame errors on SMEs final estimates by including or not the non respondent units for frame errors into the SMEs final estimation process. This way it is possible to get an indirect measure of the BR quality. Frame errors and their impact on the overall error have been classified according to the following: i) *under-coverage*: BR does not reflect businesses within scope for the survey (estimates’ bias increases); ii) *over-coverage*: BR considers in scope businesses that are not (estimates’ bias increases; moreover ceased units consequently reduce sample size and increase sampling error). Each SMEs contacted unit provides direct information about the correctness of frame variables; these data are then encoded according to some modalities, among whom the following typologies, “Rejected, unknown, moved (blank questionnaires came back)” and “Ceased units, not active, in bankruptcy, ...” may be associated to a lack of quality in the BR (over-coverage). They will be treated as frame errors. A time lag exists between the reference year of the SMEs and that of the BR. For example, with reference to SMEs (year of reference  $t$ ) sample is selected at time  $t+1$  from BR (updated at time  $t-1$ ). After data collection, estimation is carried out at time  $t+2$  using BR updated at time  $t$ . The lag between BR and SMEs estimates is zero (year of reference is  $t$  for both data sources).

Sampling design of the SMEs is a one stage stratified simple random sampling. Population of interest is represented by active enterprises in the BR with less than 100 persons employed ( $N \approx 3,723,000$  units for year of reference 2002). They cover 99.8% of the total enterprises and 73.0% of the total employment. Strata are defined according to the NACE Rev.1.1 economic activity, size class and region. The sample size is  $n \approx 122,000$  units that are selected according to multi-variable and multi-domain sample allocation (Bethel, 1989) and negative coordination with previous SME surveys (Ohlsson, 1995). Correction factors for initial sampling weights because of unit non-response and under-coverage are then calculated. The calibration estimators methodology (Deville and Särndal, 1992; Laiho, 2002), implemented in SAS (ISTAT, 2005-2), is used for the estimation process. This methodology may be applied because auxiliary variables are strictly correlated to the variables of interest. The estimates of auxiliary variables equal the known totals of such variables (number of enterprises and number of persons employed) and limit the bias effect due to the non-response and under-coverage phenomena. Final sampling weights are then obtained as the result of the product between initial weights and post-stratification correction factors.

## 3. Empirical results

Let  $A$  be the set of respondent units to SMEs ( $A \neq \emptyset$ ), and  $B \equiv A \cup F$ , ( $A \cap F \equiv \emptyset$ ) the data set of respondent units to SMEs ( $A$ ) and non-respondent units for frame errors

to SMEs ( $F$ ). Let  $n_A$ ,  $n_B$ ,  $\tilde{Y}_A$ ,  $\tilde{Y}_B$ ,  $Var(\tilde{Y}_A)$  and  $Var(\tilde{Y}_B)$  be respectively sizes, calibration estimates and sampling variances of the total BFT for  $A$  and  $B$ . Bias of  $\tilde{Y}_A$  and bias of  $\tilde{Y}_B$  are estimated with respect to  $Y$ , the true value in the population  $P$  (represented by the BR). Relations between the sample size increment  $[(n_B/n_A) - 1] \cdot 100$ , estimated coefficients of variation ( $C\tilde{V}$ ) and mean square errors ( $MSE$ ) are then detected. Results show that for 59.4% of the NACE Rev. 1.1 groups considered, sample size increment (i.e.: from  $n_A = 53,803$  to  $n_B = 53,803 + 6,236 = 60,039$ ) generates a better approximation of the estimate to the true value (bias reduction). The variation of  $\tilde{Y}_A$  and  $\tilde{Y}_B$  with respect to  $Y$  is obtained according to the following:

$$\tilde{B}_A = 100(\tilde{Y}_A Y^{-1} - 1) \quad \tilde{B}_B = 100(\tilde{Y}_B Y^{-1} - 1)$$

As it is shown in table 1,  $\tilde{Y}_B$  is 0.5% lower than  $\tilde{Y}_A$  for the overall BFT total. Estimates look better, both with reference to the economic activity sector and size class, if non-respondent units for frame errors are included into the estimation.

**Table 1:** Total BFT in population, estimates and differences by activity or size class

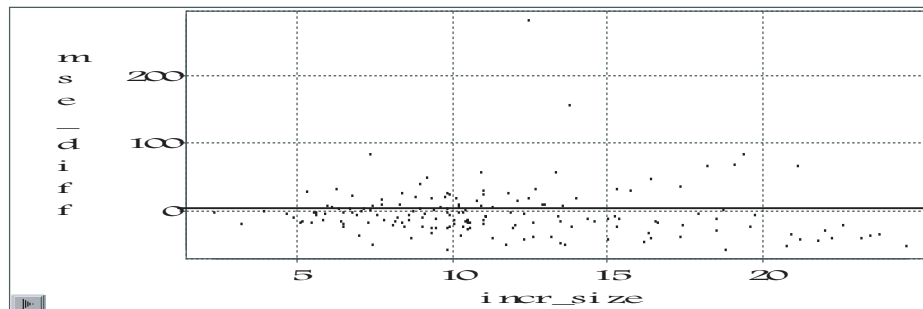
| Activity sector | $Y \cdot 10^{11}$ | $\tilde{Y}_A \cdot 10^{11}$ | $\tilde{Y}_B \cdot 10^{11}$ | $\tilde{B}_A$ | $\tilde{B}_B$ |
|-----------------|-------------------|-----------------------------|-----------------------------|---------------|---------------|
| Industry        | 4.072             | 4.020                       | 4.025                       | -1.3          | -1.1          |
| Construction    | 1.529             | 1.722                       | 1.721                       | 12.7          | 12.6          |
| Services        | 8.234             | 8.803                       | 8.726                       | 6.9           | 6.0           |
| Size class      |                   |                             |                             |               |               |
| 1               | 2.205             | 2.620                       | 2.580                       | 18.8          | 17.0          |
| 2-9             | 4.662             | 5.131                       | 4.994                       | 10.1          | 7.1           |
| 10-99           | 6.968             | 6.795                       | 6.898                       | -2.5          | -1.0          |
| Total           | 13.835            | 14.545                      | 14.472                      | 5.1           | 4.6           |

Let  $C\tilde{V}(\tilde{Y}_A) = 100\sqrt{Var(\tilde{Y}_A)}/\tilde{Y}_A$  and  $C\tilde{V}(\tilde{Y}_B) = 100\sqrt{Var(\tilde{Y}_B)}/\tilde{Y}_B$  be the coefficients of variation for  $\tilde{Y}_A$  and  $\tilde{Y}_B$ . With reference to the total of enterprises, for 82.4% of NACE groups,  $C\tilde{V}(\tilde{Y}_B) < C\tilde{V}(\tilde{Y}_A)$  after the increment in the sample size, so the quality of the BFT estimate is improved. A measure of the overall error is given by the MSE (mean square error) defined as:

$$MSE = (bias)^2 + (sampling\ error)^2$$

In figure 1 it is shown the relation between sample size increase  $[(n_B/n_A) - 1] \cdot 100$  and  $mse\_diff = [(MSE_B/MSE_A) - 1] \cdot 100$ . Some MSE reduction consequent to the increment of the sample size is shown by 58% of NACE groups. When  $mse\_diff < 0$ , then the inclusion of the frame error units in the respondent data set brings to more accurate results. If included in the estimation process, the non-response units for frame errors contribute to improvement of the overall error.

**Figure 1:** Relation between sample size increment (%) and MSE variation (%)



#### 4. Conclusions

The impact of frame errors on sample-based estimates of business fiscal turnover is measured by the MSE under the hypotheses that frame errors units are included or not in the estimation process. The increment of sample size due to the inclusion of frame errors units is 11.6%. Its impact on the final estimates mainly depends on the frequency distributions of respondent units by size class. Frame errors units are for the most part present in the smallest size class (47.7% of frame error units have 1 person employed whereas in the respondent data set this percentage is 30.9%). Results show that both estimates overestimate the population total of the variable of interest. Moreover, if frame error units are considered for the estimation, a reduction in the estimate of total occurs. Also CVs decrease even if in a limited way. So, MSE is generally lower when frame errors units are included in the estimation. In conclusion, the negative impact of the presence of frame error on final estimates is almost totally shown by bias increment.

#### References

- Bethel J. (1989) Sample Allocation in Multivariate Surveys, *Survey Methodology*, 15, 47-57.
- Commission Regulation EC (1998) No 2700/98 of 17 December 1998 concerning the definitions of characteristics for structural business statistics. *Official Journal of European Communities* No L 344, 18.12.98.
- Deville J.-C., Särndal C. E. (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 87, 376-382.
- Garofalo G. (2000) Progress Report: The Statistical Information System on Enterprises and Institutions, *14th International Roundtable on Business Survey Frames*.
- ISTAT (2005-1) *Conti economici delle imprese – Anno 2000*, Informazioni, 6-2005.
- ISTAT (2005-2) *GENeralised software for Sampling Estimates and Errors in Surveys, Funzione Riponderazione V.3.0*, Tecniche e Strumenti, 2-2005.
- Laiho J. (2002) Estimating the impact of survey errors on survey estimates, in: *The International Conference on Improving Surveys*, Copenhagen.
- Ohlsson E. (1995) Coordination of Samples Using Permanent Random Numbers, in: *Business Survey Methods*, Brenda G. Cox et al. (Eds.), Wiley, 153-169.