# EXPERIMENTAL ANALYSIS IN THE ESTIMATION OF SBS VARIABLES FOR SMALL FIRMS BY USING ADMINISTRATIVE DATA

*M.C. Casciano, V. De Giorgi, F. Oropallo, G. Siesto*

## Abstract

*In last years the availability of large administrative data sources together with the purpose of reducing statistical burdens has encouraged the launch of various projects in restructuring the production of business statistics. The issue of missing response, the difficulty to handle with bias-effects in sample surveys as well has brought to exploit the possibilities offered by external sources as a benchmark and a way of statistical estimation. By using fiscal agency sources, together with financial statements, an experimental estimate of main Structural Business Statistics (such as turnover, value of production, intermediate costs, value added, labour costs and gross operating surplus) has been done. The study may be summarized in these groups of activities: first, the analysis of the population coverage, using classification variables, such as the typology of economic activity, the size, the legal type of; second, the comparisons of the meaning and the contents of the variables from fiscal questionnaires with the respective SBS variables and the reconstruction of summarized variables, via respondent unit; third, the imputation of the main variables, through the fiscal ones, also to the subset of non respondent units and the re-weighting of the final sample, thus obtaining a new (secondary) estimate for 2007. Finally, the primary (from sample weight calibration only) versus secondary estimates has permitted to evaluate an estimation discrepancy distinguishing a source effect from the non response effect and to measure the reduction of the sampling error.*

**Keywords**: Missing response, Integration, Administrative data, Calibration estimators, Bias Estimation, Sampling error

## Introduction

Within a general plan moving towards a modernization of the production of structural business statistics, Italian NSI decided to intensely use all available administrative sources, having relevant economic information, in the statistical production process in order to reduce statistical burden on enterprises and to enhance the statistical quality of surveys in terms of comparability with other sources and reduction of non response bias.

First, a review on the availability and on the quality of administrative sources has been done in order to use this information from a new point of view. Administrative data should be used not only for imputing missing non-response items and non-response units, but also for sampling designs and for an integrated survey system, in order to obtain coherent final estimates. The Statistical Business Register (SBR) is obviously correlated to this process by means of its role of both list frame for surveys and as a record linkage basis.

In this document, after the description of the Istat current process for Small and Medium-sized Enterprise sample survey (SME), in which only financial statements are used to impute non-responses, we describe a new integration process that uses other administrative sources as well, by focusing on the differences in estimation outcomes and distinguishing the sources of discrepancies.

## 1. The Small-Medium Enterprises Survey and the current sampling strategy

Small and Medium-sized Enterprises (SME) sample survey is carried out annually by sending a postal questionnaire with the purpose of investigating profit-and-loss account of enterprises with less than 100 persons employed, as requested by SBS EU Council Regulation n. 58/97 and 295/2008. The units involved in the survey have also the possibility to fill in an electronic questionnaire and trasmit it to Istat via web.

The survey covers enterprises belonging to the following economic activities according to the Nace Rev.1.1 classification:
- Sections C, D, E, F, G, H, I, J (division 67), K;
- Sections M, N and O for the enterprises operating in the private sector.
Main variables of interest asked to the SME sampled enterprises are Turnover, Value added at factor cost, Employment, Total purchases of goods and services, Personnel costs, Wages and salaries, Production value. They are also asked to specify their economic activity sector and geographical location in order to test the correctness of the frame with respect to these information. Totals of variables of interest are estimated with reference to three typology of domains of study.

## 1.1. Frame of interest

The frame for SME survey is represented by the Italian Statistical Business Register (hereafter Asia[1]). It results from the logical and physical combination of data from both statistical sources (surveys) and administrative sources (Tax Register, Register of Enterprises and Local Units, Social Security Register, Work Accident Insurance Register, Register of the Electric Power Board) treated with statistical methodologies. Variables in the register are both quantitative (Average number of employees in the year $t$-1, Number of employees in date 31/12/year $t$-1, Independent employment in date 31/12/year $t$-1, Number of enterprises) and qualitative (Geographical location, Economic activity according to Nace Rev.1.1- 4 digit). From the Fiscal Register is also provided the VAT Turnover, which represents a good proxy of the variable Turnover asked to the sampled enterprises by questionnaire.

The population of interest for SME sample surveys is about 4.5 millions active enterprises for the year of reference 2007.

The survey is launched in June of the year $t+1$ on the basis of the year $t$-1 BR year $t$ being the reference year of the survey. The updated frame is available for the estimation phase only 15 months after the end of year $t$. New enterprises (births) are not included in the BR with which the survey has been launched (year t-1), while they are surely present in the updated Asia of the year t. Errors in coverage of the BR with respect to new businesses may lead to estimates bias. Also errors due to BR lag have an impact on SME final estimates.

## 1.2. Sampling design (allocation and domain of estimates)

SME is a multi-purpose and multi-domain survey and it 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 Tables 1 and 2).

Table 1: *Types of SME Survey domains*

| 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

Table 2: *Definition of Size-classes of employment for domain DOM3 of SME Survey*

| Nace Rev.1.1 2-digit level | Size-classes of employment |
|---|---|
| 10-45; | 1-9; 10-19; 20-49; 50-99; |
| 50-52; | 1; 2-9; 10-19; 20-49; 50-99; |
| 55;60-64;67;70-74; | 1; 2-9; 10-19; 20-49; 50-99; |
| 80; 85; 90; 92; 93; | 1-9; 10-19; 20-49; 50-99; |

Sampling design of the SME survey is a one stage stratified random sampling, with the strata defined by the combination of the modality of the characters Nace Rev.1.1 economic activity, size class and administrative region. A fixed number of enterprises is selected in each stratum without replacement and

---

[1] Archivio Statistico delle Imprese Attive - Statistical Register of Active Enterprises

with equal probabilities. The number of units to be selected in each stratum is defined as a solution of a linear integer problem (Bethel, 1989):

In particular, the minimum sample size is determined in order to ensure that the variance of sampling estimates of the variable of interest in each domain does not exceed a given threshold, in terms of coefficient of variation. In this way, about 103,000 of small and medium-sized enterprises (units) are included in the sample. The sampling units are drawn by applying JALES procedure (Ohlsson, 1995), in order to take under control the *total statistical burden*, by achieving a negative coordination among samples drawn from the same selection register.

## 1.3. The Total Missing Response

In SME survey of the reference year 2007 about 37,000 questionnaires were filled in by enterprises. That corresponds to a response rate of approximately 42% in terms of reliable replies (excluding non contacted units, out of coverage and list errors).

Actions to speed up or increase the response rate have been adopted: enterprises on delay are subjected to one reminder by post and one by phone.

The survey data have been integrated with administrative ones in the 20-99 size class for about 6,300 units, by using financial statements. In this way the estimates have been calculated on the basis of 43.701 units.

Data integration for each total missing response is done by following these steps:

° selecting randomly a donor enterprise with the same principal activity (Nace Rev.1.1 4-digit), size class and geographical area as the non-respondent unit to be imputed;

° calculating the donor per-head values;

° multipling the values obtained by the number of persons employed (as resulting from the updated frame) of the missing enterprise;

° substituting the main donor items with the actual values from the financial statement of the non respondent, and then recalculating some detailed value by pro-quota.

An enterprise can be used as a donor for not more than 5 times; if there is not any available donor, the constraints on geographic location and Nace Rev.1.1 may eventually be relaxed (from 4 to 3 digit).

**Table 3: *Sample units and respondents by sector of activity and size class***

| Sector | n | m | TMR* |
|---|---|---|---|
| Mining | 1.027 | 425 | 61,5 |
| Manufacturing | 39.215 | 16.845 | 60,6 |
| Energy | 1.130 | 540 | 50,8 |
| Construction | 4.882 | 2.066 | 59,2 |
| Trade | 18.344 | 8.400 | 54,2 |
| Hotel, Restaurant | 2.960 | 1.066 | 62,8 |
| Transport | 7.142 | 2.530 | 61,7 |
| Financial services | 1.453 | 598 | 50,8 |
| Business services | 16.483 | 7.202 | 53,2 |
| Social services | 10.403 | 4.029 | 58,6 |
| **Size class** | **n** | **m** | **TMR** |
| 1-9 | 67.058 | 24.778 | 57,5 |
| 10-19 | 17.085 | 5.983 | 56,0 |
| 20-49 | 12.751 | 8.872 | 60,5 |
| 50-99 | 6.145 | 4.068 | 62,7 |
| **Total** | **103039** | **43701** | **57,9** |

## 1.4. The weighting procedure

Correction factors for initial sampling weights for unit non-response and under-coverage are calculated in the estimation phase by applying the methodology based on calibration estimators (Deville and Sarndal, 1992).

The estimator of the total *Y(D)* referred to the domain *D* is:

$$\tilde{Y}_{(D)} = \sum_{k \in s_r} w_k \ y_k \ I_k(D)$$

where $s_r$ is the set of respondent units (respondent and imputed); $k$ is the unit index, $w_k$ is the final weight, $y_k$ is the observed (or imputed) value of the variables of interest; $I_k(D)$ equals 1 if the unit $k$ belongs to domain $D$, and 0 otherwise

The final weight $w_k$ is obtained as a product of three factors:

$$w_k = d_k \gamma_{1,k} \gamma_{2,k}$$

where

- ◦ $d_k$ is the direct weight (the reciprocal of the inclusion probability);
- ◦ $\gamma_{1,k}$ is the total non-response correcting factor ;
- ◦ $\gamma_{2,k}$ is the "post-stratification" factor.

After calculating the total non response correcting factors as the ratio of the number of sampled units and the number of respondent units belonging to appropriate "weighting adjustment cells", the weight of every single enterprise is furtherly modified in order to match known or alternatively estimated population totals called benchmarks. In particular, known totals of selected auxiliary variables on Asia Register (Average number of employees in the year $t-1$, Number of enterprises) are currently used to correct for sample-survey nonresponse or for coverage error resulting from frame undercoverage or unit duplication.

## 2. The matching between SME survey sample units and the Fiscal Authority survey units

### 2.1. Description of administrative sources used in the process

The sources used in this experimental analysis are the Financial statements and the Tax Authority sources (Tax returns forms and one Fiscal Authority survey) all linked with the BR identifying code. The whole enterprise population of the BR is about 4.5 million of enterprises which employ approximately 17.6 million persons. Only a part of it, the companies, is liable to fill in the financial statement: they are less than 20%, although they are about 57% in terms of persons employed. This source is the best harmonized with the SBS Regulation definitions. All other enterprises are obliged to declare their taxable income to the Fiscal Authority by filling in tax forms. Based on their legal type and accountancy regime, enterprises have to fill in different types of tax forms. According to the simplified accountancy regime, sole proprietorships (PF) have to fill in either the Pf-Re, if they are freelances, or the PF-RG form, if they are firms in a simplified accounting regime; the unincorporated firms (SP) are liable to fill in the SP-RG form, and the corporate ones (SC) have to compile the SC-RS.

Besides tax returns and financial statements data, Istat acquires directly from the Tax Authority the Sector Studies survey (Fiscal Authority Survey) source: it is a fiscal survey aiming to evaluate the capacity of enterprises to produce income and to know whether they pay taxes correctly. In spite of some exclusion and non-enforceability principles, almost all enterprises are obliged to fill in the Sector Studies survey form together with the tax return one and to declare in detail costs and income items. As the common part of all sector studies questionnaires is like a financial statement, it can be used in a more effective way than tax return data.

Since we have different types of sources to be used for recovering information about non-respondents units of the initial sample, we have to determine priorities in using only one of them to impute each unit. Based on the universe coverage, the number of available comparable variables, and the coherence to the SME survey variables in terms of number of effective Kolmogorov-Smirnoff tests, we decide to use the priority scheme defined in Table 4.

Table 4: *Kolmogorov Smirnoff test satisfied on comparable variables by sources*

|  | Comparable | Test KS |
|---|---|---|
| Financial Statements | 21 | 13 |
| Fiscal Authority Survey | 15 | 8 |
| Tax Return data - PF-RE | 13 | 6 |
| Tax Return data - PF-RG | 14 | 6 |
| Tax Return data - SP-RG | 14 | 6 |
| Tax Return data - SC-RS | 16 | 2 |

So, we use first the financial statement source (because of its almost total coverage of corporate firms, and the greatest number of variables comparable with SME survey, 13 out of 21 with a similar distribution), then the Sector Studies survey (15 comparable variables, 8 of them with a similar distribution), and last the tax return data (6 coherent variables), leaving the SC-RS variables in case of no other available data.

## 2.2. The coverage of available Administrative data

After the analysis of coherence among subsidiary sources (Table 4), we can outline the coverage of the 2007 SME survey sample in terms of number of units and of their information contents. Unless coverage list errors, Financial statements and Sector Studies together with Tax return modules, cover almost all sample enterprises: what remains are only the large and very small sole proprietorships. The large ones (with an ordinary accountant regime) are asked to fill the RF form of Tax return module which is not comparable with the profit and loss scheme. The very small ones, called minimum taxpayer, only from 2008 are liable to compile a special tax return form named CM.

**Figure 1**: *Coverage analysis by legal type and size class*
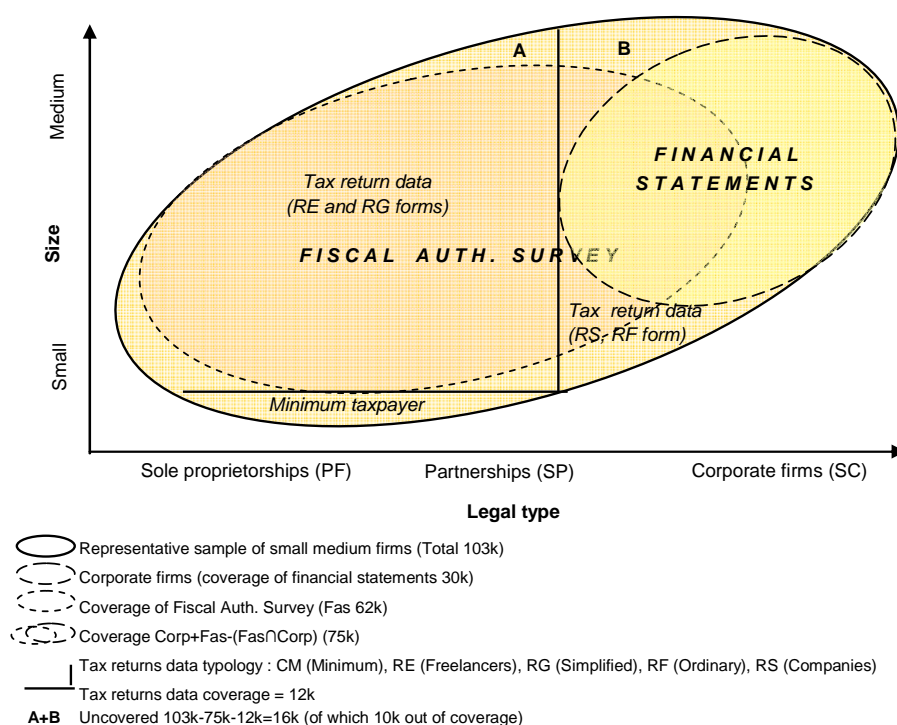


In Table 5, the sample coverage figures are showed in terms of the administrative source used, by following the priority rule defined before.

**Table 5**: *Coverage of the initial sample by type of response and administrative data*

| Source | Initial theoretical sample | | |
| | Non respondents | Respondents | Total |
| --- | --- | --- | --- |
| Financial Statements | 10.370 | 19.739 | 30.109 |
| Fiscal Authority Survey (F) | 24.655 | 17.798 | 42.453 |
| Fiscal Authority Survey (G) | 1.343 | 1.223 | 2.566 |
| Tax Return data - PF-RG | 2.312 | 990 | 3.302 |
| Tax Return data - PF-RE | 747 | 483 | 1.230 |
| Tax Return data - SP-RG | 810 | 378 | 1.188 |
| Tax Return data - SC-RS | 4.546 | 1.839 | 6.385 |
| From survey only | - | 1.251 | 1.251 |
| Total | 44.783 | 43.701 | 88.484 |
| | | | |
| Out of coverage and list errors | | | 10.218 |
| No sources | | | 4.337 |
| Total sample units | | | 103.039 |

5

If we do not take into account list errors and units out of SME frame, the total coverage is about 95%, half from respondent units and half from administrative sources.

Although we have a very high coverage level, some problems about the total number of variables to be used and the definitions harmonization among sources occur.

## 2.3 The information contents and harmonization among sources

The variables common to all sources (Financial Statement, Sector Studies survey, Tax Return data) are: Income from sales and Services (Turnover), Changes in stock, Changes in contract work in progress, Other income and earnings (neither financial, nor extraordinary), Purchases, Purchases of goods and services, Services (Total), Use of third party assets, Value adjustments, Fund allocations, Other operating charges, Personnel costs. Moreover there are two further variables, Value added and Gross operating value that can be calculated with the previous ones.

The variables content comparability has been assessed by comparing both their definitions and values in frequency distribution with survey variables. For instance, once assessed the contents are defined in a similar way, we can draw the distribution of the difference ranges between sample survey variables and source variables (Financial Statements, Fiscal Authority Survey, Tax Return PF and Tax Return SP), as it is showed in Figure 2 for Turnover .

**Figure 2**: *Distribution of respondents units linked with administrative data by range of differences for Turnover*
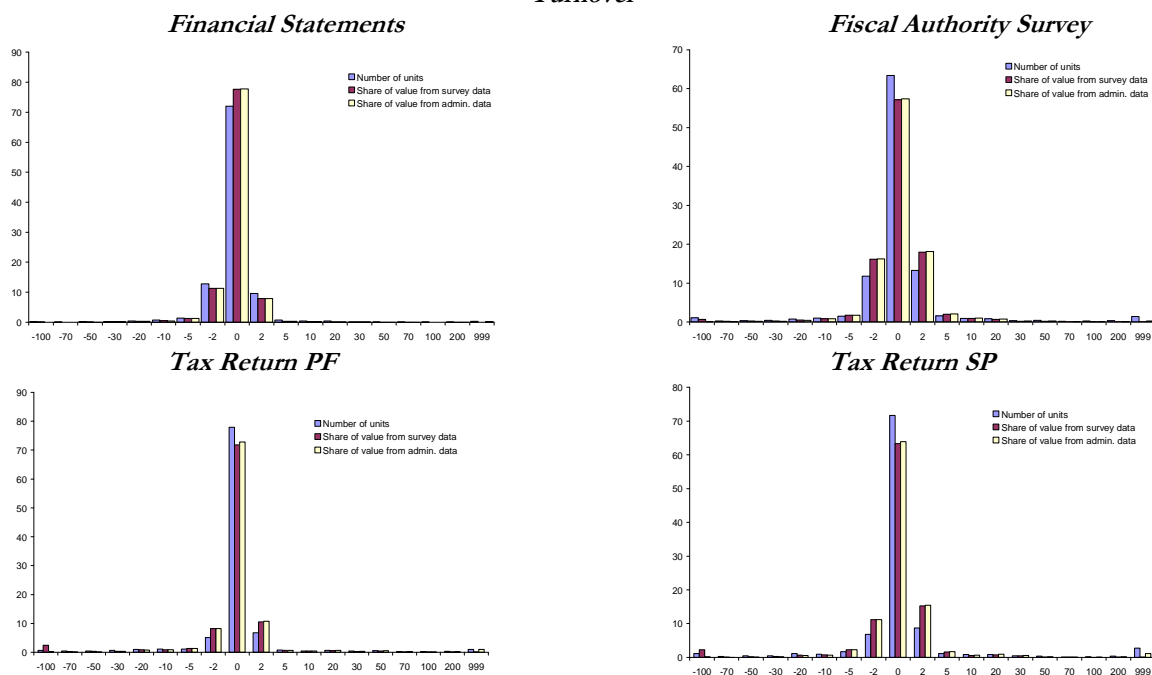


The distribution is 0-balanced and positive asymmetric, which means that the source and the survey variables correspond, although the administrative one tends to be higher than the correspondent in the suvey.

# 3. Results of integration of SME survey with Administrative data and re-estimation

## 3.1. Imputation of TRM through administrative data

Administrative sources permit to cover almost all sample units of the SME survey, so it has been possible to extend the reconstruction of SBS variables both for respondent units or non respondent ones.

The initial survey sample has been distinguished, on the y-axis, among respondents (47%) and non respondents (53%). The x-axis represents the dimension of the information content. Through

administrative data (Tax returns, Fiscal survey and Financial statements) we cover only a part of information contained in the survey scheme. For this part it is possible to compare SBS variables from survey data with SBS reconstructed variables from fiscal and financial data.

**Figure 3 – *Integration scheme among Administrative and Survey data (Year 2007)***
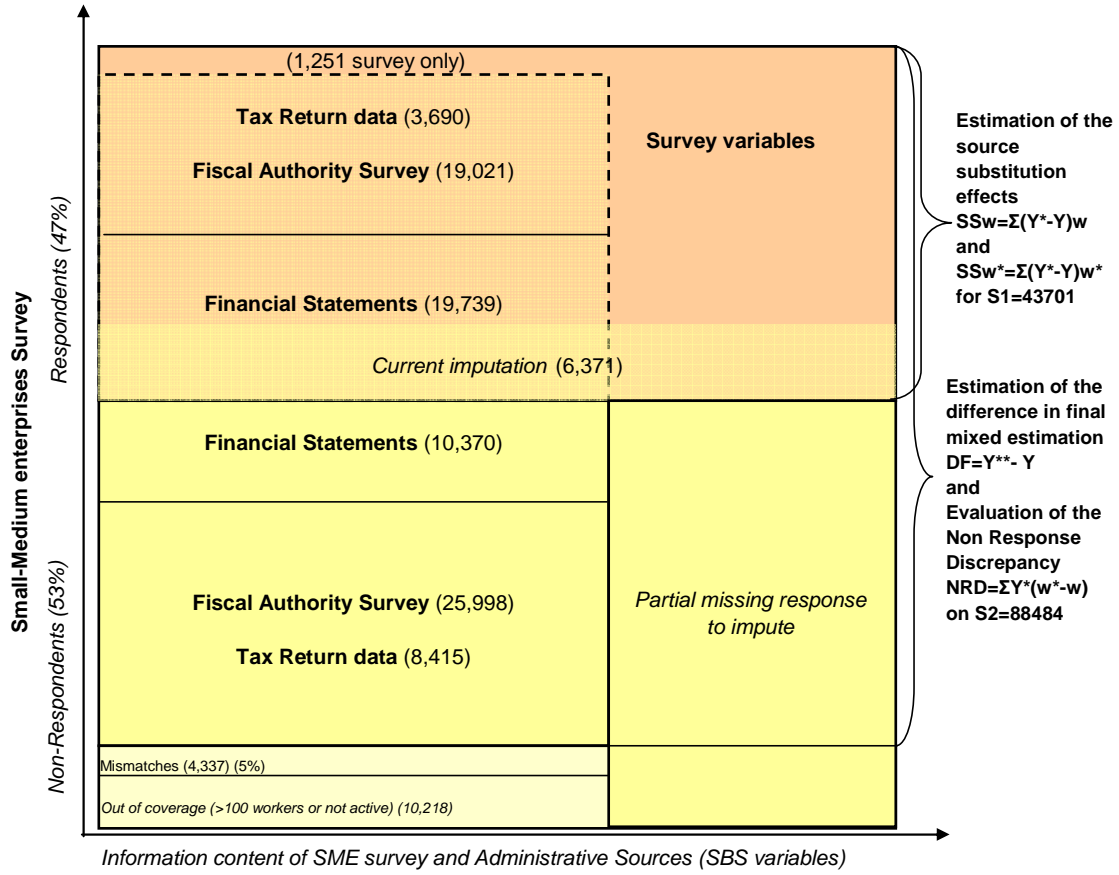


The initial estimates, based on the subset of respondents (S1) is:

$$\tilde{Y}_\alpha = \sum_{S1} y_k \, w_k$$

Final estimates on the integrated sample (S2), with a new set of final weights, is:

$$\tilde{Y}_\alpha^* = \sum_{S2} y_k^* \, w_k^*$$

Final mixed estimate on the integrated sample (S2) is:

$$\tilde{Y}_\alpha^{**} = \sum_{S1} y_k \, w_k^* + \sum_{S2-S1} y_k^* \, w_k^*$$

That can be written as:

$$\tilde{Y}_\alpha^{**} = \sum_{S2} y_k^* \, w_k^* + \sum_{S1} y_k \, w_k^* - \sum_{S1} y_k^* \, w_k^* = \tilde{Y}_\alpha^* - \sum_{S1} (y_k^* - y_k) w_k^*$$

In this way the mixed (survey/administrative) estimate is equal to the estimate based on administrative data with new weights minus the weighted discrepancy among sources.

## 3.2. Calculation of the estimation discrepancy

Starting from the previous formulas, we can calculated various component of the final estimation difference.

The difference between the administrative based and survey based estimate is equal to:

$$\tilde{Y}_\alpha^* - \tilde{Y}_\alpha = \sum\nolimits_{S2} y_k^* \, w_k^* - \sum\nolimits_{S1} y_k \, w_k$$

Where $\tilde{Y}_\alpha^*$ is the new variable calculated with the administrative source and $w_k^*$ the new vector of final weighs is obtained after the new calibration procedure: $w_k^* = d_k \, \gamma_{1,k}^* \, \gamma_{2,k}^*$

This new version of final weights fundamentally reduce the role of the adjustment for total non responds: $\gamma_1^* < \gamma_1$. On the overall initial sample of roughly 93 thousands units, the new respondents are roughly 88,5 thousands unit, with an average correction factor $\gamma_1^* = 1.05$. Instead of the previous correction factor $\gamma_1 = 2.12$.

If we add $\sum\nolimits_{S1} y_k^* \, w_k$ and subtract $\sum\nolimits_{S2} y_k^* \, w_k$ in the difference formula, where $w_k$ is zero for all units of S2 not included in S1, we obtain:

$$\tilde{Y}_\alpha^* - \tilde{Y}_\alpha = \sum\nolimits_{S2} y_k^* \, w_k^* - \sum\nolimits_{S1} y_k \, w_k + \sum\nolimits_{S1} y_k^* \, w_k - \sum\nolimits_{S2} y_k^* \, w_k^*$$

That can be grouped in the following two components:

$$\tilde{Y}_\alpha^* - \tilde{Y}_\alpha = \sum\nolimits_{S1} \left( y_k^* - y_k \right) w_k + \sum\nolimits_{S2} y_k^* \, (w_k^* - w_k)$$

Moreover if we consider the mixed estimates $\tilde{Y}_\alpha^{**}$, as the final new estimates, we can calculate the total difference between initial and final mixed estimates. So, considering that:

$$\tilde{Y}_\alpha^{**} = \tilde{Y}_\alpha^* - \sum\nolimits_{S1} ( y_k^* - y_k ) w_k^*$$

We can introduce another component in the final difference, that is the "source substitution effect" evaluated with the new weights : $\sum\nolimits_{S1} \left( y_k^* - y_k \right) w_k^*$

$$\tilde{Y}_\alpha^{**} - \tilde{Y}_\alpha = \tilde{Y}_\alpha^* - \tilde{Y}_\alpha - \sum\nolimits_{S1} ( y_k^* - y_k ) w_k^*$$

In this way the difference (DIFF) is equal to:

$$DIFF = \tilde{Y}_\alpha^{**} - \tilde{Y}_\alpha = \sum\nolimits_{S1} \left( y_k^* - y_k \right) w_k - \sum\nolimits_{S1} ( y_k^* - y_k ) w_k^* + \sum\nolimits_{S2} y_k^* \, (w_k^* - w_k)$$
$$DIFF = SSw - SSw* + NRD$$

in which we can distinguish three sources of differences due to:

- The source substitution effect for S1 (with old weights) $SSw = \sum\nolimits_{S1} \left( y_k^* - y_k \right) w_k$
- The source substitution effect for S1 (with new weights): $SSw* = \sum\nolimits_{S1} \left( y_k^* - y_k \right) w_k^*$
- The non response effect for S2: $NRD = \sum\nolimits_{S2} y_k^* \, (w_k^* - w_k)$

### 3.3 Measurement of coherence among different sources and evaluation of the TMR effect

The analyses here described have been carried out for the variable turnover. Differences between survey estimates and the new ones are showed in Table 6.

**Table 6 -** *Turnover estimates, analysis of the differences between sources and evaluation of the TMR effect by sectors of activity and size class – Year 2007*

**%DIFF - Total difference in final estimates (Y**-Y)/Y%**

| Sectors | 1-9 | 10-19 | 20-49 | 50-99 | Total |
|---|---|---|---|---|---|
| Industry | 4.33 | -0.89 | -0.66 | 3.13 | **1.36** |
| Constructions | -6.92 | -7.86 | 16.01 | -3.48 | **-3.53** |
| Services | -0.81 | -0.04 | 0.31 | 3.99 | **0.06** |
| **Total** | **-0.96** | **-1.25** | **1.15** | **3.21** | **0.03** |

**SSw - Source substitution effect for S1 (with old weights)**

| Sectors | 1-9 | 10-19 | 20-49 | 50-99 | Total |
|---|---|---|---|---|---|
| Industry | -0.23 | -3.30 | -0.82 | -0.54 | **-1.16** |
| Constructions | -0.67 | -0.09 | -2.62 | 0.60 | **-0.76** |
| Services | -0.64 | -1.21 | -2.71 | -0.70 | **-1.09** |
| **Total** | **-0.59** | **-1.77** | **-1.89** | **-0.56** | **-1.07** |

**SSw* - Source substitution effect for S1 (with new weights)**

| Sectors | 1-9 | 10-19 | 20-49 | 50-99 | Total |
|---|---|---|---|---|---|
| Industry | -0.43 | -1.41 | -0.74 | -0.40 | **-0.73** |
| Construction | -0.46 | -0.23 | -3.29 | 0.67 | **-0.82** |
| Services | -0.03 | -0.93 | -2.15 | -0.59 | **-0.60** |
| **Total** | **-0.15** | **-1.02** | **-1.65** | **-0.44** | **-0.66** |

**NRD - Difference due to TMR**

| Sectors | 1-9 | 10-19 | 20-49 | 50-99 | Total |
|---|---|---|---|---|---|
| Industry | 4.13 | 1.00 | -0.58 | 3.27 | **1.80** |
| Constructions | -6.71 | -8.00 | 15.34 | -3.40 | **-3.59** |
| Services | -0.20 | 0.25 | 0.87 | 4.10 | **0.55** |
| **Total** | **-0.53** | **-0.50** | **1.39** | **3.33** | **0.44** |

The total difference is very close (+0,03%) to the initial estimate produced with an higher TMR, although there is an high variability in results when we breakdown economic activities and size classes. The source substitution effect is -1.07 (old weights) and -0.66 per cent (new weights) respect to the previous estimate and it is greater for small firms (10-19 and 20-49). In particular for industry with 10-19 persons employed and construction and service activities with 20-49 p.e. The difference due to the TMR is higher in construction activities and lower in service sector. For that non response has produced an higher estimate of turnover for micro and small firms (new weighs decrease of 0.5 per cent the previous estimates) and a strong under estimate for medium enterprises (3.33) especially of service sectors.

Finally, a simulation study has permitted us to evaluate a gain in the efficiency of the estimator. After 1000 replication of the estimation procedure, by using a jacknife technique that take the 75% of the sample of respondents, with the same stratification of the initial sample, we have calculated that the absolute relative bias (Arb) and the root mean square error (Rmse) are in both cases reduced of roughly 1 per cent.

**Table 7 –** *Measurement of the efficiency of new estimates after 1000 sampling replication*

| | Variable (Y=Turnover - S1) | | Variable (Y**=Turnover - S2) | |
|---|---|---|---|---|
| **Sector** | **ARB** | **RMSE** | **ARB** | **RMSE** |
| Mining | 7.30 | 8.54 | 11.62 | 13.17 |
| Manufacturing | 7.82 | 9.52 | 6.11 | 7.50 |
| Energy | 13.47 | 16.18 | 14.06 | 16.81 |
| Construction | 7.06 | 8.59 | 6.12 | 7.50 |
| Trade | 5.83 | 7.17 | 6.09 | 7.41 |
| Hotel, Restaurant | 9.49 | 11.65 | 5.41 | 6.70 |
| Transport | 15.06 | 18.12 | 14.63 | 17.34 |
| Financial services | 19.67 | 22.75 | 15.66 | 18.47 |
| Business services | 9.58 | 11.58 | 8.08 | 9.84 |
| Social services | 7.40 | 9.09 | 6.64 | 8.15 |
| | | | | |
| **Size class** | | | | |
| 1-9 | 7.73 | 9.38 | 6.08 | 7.42 |
| 10-19 | 11.06 | 13.34 | 8.22 | 9.97 |
| 20-49 | 7.58 | 9.17 | 7.86 | 9.44 |
| 50-99 | 10.12 | 12.29 | 10.11 | 12.13 |
| | | | | |
| **Total** | **9.11** | **11.03** | **8.07** | **9.74** |

# Conclusions

The unsatisfactory sampling survey response rate together with the availability of a huge amount of administrative information (Financial statement, Sector studies, Tax returns) has led the ongoing analyses to both have a better coverage of theoretical sample, by using administrative data in the estimation process, and evaluate discrepancies in estimates due to the missing responses. The experimental study reveals that, for turnover aggregate, administrative data produce coherent estimates to survey ones at a high level of aggregation. The discrepancy obtained by using the variable integrated with administrative information instead of the current SME survey variable is 0.03%, though it is higher when we drill down for economic activities and size classes. The effect of total missing response is higher in some sectors, and demonstrates that mechanism besides the TMR is not random but is concentrated in some particular subsets of the population (micro-small firms of services and constructions). The enlargement of the subset of sample units used in estimation would also generally decrease the errors of the estimators of 1% about. Next steps of this study are to produce more disaggregated analyses (Nace at 4 digits, Nace at 3 digits and size classes) and to extend to all other SBS variables. Besides, since only some SBS variables can be derived from administrative sources, it will be necessary to develop methodologies for estimating the variables not provided by administrative sources. It also needs to remark that for a complete use of administrative sources for statistical purpose, it is necessary that the administrative data flow is continuous and stable over time and on line with the timeliness required by the statistical process to meet the deadlines of the SBS regulation. In addition, the administrative data check, edit and imputation phases of the statistical process should be implemented. Finally it will be important that Istat has a more active role in the design of administrative forms, including variables useful for statistical purposes.

# References

Bethel, J. (1989) "Sample allocation in multivariate surveys". *Survey methodology*, 15 (1989): 47-57.

Casciano, C., e P.D. Falorsi e S. Filiberti e A. Pavone e G. Siesto (2006). "Principi e metodi per il calcolo delle stime finali e la presentazione sintetica degli errori di campionamento nell'ambito delle rilevazioni strutturali sulle imprese". Rivista di Statistica Ufficiale, n. 1 (2006): 67-102.

Commission Regulation on Statistical Units: n. 2186/93 and n. 696/93; on Structural Business Statistics: n. 58/97, 410/98, 2700/98, 2056/02, 1670/03; on Evaluation of Quality of Structural Business Statistics: n. 1618/99.

Deville, J.C., Särndal, C.E. (1992) "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, 87 (1992): 376-382.

Eurostat, European Commission (1999) "Use of Administrative Sources for Business Statistics Purpose" Handbook on good practices 1999 Edition.

Eurostat, European Commission (2003) "Manual for Structural Business Statistics" Directorate D: Business statistics - Unit D-2: Structural business statistics - July 2003

Falorsi, P.D., Ballin, M., De Vitiis, C., Scepi, G. (1998) "Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'Istat". *Statistica Applicata*, 20, n. 2 (1998).

ISTAT, (2009), Rapporto Annuale 2008, Roma.

ISTAT, (2007), Conti economici delle imprese - Anno 2003.

Oropallo, F., Inglese, F. (2004), The Development of an Integrated and Systematized Information System for Economic and Policy Impact Analysis", *The Austrian Journal of Statistics* Vol 33/2004 N. 1&2

Ohlsson, E. (1995), Coordination of PPS Samples Over Time, Stockholm University Mathematical Statistics, Stockholm University, S–106 91 Stockholm, Sweden

Särndal, C.E., Swensson, B., Wretman, J. (1992) "Model Assisted Survey Sampling". New York: Springer-Verlag, 1992

Yung, W., Lys P. (2008) "Use of Administrative Data in Business Surveys - The Way Forward" Statistics Canada - IAOS Conference on Reshaping Official Statistics - Shanghai, 14-16 October 2008