Estimation of Structural Business Statistics for Small Firms by Using Administrative Data

Maria Cristina Casciano, Viviana De Giorgi, Filippo Oropallo e Giampiero Siesto[*]

The need of handling missing response and potential bias in the estimation of business statistics has brought to exploit new possibilities offered by administrative sources in the estimation of Structural Business Statistics (SBS) for small-medium enterprises. The joint use of fiscal agency sources and balance sheet data with survey data, has led to a more integrated approach in the production process. The purpose of this work is to describe the new integrated approach and of measuring the reduction of the sampling error with the new estimation process by distinguishing a source effect from the non response effect. The paper is articulated in the following steps: (i) analysis of the coverage of the survey units through administrative sources by stratification variables, such as the economic activity, the number of persons employed and the legal form; (ii) comparison of meaning and content of the variables from balance sheets and tax data with the corresponding SBS variables; (iii) imputation of unit non-response through administrative data and re-weighting of the final sample to obtain a new estimate for 2007; (iv) evaluation of the discrepancy between the old and the new estimates and evaluation of the non response effect.

Keywords: Unit non-response, Micro integration, Administrative sources, Calibration estimators, Sampling error.

Introduction[1]

Within a general plan moving towards a modernization of the production of structural business statistics at a European level, many National Statistical Institutes have indicated a willingness to increase the use of all available administrative sources, which have 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 missing response bias (Yung, 2008), (Eurostat, 1999).

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

[*] M. C. Casciano ricercatore (Istat), e-mail: casciano@istat.it; V. De Giorgi Ricercatore (Istat), e-mail: degiorgi@istat.it; F. Oropallo ricercatore (Istat), e-mail: oropallo@istat.it; G. Siesto ricercatore (Istat), e-mail: siesto@istat.it.

[1] We are grateful for useful comments from an anonymous referee. A preliminary version of the paper was presented at the Seminar "Using Administrative Data in the Production of Business Statistics: Member States experiences", organized by the Eurostat ESSnet Project *AdminData*, held at Istat in Rome the 18th and 19th of March 2010.

This paper illustrates the statistical production process for Small and Medium-sized Enterprise sample survey (SME) until the reference year of 2007, in which balance sheets are only used for imputing non-response. Afterwards it shows the new integration process (tested on 2007 and adopted from 2008 onward) in which all administrative sources are combined with survey information. Finally it investigates, for 2007 data, the differences in estimation outcomes by distinguishing the sources of discrepancies.

1. The Small-Medium Enterprises Survey

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 (Eurostat, 2003) and n. 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, Purchases of goods and services, Personnel costs, Wages and salaries, Investments, Employments and other variables useful for calculating Value added at factor cost and 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 (BR). 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 persons employed in the year *t*-1, Number of persons employed in date 31/12/year *t*-1, Number of unpaid persons employed 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 reference year 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

BR 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 time 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) (Falorsi et al., 1998 Istat, 2009).

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) (a)	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

Source: Survey of Small-Medium Enterprises
(a) 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 (Number of persons employed)
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;

Source: Survey of Small-Medium Enterprises

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 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 unit non-response

In SME survey of the reference year 2007 about 37,000 questionnaires were filled in by enterprises. The response rate is 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 balance sheets. In this way the estimates have been calculated on the basis of 43.701 units (response rate of roughly 47%).

Data imputation for unit non responses is done as follows:

1. 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;
2. calculating the donor per-head values;
3. multipling the values obtained by the number of persons employed (as resulting from the updated frame) of the missing enterprise;
4. After the step number 3 the unit non-response dataset is linked with administrative source (balance sheet database) and the missing value (estimated by donor methods) is replaced with balance sheet value if available.

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 (Nace Rev.1.1 from 4 to 3 digit).

Table 3 - Sample units,* respondents and unit non-response rate by sector of activity and size class

	Sample units (n)	Respondents (m)	Unit non-response rate %
Economic activity			
Mining	927	425	54.2
Manufacturing	35372	16845	52.4
Energy	1013	540	46.7
Construction	4447	2066	53.5
Trade	16995	8400	50.6
Hotel, Restaurant	2586	1066	58.8
Transport	6107	2530	58.6
Financial services	1328	598	55.0
Business services	14967	7202	51.9
Social services	9079	4029	55.6
Size class (Number of persons employed)			
1-9	61480	24570	60.0
10-19	14541	6237	57.1
20-49	11720	8829	24.7
50-99	5080	4065	20.0
Total	**92821**	**43701**	**52.9**

Source: Survey of Small-Medium Enterprises

(*) Net of inactive units and list errors.

Table 3 shows the unit non-response rates by activity sector and by size class, that is equal to the difference between the sample units (n) and respondents (m) divided by the sample units (n). The missing response seems to be correlated to the enterprise size class and it is more concentrated in sectors such as Transport, Hotel and Restaurant and other service activities. So the use of traditional estimator could generate biases.

1.4 The weighting procedure

Correction factors for theoretical 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 Särndal, 1992).

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

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

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 further modified in order to match known or alternatively estimated population totals called benchmarks. In particular, known totals of selected auxiliary variables on BR (Average number of persons employed in the year t-1, Number of enterprises) are currently used to correct for sample survey non-response or for coverage error resulting from frame undercoverage or unit duplication (Casciano et al., 2006).

2. The matching between SME survey sample and administrative sources

2.1 The administrative sources used in the process

The sources used in this experimental analysis are the balance sheets and the tax revenue sources (Sector Studies and Tax returns) all linked with the BR identifying code (Bernardi et al., 2010). The whole population of the BR is about 4.5 million of enterprises which employ approximately 17.6 million average annual workers. Only a part of it, the companies, is liable to fill in the balance sheet: they are less than 20%, although they represent about 57% of persons employed. This source is the best one 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 form and of the 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 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) must fill in the SP-RG form, and the corporate ones (SC) have to compile the SC-RS.

Besides tax return and balance sheet sources, Istat also acquires, directly from the Tax Authority, the Sector Studies (Fiscal Authority Survey): 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 questionnaire 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 balance sheet, it can be used in a more effective way than tax return data.

Since different types of data sources have been used for recovering information about non-respondents units of the initial sample, it has been necessary to determine priorities in using only one of them in the imputation process. For that, it has been defined a ranking priority among the different sources, shown in Table 4, based on the number of available comparable variables and on the coherence to the SME survey variables in terms of number of effective Kolmogorov-Smirnov tests (KS), that have been made on the distribution of similar variables across different sources.

Table 4 - Comparable variables and Kolmogorov-Smirnov test by sources

ADMINISTRATIVE SOURCE	Comparable variables	Test KS (variables with similar distributions)
Balance sheets	21	13
Sector Studies	15	8
Tax Return - PF-RE	13	6
Tax Return - PF-RG	14	6
Tax Return - SP-RG	14	6
Tax Return - SC-RS	16	2

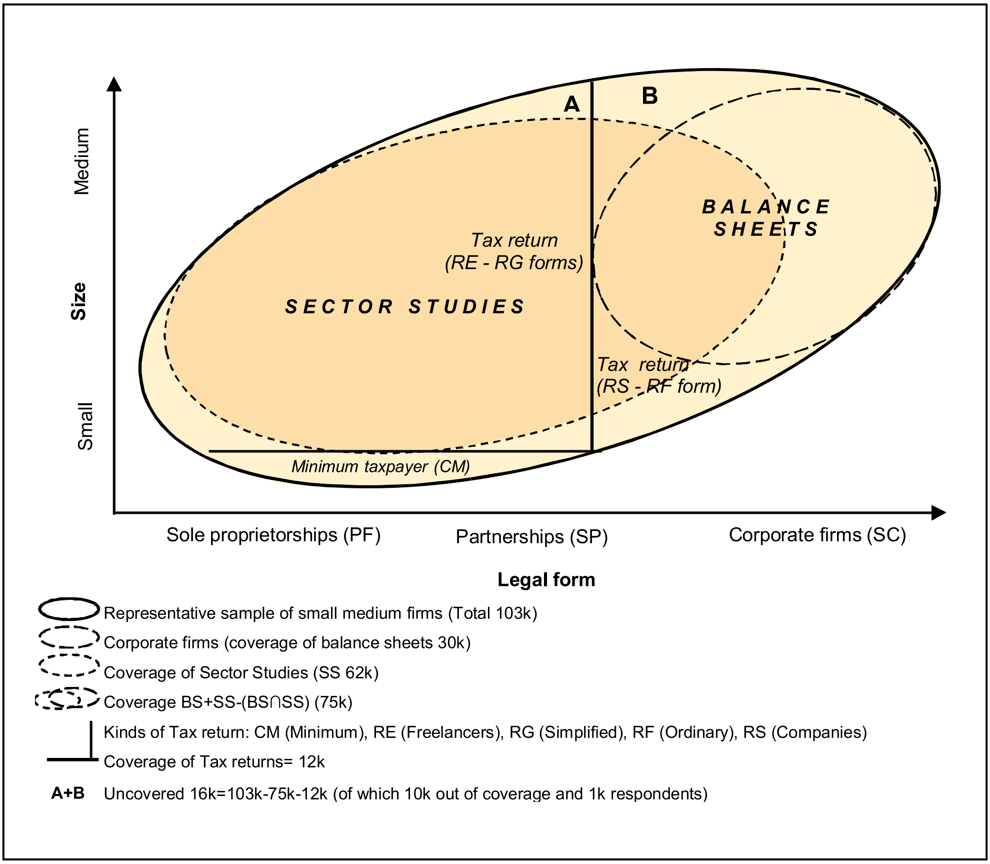
Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

High priority is given to balance sheets. Indeed, among 21 comparable variables, that refer to the same definition of SBS, 13 of them have a similar distribution. Then the Sector Studies survey (15 comparable variables, 8 of them with a similar distribution) and last the tax return source: 6 coherent variables for PF kinds and 2 coherent variables for SC kinds.

2.2 The coverage of available Administrative sources

After the coherence analysis among subsidiary sources (Table 4), it has been done an analysis of the coverage of the sample of the SME survey of 2007 in terms of number of units and of their information contents. Unless coverage list errors, balance sheets and Sectors studies together with Tax returns, cover almost all sampled enterprises. The coverage (Figure 1) of balance sheet units amounts to about 30 thousands units on a theoretical sample of 103 thousand. The additional coverage from Sector studies adds up roughly to 45 thousand and the supplementary coverage of tax returns data amounts to 12 thousand units. The residual, not covered, sample units (16 thousand) represent some large and very small sole proprietorships. The large ones (with an ordinary accountant regime) are obliged to fill the RF form of Tax return module which is not comparable with the profit & 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 form and size class of the enterprise - Year 2007


A
B
Medium
Size
Small
BALANCE SHEETS
Tax return (RE - RG forms)
SECTOR STUDIES
Tax return (RS - RF form)
Minimum taxpayer (CM)
Sole proprietorships (PF)
Partnerships (SP)
Corporate firms (SC)
Legal form
Representative sample of small medium firms (Total 103k)
Corporate firms (coverage of balance sheets 30k)
Coverage of Sector Studies (SS 62k)
Coverage BS+SS-(BS∩SS) (75k)
Kinds of Tax return: CM (Minimum), RE (Freelancers), RG (Simplified), RF (Ordinary), RS (Companies)
Coverage of Tax returns= 12k
A+B Uncovered 16k=103k-75k-12k (of which 10k out of coverage and 1k respondents)

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

Table 5 shows the sample coverage figures according to the administrative source used and following the priority rule defined before.

Table 5 - Coverage of the theoretical sample by kind of response and administrative source - Year 2007

ADMINISTRATIVE SOURCE	Sample of respondents	Unit non-response	Total sample
Balance sheets	19,739	10,370	30,109
Sector Studies (F)	17,798	24,655	42,453
Sector Studies (G)	1,223	1,343	2,566
Tax Return - PF-RG	990	2,312	3,302
Tax Return - PF-RE	483	747	1,230
Tax Return - SP-RG	378	810	1,188
Tax Return - SC-RS	1,839	4,546	6,385
From survey only	1,251	0	1,251
Total	**43,701**	**44,783**	**88,484**
No sources			4,337
Out of coverage and list errors			10,218
Total sample units			**103,039**

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

Unless of list errors the total coverage is about 95%, half from the sample of respondents and half from administrative sources.

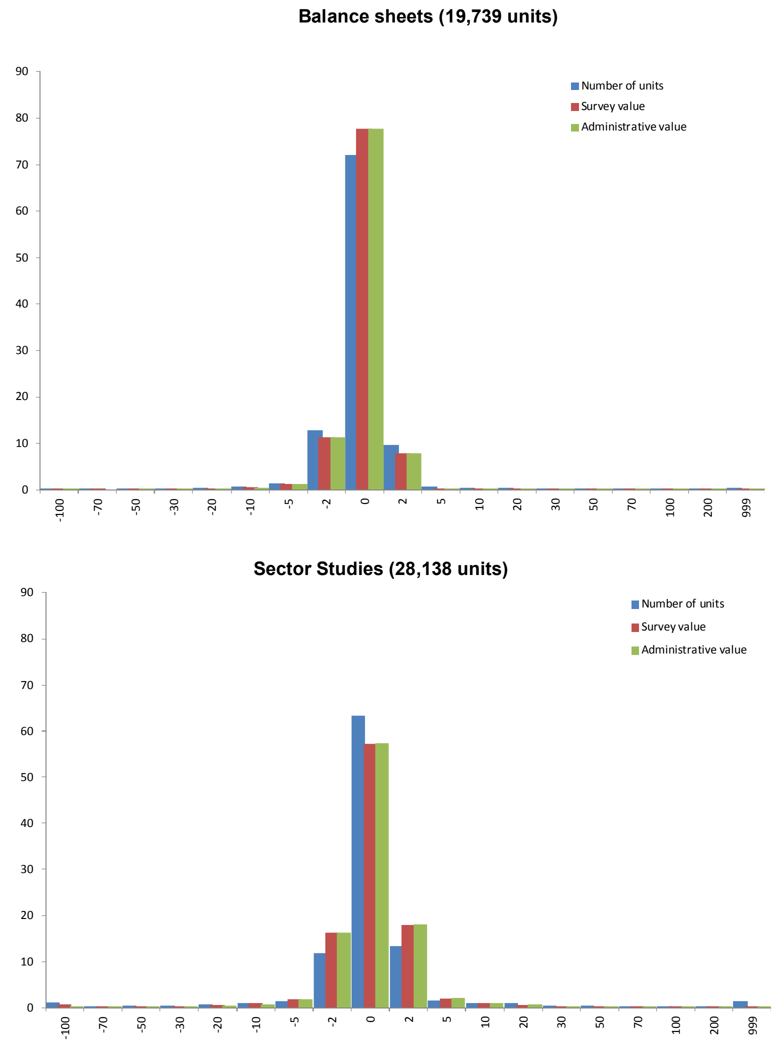
Despite of the very high level of coverage, several issues still remain about the total number of variables to be used and the harmonization of definitions among sources.

2.3 The information contents and harmonization among sources

The variables common to all sources (Balance sheets, Sector Studies survey, Tax Return data) are the following: 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.

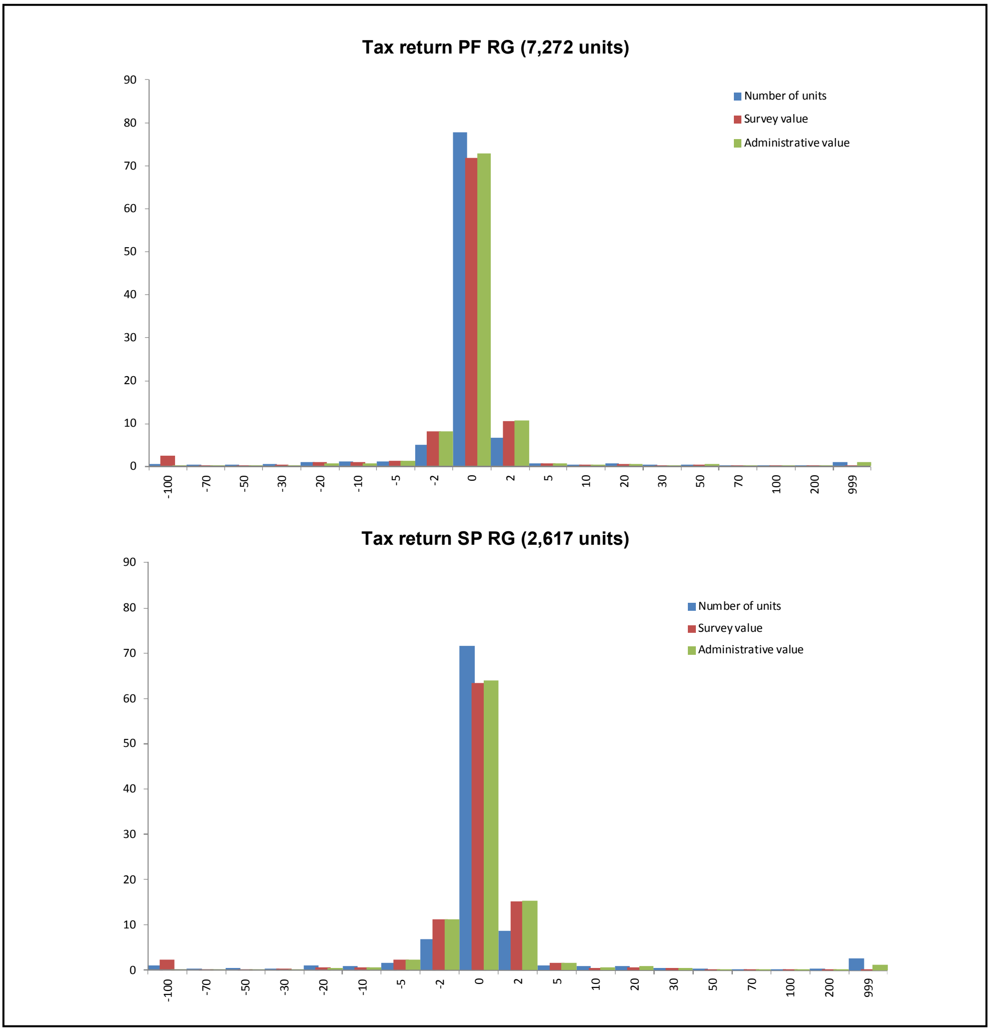
The variables content comparability has been assessed by comparing both their definitions and values in frequency distribution with survey variables (Oropallo and Inglese, 2004). For instance, once assessed the contents are defined in a similar way, the distribution by classes of differences between sample survey variables and source variables (Balance sheets, Sector Studies survey, Tax Return PF and Tax Return SP) can be drawn, as it is showed in Figure 2 for Turnover.

Figure 2 - Turnover, distribution of respondent units linked with administrative sources by classes of differences - Year 2007 *(percentage values)*



Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

Figure 2 segue - **Turnover, distribution of respondent units linked with administrative sources by classes of differences - Year 2007** *(percentage values)*



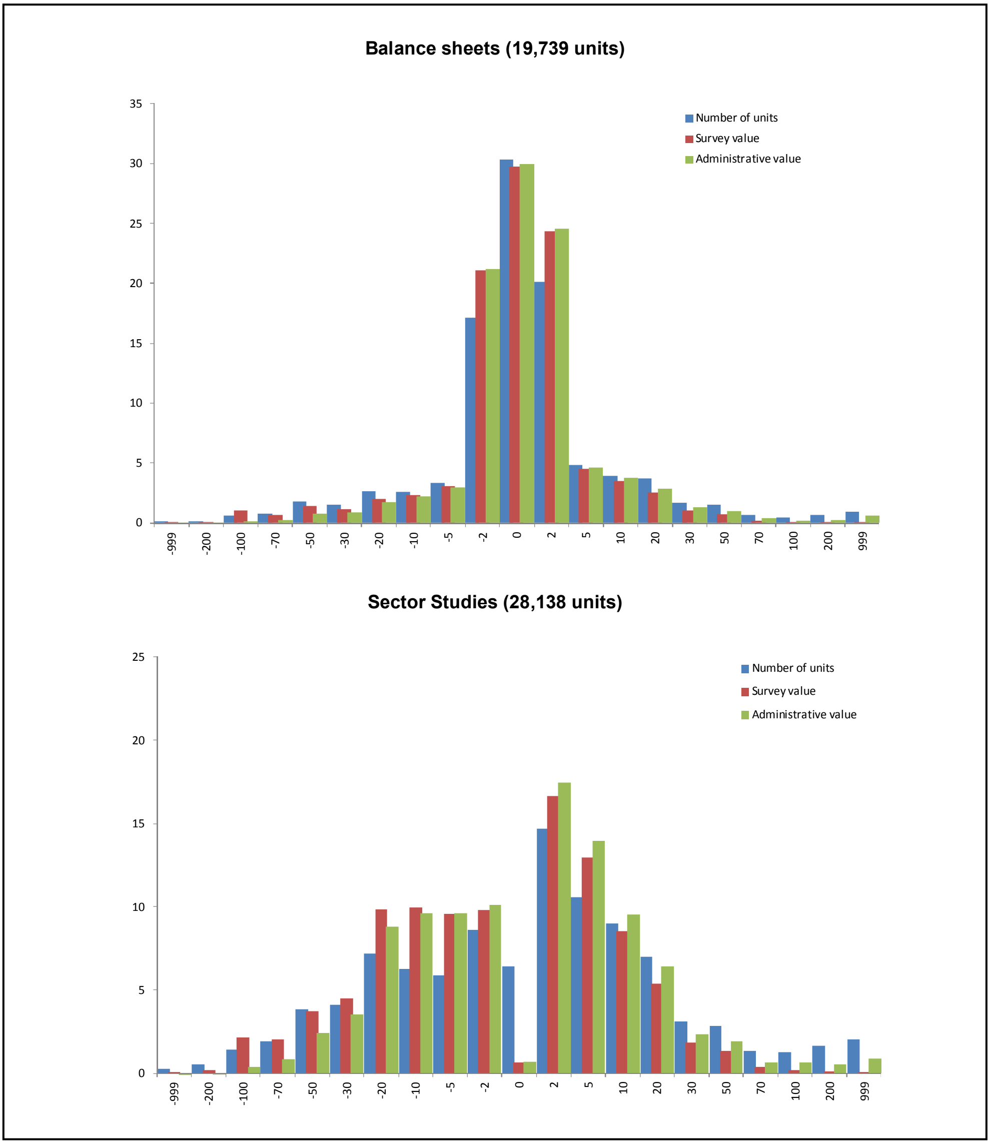
Tax return PF RG (7,272 units)

Tax return SP RG (2,617 units)

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

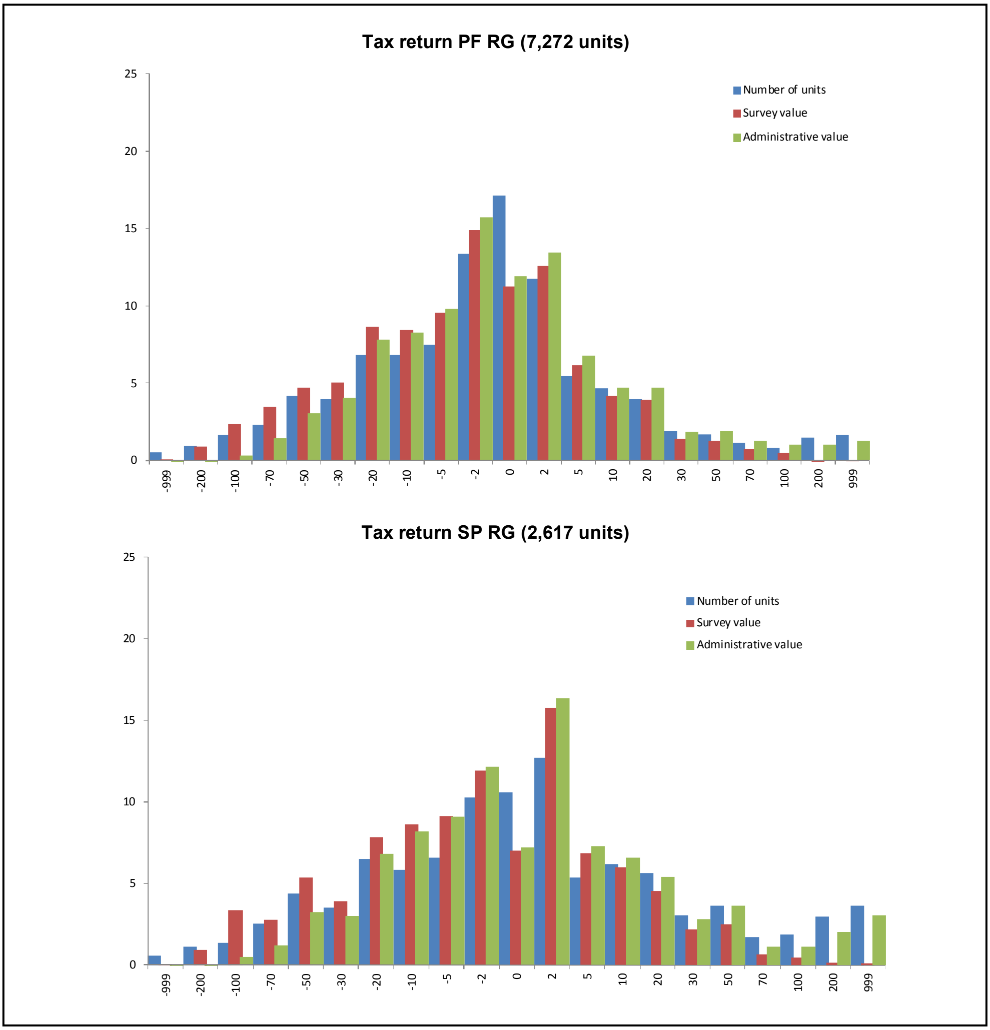
About 94% of total observations, linked with the balance sheets (19,739 units), show differences in the value of turnover lower than 2% and 70% with identical values. For the fiscal sources the analysis is satisfying too. For the first one (Sector Studies) a percentage of 88% of observations lies between the range of +/- 2%, with a 63% with identical values. For the other sources (Tax return modules RG (simplified accounting scheme) for sole proprietorships (PF) and partnerships (SP)) concordance of turnover value, between +/- 2%, is above the 80%. For all cases it is observed a symmetric distribution of differences, that give an evidence of the randomness of errors and the normality test on the distribution of the differences is statistically significant (the Kolmogorov-Smirnov is good and the statistic is equal to 0.4 with a p-value lower than 1%).

Figure 3 - Value added, distribution of respondent units linked with administrative sources by classes of differences - Year 2007 *(percentage values)*



Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

Figure 3 segue - **Value added, distribution of respondent units linked with administrative sources by classes of differences - Year 2007** *(percentage values)*



Tax return PF RG (7,272 units)

Tax return SP RG (2,617 units)

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

The reconstruction for each observation of the variable Value added through the administrative data has been more complex because it is a result of an algebraic sum of more budget items. In the balance sheets case the situation is better than the fiscal sources. In this latter cases there have been encountered difficulties in the perfect reconstruction of cost items that are affected by the fiscal legislation. Nevertheless for the Sector Studies at least 46% of linked observations and 55% of the Value Added lie between an error of +/- 5% and at least 60% of linked observations and 72% of the Value added lies between an error of +/- 10%. Also in this case the normality test on the distribution of the differences is statistically significant for balance sheets (the statistic Kolmogorov-Smirnov is equal to 0.4 with a p-value lower than 1%).

Finally the comparative analysis shows in each case zero-balanced and symmetric distribution of differences that can be assimilated to a random error.

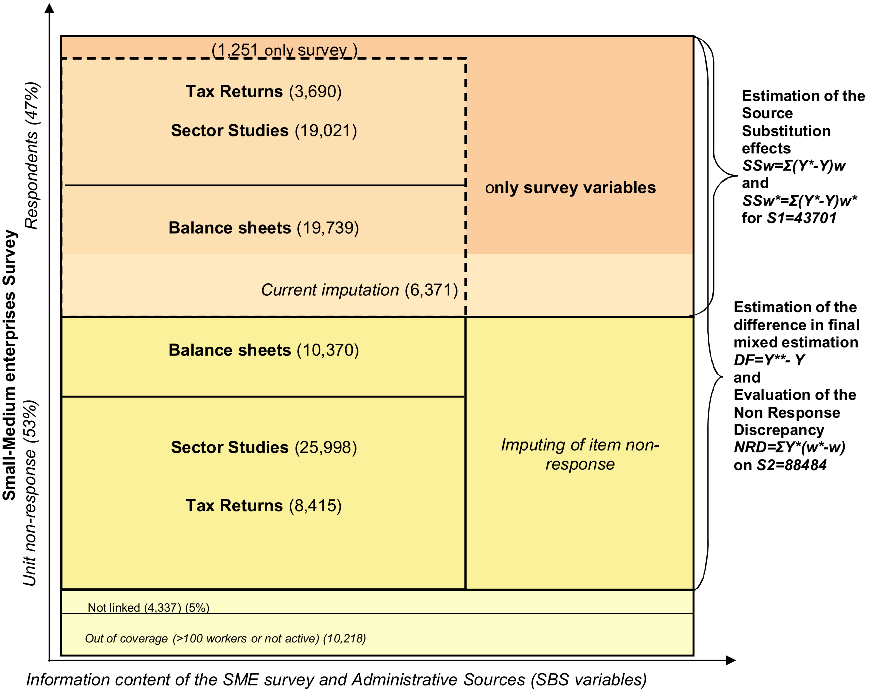
3. Integration of SME survey with administrative sources and re-estimation

3.1 Imputation of unit non-response with data from administrative sources

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.

On the y-axis (Figure 4) the theoretical sample is broken down between respondents (47%) and unit non response (53%). The x-axis represents the dimension of the information content. Administrative sources (Balance sheets and Tax data) cover only a part of the information contained in the survey questionnaire. For the covered part it is possible to compare survey variables (Y) with SBS variables calculated with balance sheets, sector studies and tax returns (Y^*).

Figure 4 - Integration frame of the SME Survey with Administrative sources - Year 2007


Small-Medium enterprises Survey
Respondents (47%)
Unit non-response (53%)
(1,251 only survey)
Tax Returns (3,690)
Sector Studies (19,021)
Balance sheets (19,739)
Current imputation (6,371)
only survey variables
Balance sheets (10,370)
Sector Studies (25,998)
Tax Returns (8,415)
Imputing of item non-response
Not linked (4,337) (5%)
Out of coverage (>100 workers or not active) (10,218)
Information content of the SME survey and Administrative Sources (SBS variables)
Estimation of the Source Substitution effects SSw=Σ(Y*-Y)w and SSw*=Σ(Y*-Y)w* for S1=43701
Estimation of the difference in final mixed estimation DF=Y**- Y and Evaluation of the Non Response Discrepancy NRD=ΣY*(w*-w) on S2=88484

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

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

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

Estimate based on administrative sources (*S2*), with a new set of final weights, is:

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

Final 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 integrated estimate (survey plus administrative) is equal to the estimate based on administrative sources with new weights minus the weighted discrepancy between sources.

3.2 Calculation of the estimation discrepancy

Starting from the previous formulas, various component of the final estimation difference can be highlighted.

The difference between the administrative (Y^*) and survey (Y) estimate is equal to:

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

where $\tilde{Y}_\alpha^*$ is the new variable calculated with the administrative source and w_k^* the new vector of final weights 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 response: $\gamma_1^* < \gamma_1$. On the overall initial sample of roughly 93 thousands units, the new respondents are roughly 88.5 thousand units, with an average correction factor $\gamma_1^* = 1.05$. Instead of the correction factor described in par. 1.4 $\gamma_1 = 2.12$.

Adding $\sum_{S1} y_k^* \; w_k$ and subtracting $\sum_{S2} y_k^* \; w_k$ in the difference formula, where w_k is zero for all units of S2 not included in S1, the result is the following:

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

That can be grouped in the following two components:

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

Moreover considering the integrated estimate $\widetilde{Y}_\alpha^{**}$, as the final new estimate, it is possible to calculate the total difference. So, considering that:

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

and introducing the component "source substitution effect" evaluated with the new weights $\sum\nolimits_{S1}(y_k^* - y_k)\,w_k^*$ in the final difference:

$$DIFF = \widetilde{Y}_\alpha^{**} - \widetilde{Y}_\alpha \;= \widetilde{Y}_\alpha^* - \widetilde{Y}_\alpha \; - \sum\nolimits_{S1}(\,y_k^* - y_k)w_k^*$$

the previous difference (DIFF) becomes:

$$DIFF = \widetilde{Y}_\alpha^{**} - \widetilde{Y}_\alpha \;= \sum\nolimits_{S1}(y_k^* - y_k)\,w_k - \sum\nolimits_{S1}(\,y_k^* - y_k)w_k^* + \sum\nolimits_{S2} y_k^*\,(w_k^* - w_k)$$

$$DIFF = SSw - SSw^* + NRD$$

in which the total difference can be decomposed into three effects:

- The source substitution for S1 with old weights: $SSw = \sum\nolimits_{S1}(y_k^* - y_k)w_k$;
- The source substitution for S1 with new weights: $SSw^* = \sum\nolimits_{S1}(y_k^* - y_k)w_k^*$;
- The unit non-response discrepancy for S2: $NRD = \sum\nolimits_{S2} y_k^*(w_k^* - w_k)$.

3.3 Measurement of coherence among different sources and evaluation of the missing response effect

The analysis of differences in the final estimate have been evaluated for the following economic variables: Turnover and Value Added.

Table 6 shows the decomposition of differences between the official estimate of Turnover *Y* and the new estimate of Turnover Y^{**} obtained combining survey and administrative sources.

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

Table 6 - Turnover estimates, analysis of the differences between sources and evaluation of the unit non-response effect by sectors of activity and size class - Year 2007 *(percentage difference)*

%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 unit non-response

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**

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

Table 7 shows the decomposition of differences between the old estimate Y of Value Added and the new estimate Y^{**} of Value Added obtained combining survey and administrative.

Table 7 - Value Added estimates, analysis of the differences between sources and evaluation of the unit non-response effect by sectors of activity and size class - Year 2007 *(percentage difference)*

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

SECTORS	1-9	10-19	20-49	50-99	Total
Industry	-3.24	-3.16	-2.37	1.90	-1.88
Constructions	-8.16	-8.93	2.19	-1.51	-6.35
Services	-5.99	-5.62	-2.79	-3.91	-5.32
Total	**-5.95**	**-5.25**	**-2.03**	**-0.84**	**-4.50**

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

SECTORS	1-9	10-19	20-49	50-99	Total
Industry	-1.22	-3.30	-0.41	-0.58	-1.33
Constructions	0.21	6.51	-2.49	-9.12	0.51
Services	1.53	-0.86	-5.00	1.13	0.26
Total	**0.94**	**-0.57**	**-2.62**	**-0.51**	**-0.15**

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

SECTORS	1-9	10-19	20-49	50-99	Total
Industry	-1.06	-1.76	-0.31	-0.55	-0.89
Construction	0.19	3.08	-2.30	-10.02	-0.38
Services	0.89	-1.20	-3.56	1.14	-0.03
Total	**0.50**	**-0.83**	**-1.94**	**-0.57**	**-0.34**

NRD - Difference due to unit non-response

SECTORS	1-9	10-19	20-49	50-99	Total
Industry	-3.08	-1.61	-2.27	1.93	-1.43
Constructions	-8.19	-12.35	2.38	-2.41	-7.24
Services	-6.63	-5.96	-1.35	-3.90	-5.62
Total	**-6.39**	**-5.51**	**-1.35**	**-0.90**	**-4.68**

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

For the Value Added there are higher differences in the size classes 1-9 and 10-19 especially for construction sectors. In any case total difference is more affected by unit non-response than the substitution of data sources (-0.15 and -0.34). The substitution effect is very high in the medium firms of construction (-9 and -10%).

Finally, a simulation study has permitted to evaluate a gain in the efficiency due to the estimator. Performing 1000 jackknife replications for selecting randomly 75% of the sample of respondents with the same stratificated design adopted by the initial sample, the absolute relative bias (ARB) and the relative root mean square error (RMSE) can be expressed in the following way:

$$ARB = \left| \frac{1}{1000} \sum_{r=1}^{1000} \frac{\hat{Y}_r - Y}{Y} \right|$$

$$RRMSE = \frac{1}{Y} \sqrt{\frac{1}{1000} \sum_{r=1}^{1000} (\hat{Y}_r - Y)^2}$$

As for the estimation of Turnover, errors are always lower of more than 1 percentage point. The error has been reduced in the service activities and in the size class 10-19.

Table 8 - Measurement of the efficiency of new estimates of Turnover after 1000 sampling replications - Year 2007 *(mean values of strata estimates)*

	Y=Turnover (dataset S1)		*Y***=Turnover (dataset S2)	
	ARB	*RRMSE*	*ARB*	*RRMSE*
Sector				
Industry	8.22	9.97	7.11	8.62
Constructions	7.06	8.59	5.95	7.35
Services	10.36	12.52	8.97	10.82
Size class				
1-9	7.76	9.40	6.18	7.55
10-19	11.15	13.43	7.83	9.55
20-49	7.40	9.02	7.29	8.92
50-99	10.43	12.65	9.49	11.40
Total	**9.18**	**11.13**	**7.70**	**9.35**

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

As for the Value Added, errors are also reduced of roughly 2 percentage point. High gain in efficiency has been verified in all economic activities and for all size classes.

Table 9 - Measurement of the efficiency of new estimates of Value added after 1000 sampling replications - Year 2007 *(mean values of strata estimates)*

	Y=Value added (dataset S1)		*Y***=Value added (dataset S2)	
	ARB	*RRMSE*	*ARB*	*RRMSE*
Sector				
Industry	6.81	8.31	5.76	7.04
Constructions	7.70	9.31	6.62	8.14
Services	9.19	11.24	9.31	11.52
Size class				
1-9	7.18	8.82	6.08	7.59
10-19	9.04	10.99	7.65	9.32
20-49	6.75	8.23	6.01	7.36
50-99	8.49	10.49	8.73	10.94
Total	**7.87**	**9.63**	**7.12**	**8.80**

Source: Elaborations on data from Administrative Sources and from Survey of Small-Medium Enterprises

Conclusions

The unsatisfactory sampling survey response rate together with the availability of a huge amount of data from administrative sources (balance sheets and tax data) has suggested some adjustments in the SME production process. The integration of the original SME sample with administrative sources has allowed both to increase the response rate from about 47% to roughly 88% and to measure the discrepancies in the final estimation due to unit non-response.

The experimental study shows that the estimates of the previous survey not always are coherent with the new estimates based on the integrated dataset for the presence of a source substitution effect (i.e. discrepancy of the same variable in different sources) and of a missing response effect after the widening of the respondent sample. The discrepancy calculated between the variables of the previous SME survey and the variables of the new integrated dataset is 0.03% for the Turnover and -4.5% for the Value added. Particularly the difference due to unit non-response is higher in some strata (micro-small firms of service and construction sectors) reflecting the fact that the missing response mechanism is not based on a random process. Moreover, the enlargement and the integration of the final dataset, used in the new process of estimation, could reduce the errors of the estimators of roughly 2%.

A further work needs to be done like a more disaggregated analysis (Nace at 4 digits, Nace at 3 digits and size classes) in order to detect errors, to better harmonize tax data for statistical purpose and to reduce the source effects (measurement error). A further analysis on the informative contents of tax data could permit to extend this experiment to other SBS variables. While for other SBS variables which cannot be obtained from administrative sources it will be necessary to develop specific statistical imputation methods. For that aim, it could be desirable that Istat should have an active role in designing tax forms harmonizing concepts and adding some information useful for statistical purposes. Finally it needs to remark that the use of administrative sources for statistical purpose will imply the continuity and the stability over time of the data flow in order to guarantee the requirements of the Eurostat SBS regulation.

References

Bernardi A., Cerroni F., De Giorgi V. (2010), "Analysis on economic fiscal data for a statistical use" Working paper presented at the Seminar "Using Administrative Data in the Production of Business Statistics: Member States experiences", organized by the Eurostat ESSnet Project *AdminData*, held at Istat in Rome the 18th and 19th of March 2010.

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

Casciano C., Falorsi P.D., Filiberti S., Pavone A., Siesto G. (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.

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), *Conti economici delle imprese - Anno 2005.*

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.

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.