Estimating structural business statistics based on administrative data: the case of the Italian small and medium enterprises

Luzi O., Rinaldi M., Seri G., Guarnera U., De Giorgi V.
Italian National Statistical Institute (Istat)

1. Introduction

Following the most recent EU Regulations on structural business statistics (SBS) establishing that, in order to estimate information on the structure of National production systems, Member States can integrate data available in different information sources (including administrative data), the Italian National Statistical Institute (Istat) has recently started a re-design project concerning the System of Enterprises Accounts Surveys with the main objective of reducing burden on enterprises, production costs and non response rates, while maintaining high data quality levels. The objective is widening the use of administrative data in this area and gradually move from a traditional *stove pipe model* for the statistical production to an integrated model where administrative data represent the information *core*, and statistical surveys are conducted to estimate specific sub-populations/variables which are not available in external archives.
In Italy, SBS on enterprise accounts are currently obtained from two annual statistical surveys, partly integrated with administrative data. The main administrative sources used in this context are *Financial statements* (*FS*) and *Tax Authority* data, in particular the *Sector Studies Survey* data (*SS*).
In this paper, besides the analysis of the usability of *FS* and *SS* in terms of completeness and coverage, experimental evaluations of the potential biasing effects of integrating these sources with SBS survey data are illustrated. Some results for key variables directly available in these administrative sources (*Turnover*, *Purchases of goods and services*, *Personnel costs*) are illustrated. Variables which are not directly available from external sources require a higher modelling effort: some experiments on the components of *Change in stocks of goods and services* (Casciano et al., 2011; Luzi et al., 2011) have been already performed in the context of the *ESS-Net on the use of Administrative data for business statistics* (http://essnet.admindata.eu/)[1].
The paper is structured as follows: in Section 2, the current Istat surveys on SBS and the administrative data useful in this context are illustrated. Section 3 contains some analyses on the quality of the external data sources, and the results of the performed experimental studies. Final remarks are reported in Section 4.

2. Structural Business Statistics in the Italian survey system

In this paper we refer to the Italian system on SBS as to the set of annual surveys aiming at investigating mainly profit-and-loss accounts of enterprises. The SBS population (reference year 2008) is composed by about 4.5 million enterprises. The Italian enterprises' structure is characterised by a strong presence of small and very small units: in effect, the enterprises with 1-9 persons employed are about 4.2 million, and account for 33.2% of total value added and for 47.2% of total employment. On the opposite side, the enterprises with 100 or more persons employed are about 11,000: they account for 37.2% of total value added and 25.3% of total employment.
In order to meet SBS regulation, Istat carries out two distinct annual surveys: 1) the *Annual Survey on the Economic Accounts of Enterprises* (*SCI* in the following), which is a total survey on the enterprises with 100 or more persons employed, and 2) the sample survey on *Small and Medium Enterprises* survey (*SME* in the following) upon enterprises with less than 100 persons employed. Both surveys involve units belonging to the industrial, construction, trade and services economic activities and collect information concerning profit-and-loss statements and balance sheets, as well as information regarding employment, investment, personnel costs and the regional breakdown of some variables. For both surveys the frame is represented by the *Italian Business Register* of active enterprises (*BR* in the following), resulting from the combination of both statistical and administrative information. Target parameters are estimated by publication domains in accordance with the *SBS* Regulation[2].

---

[1] funded by Eurostat in 2009 in the framework of the *Modernisation of European Enterprises and Trade Statistics* (*MEETS*) Program.
[2] The data domains are: 1) class of economic activity (4 Nace-code digits); 2) economic activity (3 Nace-code digits) by size (classes of persons employed); 3) economic activity (2 Nace code digits) by regions (Nuts2 level).

The 2008 *SME* sample consists of about 105,000 enterprises, selected based on a one stage stratified random sampling without replacement and with equal probabilities, with the strata defined in terms of economic activity, size classes of persons employed and regions. The sample selection uses the JALES procedure (Ohlsson, 1995) to achieve a negative coordination among samples drawn from the same selection register and thus reduce response burden. Standard calibration estimators (Deville et al., 1992) and post-stratification are adopted at the estimation stage. The non response rate was about 60%: for a subset of unit non responses, the main balance sheet variables have been integrated by direct replacement of values from either *FS* (48%) or *SS* (52%). Item non responses have been integrated by within-cells donor imputation, where cells are defined in terms of economic activity, size class and geographical area.
Concerning SCI, whose non response rate in 2008 has been of about 55%, unit non responses were integrated by direct replacement of values from *FS*, and item non responses have been imputed by a within-cells donor method analogous to the one adopted for SMEs.
As to administrative sources, out of the 4.4 million frame enterprises, the corporate companies (about 15,7% in terms of enterprises and 37,2% in terms of persons employed) are liable to fill in *FS*. This is the best harmonized source with the SBS Regulation definitions. All other enterprises have to declare their taxable income to the Fiscal Authority by filling in tax forms. In particular, the *SS* is a fiscal survey involving enterprises with specific economic activities and turnover in the interval [30,000 – 7.5 million] Euros: almost all Italian enterprises fill in the *SS* form together with the tax return one, thus declaring in detail costs and income items: as the *SS* form actually looks like a financial statement (Bernardi et al., 2010), it provides more effective data than tax return forms.

3. Integrating administrative and survey data: data analysis and quality evaluations

Using administrative data for statistical purposes poses the additional problem of assessing their quality and usability in the statistical process. In this paper we refer to the definitions adopted in the Eurostat quality framework for administrative data (Eurostat 1999, 2003), in which the quality concepts adopted for statistical data (e.g. relevance, accuracy, accessibility and clarity, timeliness, coherence and consistency) are "adapted" to the characteristics of administrative data.
The coverage of the theoretical sample (SCI+SME) and of the population has been analysed in terms of number of units (Tables 1). Moreover we also considered the coverage in terms of number of persons employed by NACE sections (Table 2).

Table 1: SBS sample (SCI+ SME) and population coverage by type of response and administrative data - year 2008

Source Coverage	SBS (SCI+PMI)	Respondents	Theoretical sample	Population
PMI-respondents	36850 (35%)			
SCI-respondents	4795 (43%)			
FS	32320 (28%)	≈20.5k (49%)	≈53k (46%)	≈650k (14%)
SS	14635 (13%)	≈25k (60%)	≈64k (55%)	≈3480k (77%)
MRT out of coverage	23136 (20%)			
Total number of obs	116137	≈41.5k	≈116k	≈4500k

Table 2: SCI+ SME sample and population coverage (%) of the available administrative sources, in terms of number of enterprises (ENT) and number of employees (EMP) by economic activity - Year 2008.

	Sample (respondents)						Population					
	FS		SS		Total		FS		SS		Total	
NACE Sections	ENT	EMP	ENT	EMP	ENT	EMP	ENT	EMP	ENT	EMP	ENT	EMP
B mining and quarrybg	68.3	95.6	72.8	9.9	94.7	98.0	53.1	82.9	76.9	40.0	87.8	95.8
C Manufacturing	56.6	93.5	57.0	7.4	89.6	95.8	24.1	69.0	76.1	44.6	83.1	92.4
D Electricity, gas, steam and air conditioning supply	84.6	93.7	3.1	0.1	85.3	93.7	72.1	92.3	2.0	0.3	72.4	92.3
E Water supply, sewerage, waste management and remediation activities	64.7	94.5	22.4	2.2	77.7	95.2	49.6	82.4	29.3	9.2	66.9	86.0
F Construction	45.3	91.8	75.4	17.4	93.2	97.8	15.3	37.6	78.4	74.6	81.0	87.7
G Wholesale and retail trade; repair of motor vehicles and motorcycles	46.8	92.4	71.1	7.6	94.5	95.4	12.0	42.4	83.0	65.0	85.2	90.6
H Transportation and storage	58.3	97.2	58.0	3.7	91.4	97.7	16.1	71.2	78.3	33.1	84.2	89.7
I Accommodation and food service activities	40.6	83.3	76.3	8.3	91.9	85.9	10.9	32.6	80.3	75.0	82.2	88.8
J Information and communication	60.3	89.2	62.8	5.7	86.9	89.9	30.9	74.6	71.8	40.4	78.4	89.5
K Financial and insurance activities	18.0	37.4	74.6	51.8	79.8	65.7	6.0	16.5	80.2	77.1	81.1	82.7
L Real estate activities	42.0	77.0	86.7	45.9	93.6	97.2	39.2	37.5	90.0	87.8	93.3	93.5
M Professional, scientific and technical activities	32.4	87.1	67.4	16.9	83.5	93.0	6.2	25.4	74.9	68.8	76.7	83.3
N Administrative and support service activities	52.5	96.2	52.5	11.8	80.3	97.7	23.5	72.8	58.2	32.5	67.1	87.0
P Education	53.4	94.4	26.0	3.6	74.0	95.5	6.8	51.2	70.9	40.0	75.5	85.8
R Arts, entertainment and recreation	35.1	82.2	36.6	15.7	60.3	87.3	16.5	46.5	33.5	35.7	43.4	67.0
S Other service activities	19.3	80.5	77.9	37.1	84.3	96.7	4.3	14.9	71.2	77.1	72.1	82.5
Total	**49.4**	**93.1**	**60.5**	**7.6**	**87.4**	**95.3**	**14.4**	**51.8**	**77.2**	**55.4**	**80.7**	**89.1**

It appears that the two administrative sources cover around 80% and 90% of the population in terms of units and persons employed respectively. These percentages increase if we refer to the sample (of respondents). The coverage is extremely high even if we consider NACE sections independently.
Exploratory analysis aiming at assessing the consistency of the target variables have been conducted. In Table 3 the results of the Kolmogorov-Smirnov tests for the variables *Purchases of goods and services*, *Turnover*, *Personnel costs* and *Changes in stocks* (this last has been considered in order to test a variables that presents a high number of zeros observed and negative values). Results show that FS appears to be more consistent to the survey data with respect to the SS data but also that the two administrative sources are consistent between them in the subset of the respondents.

Table 3: Kolmogorov-Smirnov tests for the variables *Purchases of goods and services*, *Turnover*, *Personnel costs* and *Changes in stocks* in the subpopulations of respondents belonging to the administrative sources FS, SS and both FS and SS respectively: light and dark grey highlights for null hypothesis accepted at 0.05 and 0.01 alpha level respectively - Year 2008.

Sources (n of bservations) Variable	Survey-FS (20579)	Survey-SS (25185)	FF-SS (9349)
GSPurch's	2.15	3.11	0.91
Turnover	0.20	0.50	0.18
PersCosts	1.51	4.24	0.32
CoS	0.73	2.17	0.50

In terms of *completeness* (number of items directly available in administrative sources) *FS* and *SS* theoretically provide all the information needed to estimate parameters for a subset of key SBS variables: *N. of employees, Turnover*, *Changes in stocks*, *Changes in contract work in progress, Other income and earnings, Purchases of goods and services, Use of third party assets, Other operating charges, Personnel costs, Value added* and *Gross operating value*.
In next sections we focus the attention on *accuracy*, that is on the statistical adequacy of the information provided by *FS* and *SS* for estimating key SBS parameters. The performed studies assume that the available external data can be used in two different ways in a statistical process:

(a) as auxiliary information to improve the efficiency of the statistical survey process (e.g. for optimizing error detection, non response imputation, and estimation strategies, see Casciano et al., 2011);
(b) as primary source of information for estimation[3] (Casciano et al., 2010), complemented by direct surveys to estimate either non covered sub-populations, or variables which are not directly available from external sources (Luzi et al., 2011).

Concerning situation (a), the experiments made aim at evaluating the potential effects on SBS parameters' estimates of the following strategies:

1) adopting a *selective editing* approach (Latouche et al., 1992) to optimize the identification of influential measurement errors in survey data by exploiting information from *FS* and *SS*;
2) imputing survey non responses using information from *FS* and *SS*.

Under situation (b), some experimental analyses have been performed in order to evaluate the statistical effects on SBS parameters' estimates of using administrative data in place of survey data for specific sub-populations (*source effect)* of enterprises.
All studies have been performed on data referring to year 2008. In all experiments the analysis has been limited to the subset of units also available in either *FS* or *SS* archives.

3.1.Evaluating the data editing effect

In general, administrative data suffer from measurement errors similar to those affecting survey data (see for example Hoogland, 2010). As a consequence, traditional error detection methods can be adopted for this type of data. We are especially interested in exploring methods for the identification of measurement errors which are influential on target estimates, which exploit the information available in *FS* and *SS*. *Selective editing* can be useful in this case. In this paper, the results of first analyses on variables *Turnover* and *Personnel Costs* on SMEs are illustrated. For each variable *Y* and for each domain *D,* separately for each administrative source, a *local score* is defined as the weighted difference between survey and external elementary data: once units are ordered by descending values of the *local score*, the lowest *k* index is computed as the value which guarantees that, once all values of *Y* in the first *k* units are replaced by the corresponding values from the

[3] in this case, the sample survey design can also exploit the availability of the additional administrative information for improving the efficiency of the estimation process in terms of trade-off between sample dimension and accuracy.

considered source, the difference between the resulting estimate of the *Y* total differs less than 2% from the corresponding total estimated on the original (observed) *Y* values[4]. If *k* is low, few units are to be replaced by the administrative data source in order to have low discrepancies among estimates, so few discrepancies (potentially due to influential measurement errors) are found on elementary data, hence variables definitions in survey and administrative sources can be reasonably assumed as homogeneous.

As an example, in Table 4 some results of the comparison between SME and *FS* for *Turnover* and *Personnel Costs* are reported for some economic activities: since in general few units are found as responsible of discrepancies among survey and administrative totals' estimates, we are confident about the reliability of *FS* for the analysed variables at micro level, and are allowed to use this source as auxiliary information to optimize the detection of potentially influent measurement errors. Analogous considerations can be made for *SS* (results are not reported for shortness). These results encouraged us in starting further experiments in this context: in particular, the selective editing approach based on contamination models (Di Zio et al., 2008) is currently under evaluation on SME data.

Table 4 – Selective editing effects when comparing *SME* and *FS* data, by variable and economic activity

Economic Sector	2-dig. Nace code	Variable									
		Turnover					Personnel Costs				
		N	n influential	% Influential	Diff_ori (%)	Diff_rep (%)	N	n influential	% Influential	Diff_ori (%)	Diff_rep (%)
C-Manifacture	**26**	254	1	0.4	20.07	-0.14	254	3	1.2	-2.90	-1.81
	27	227	1	0.4	-2.05	-0.02	227	1	0.4	-2.61	-1.61
M- Professional, scientif. Techn. Activities	**72**	96	1	1.0	-2.04	1.60	96	2	2.1	-4.57	-0.19
D -Electricity, gas, steam and air conditioning supply	**35**	263	1	0.4	-5.40	0.71	263	3	1.1	-5.28	-1.60
E- Water supply, sewerage, waste management and remediation act..	**37**	59	1	1.7	2.24	0.48	59	3	5.1	-2.82	-1.05

3.2. Evaluating the non responses imputation

In this case, we evaluate the usefulness of administrative data as auxiliary information for predicting (imputing) survey non responses for key SBS. We focus on variables which are directly available in either *FS* or *SS*, and limit the attention on the sub-population on SMEs, which is characterized by more complex problems than larger enterprises in terms of usability of the available administrative information. For each variable *Y* and a domain *D,* target parameters are *Y* totals by publication domain *D*: $T_Y^D = \Sigma_{k \in D} \omega_k y_k$, where ω_k is the sampling weight of unit *k*. For each variable *Y*, a MonteCarlo experiment has been performed, consisting in *I*=100 iterations of the following steps: 1) simulation of missing values at random w.r.t. economic activity on the subset of responding units; 2) "non response" imputation and estimation; 3) evaluation, in terms of *Relative Root Mean Squared Error (RMSE)*, *Weighted Relative average imputation error* (*WRIE*) and *Relative estimation error due to imputation* (*REEI*) (Luzi et al., 2007), defined as follows:

$$WRIE_Y(i) = \frac{\sum_{k=1}^{n^*} \omega_k \left| y_{true,k} - y_{imp,k} \right|}{\sum_{k=1}^{n^*} \omega_k y_{true,i}} = \frac{1}{I} \sum_{i=1}^{I} WRIE_Y(i)$$, where $y_{true,k}$, $y_{imp,k}$ are the original and imputed *Y* values in unit *k*, respectively, and n^* is the number of responding units with simulated non response. Let RIE_Y be the corresponding un-weighted indicator;

$$REEI_Y^D = \frac{1}{I} \sum_{i=1}^{I} \frac{\left(\hat{T}_{Y,true}^D - {}_k\hat{T}_{Y,imp}^D(i) \right)}{\hat{T}_{Y,true}^D},$$ where $\hat{T}_{Y,true}^D, \hat{T}_{Y,imp}^D$ are the parameters estimates computed on true or imputed (at iteration *i)* values, respectively.

Within-cell regression imputation[5] is adopted for predicting artificial non responses, with cells defined in terms of *Economic activity*[6] and size class[7]. For a given response variable *Y*, the model's covariate is the *corresponding*

[4] The discrepancies are computed as $Diff\_rep = \frac{\left(\hat{Y}_{D,SME}^k - \hat{Y}_{D,ADM} \right)}{\hat{Y}_{ADM}^D}$, where *ADM=FS* or *SS* (depending on the analyzed source) and the first *k* values in the ordered list of units are replaced by the corresponding administrative ones. *Diff_ori* is the difference between estimates computed on the original SME data, and on the corresponding administrative data.

[5] Robust regression is not used since we assume that influential errors as well as discrepancies between survey and the corresponding administrative variable values have been removed at the data editing stage.

[6] 16 Sections of economic activity (corresponding to groups of 2-digits Nace code).

item directly available from either FS or SS, with priority assigned to *FS* in case of units available in both sources. Model fitting was good in all imputation cells. Different experiments have been performed for different non responses rates (5% and 10%). In Table 5, we report some results for variables *Turnover* and *Purchases of goods and services* for sections *Manifacture* and *Commerce*: observed non response rates in these sections are 5.8% and 4.9%, respectively. In the table, quality indicators by non response rate and enterprises' size[8] are shown.
As it can be seen, imputation does not significantly affect estimates for both economic activity sections and missing rates. The worst performance corresponds to the size class [1-9] and, to a lesser extent, to the size class [10-19]. This especially holds for *Commerce.* Higher effects can be seen in terms of RIE and WRIE, especially for *Commerce*: the inspection of elementary data distributions before and after imputation showed that in most cases those values which are predicted unsatisfactorily correspond to original SME values with high discrepancies w.r.t. the administrative ones: this is a further confirmation of the need a more efficient use of external information at the data editing stage.

Table 5 – Quality indicators[9] by variable, non response rate, economic activity section and size class

Variable	% non resp.	Manifacture									
		All		1-9		10-19		20-49		50-99	
		RIE	WRIE	REEI	RMSE	REEI	RMSE	REEI	RMSE	REEI	RMSE
Turnover	5%	3.7	6.9	0.43	0.51	0.22	0.23	0.05	0.09	0.06	0.11
	10%	4.3	7.3	0.90	1.01	0.39	0.44	0.11	0.17	0.16	0.26
Purchases	5%	4.3	7.4	0.45	0.58	0.26	0.28	0.08	0.13	0.07	0.16
	10%	4.6	7.9	0.87	1.11	0.49	0.54	0.12	0.18	0.12	0.24
Variable	% non resp.	Commerce									
		All		1-9		10-19		20-49		50-99	
		RIE	WRIE	REEI	RMSE	REEI	RMSE	REEI	RMSE	REEI	RMSE
Turnover	5%	8.9	17.9	0.69	0.95	0.89	0.96	0.13	0.19	0.09	0.15
	10%	8.3	18.9	1.23	1.43	1.68	1.80	0.18	0.25	0.20	0.31
Purchases	5%	8.4	17.9	0.64	0.91	0.78	0.88	0.13	0.22	0.12	0.21
	10%	8.5	18.8	1.34	1.67	1.51	1.71	0.24	0.32	0.17	0.27

3.3. Evaluating the source effect
This paragraph describes the results of an experimental study aimed at evaluating the statistical effects of using administrative data in place of survey data to estimate totals of *Turnover*, *Purchases of goods and services* and *Personnel costs* for the target population of 2008 SBS surveys (*SCI+SME*). The *source effect* is evaluated separately for *FS* and *SS* by using estimation techniques within subpopulations where *FS* and *SS* are –respectively– available, so that it is also possible to calculate the "true" totals from administrative sources. Subpopulation SP_1 includes about 650,000 corporate companies for which *FS* are available, while subpopulation SP_2 includes almost 3,500,000 enterprises filling in *SS*; both *FS* and *SS* data are available for approximately 500,000 companies.
The totals are estimated by a calibration process carried out on the 41,646 responding to SME and SCI, where the auxiliary variables are *no. of enterprises* and *no. of persons employed*, the distance function is linear[10] and the calibration domains are identified by crossing Nace activity groups (3-digits) and size classes[11]. The Italian *BR* contains all the information required to calculate the known totals of the auxiliary variables by estimation domains.
Let y_k represent the value of the variable Y collected on the k-th responding unit and y_k^{adm} the value of y_k taken from the administrative source (*FS* or *SS*). The calibrated weights w_k calculated for responding units can be used to estimate totals from both surveys and administrative data, as follows:

$$\text{(a) } \hat{Y}_{SP_i} = \sum_{k \in R \cap SP_i} w_k \cdot y_k, \quad i = 1, 2$$

$$\text{(b) } \hat{Y}_{SP_i}^{adm} = \sum_{k \in R \cap SP_i} w_k \cdot y_k^{adm} \quad i = 1, 2$$

where R is the set of respondents while the administrative source considered is *FS* for i=1 and *SS* for i=2. Totals can also be calculated (not estimated) in subpopulations where *FS* and *SS* are –respectively– available:

$$\text{(c) } Y_{SP_i}^{adm} = \sum_{k \in SP_i} y_k^{adm} \quad i = 1, 2.$$

[7] 4 size classes have been defined: [1-9], [10-19], [20-49], [50-99].
[8] Note that RIE and WRIE are computed on the overall subset of units subject to non responses simulation (independently on size).
[9] Values of indicators are multiplied by 100.
[10] Final weights are positive because ratios between final and initial weights have been imposed positive.
[11] Size classes are defined in terms of no. of persons employed: 1-9, 10-19, 20-49, 50-99, 100-249, 250 or more.

In the experiment, estimates (a) and (b) and totals (c) are computed within subpopulations, also disaggregated by Nace sections. Table 6 reports percent differences between estimates of totals of the three variables calculated from survey and administrative data that is, for each variable Y:
$\left|(\hat{Y}_{SP_i} - \hat{Y}_{SP_i}^{adm}) / \hat{Y}_{SP_i}\right| \times 100, \quad i = 1, 2$. Highlighted cells in the table correspond to absolute differences greater than 5%. Survey and administrative data produce total estimates of *Turnover* and *Personnel costs* very close in most sections (except "*M"* for *FS*), while differences between estimates of total are higher for *Purchases of goods and services*, especially when the administrative source is *SS* and for *FS* in most services activities. The overall differences are negligible for all combinations variable-administrative source, except for total purchases estimated from *SS* (about 5%).

Table 6 – Difference between total estimates from survey and administrative data, in SP1, SP2

	%diff survey vs. *FS* estim's in SP_1			%diff survey vs. *SS* estim's in SP_2		
Nace Rev 2 – Section	**GSPurch's**	**Turnover**	**PersCosts**	**GSPurch's**	**Turnover**	**PersCosts**
B mining and quarrying	- 0.23	- 0.02	0.42	- 16.94	- 0.15	2.31
C Manufacturing	- 1.57	0.07	- 1.00	- 4.44	0.31	- 0.92
D Electricity, gas, steam, etc.	- 2.70	- 1.39	11.39	- 31.10	- 2.47	- 7.31
E Water supply, sewerage, etc.	- 2.99	0.14	- 0.56	7.82	9.61	0.25
F Construction	- 2.08	0.64	- 1.80	- 2.77	0.96	- 2.46
G Wholesale and retail trade; etc.	- 4.67	- 0.93	- 1.28	- 2.41	- 0.37	- 2.90
H Transportation and storage	- 5.93	- 1.60	- 0.99	- 7.64	0.75	- 1.51
I Accommodation and food service	- 10.92	- 0.19	- 0.18	- 10.38	- 0.37	- 1.07
J Information and communication	13.05	8.60	2.66	- 6.81	- 0.60	- 2.72
K Financial and insurance	- 0.30	0.01	- 1.65	- 8.69	1.03	- 3.26
L Real estate activities	- 19.76	- 2.42	- 2.00	- 16.76	- 0.41	1.09
M Professional, scientific and technical act	- 10.61	21.99	- 1.68	- 10.26	- 1.88	- 2.80
N Administrative and support services	- 2.45	1.58	2.66	- 11.07	- 0.44	- 0.46
P Education	- 6.91	0.65	- 1.12	- 11.81	0.91	- 2.86
R Arts, entertainment and recreation	- 11.80	1.25	- 4.63	- 21.43	- 1.67	- 5.71
S Other services	- 7.80	0.71	- 2.81	- 18.14	3.21	- 2.31
Total	**- 3.12**	**0.53**	**- 0.53**	**- 4.57**	**0.02**	**- 1.83**

As using *FS* and *SS* data in place of survey data produces similar estimates of *Turnover* and *Personnel costs* totals, this can encourage in extending the use of administrative data at the estimation stage.
Furthermore, 95% confidence intervals associated to the obtained estimates have been calculated for all variables, in order to check if "true totals" of administrative data belong to these intervals. The result is that, for all variables and most sectors, true totals calculated from *FS* and *SS* by formula (c) does not belong to corresponding intervals, with the exception of *Purchases of goods and services* in sections *K-S.* This circumstance could depend on sampling elements (mainly related to the need of estimating the target parameters for a very large and complex population minimizing survey costs and response burden), and non sampling aspects (e.g. the presence among external data of measurement errors with strong effects on totals).

5. Final remarks

In the context of the Istat project on redesigning the statistical production process of SBS , a number of data analyses and experimental studies are under development in order to move from the actual production system essentially based on survey data, to a new system characterized by an extensive use of the available administrative sources of information. The results obtained for large, small and medium enterprises (year 2008) support the assumption that external data sources may be efficiently used to significantly reduce costs and response burden while preserving data accuracy. These results represent a good basis for future developments and implementations of the new system for at least a set of key variables required by the SBS Regulations.
However, some additional analyses are still in progress as some aspects need further investigations. In particular, multivariate selective editing approaches, like the one based on the use of contamination models, are under evaluation in order to verify the possibility of improving the error detection phase in SME+SCI data by exploiting all the available information from FS and SS. In general, the optimization of the error detection phase has a direct link with the estimation accuracy. Furthermore, the results relating to the source effect evaluation, and to the accuracy of the sampling estimates obtained by either survey or administrative data, reveal the need for additional analyses and experiments to better understand the reasons of some inefficiencies resulting from the sole use of administrative data for specific subpopulations.

References

Bernardi A., Cerroni F., De Giorgi V. (2010). *Analysis on economic fiscal data for a statistical use*. Working paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics: Member States experiences*, Rome, 16-18 March.

Casciano M.C., De Giorgi V., Oropallo F., Siesto G. (2010). *Experimental Analysis in the estimation of SBS variables for small firms by using administrative data,* paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences*, Rome, 16-18 March.

Casciano C., De Giorgi V., Luzi O., Oropallo F., Seri G., Siesto G. (2011). Combining administrative and survey data: potential benefits and impact on editing and imputation for a structural business survey, *UN/ECE Work Session on Statistical Data Editing*, Ljubljana, 9-11 Maggio 2011 (http://www.unece.org/stats/documents/2011.05.sde.htm)

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

Di Zio M., Guarnera U., Luzi O. (2008). Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. *UN/ECE Work Session on Statistical Data Editing*, Vienna (http://www.unece.org/stats/documents/2008.04.sde.htm).

Eurostat (2003). *Item6 - Quality Assessment of Administrative Data for Statistical Purposes*, Luxemburg.

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

Hoogland J. (2010). Editing Strategies for VAT Data, paper presented at the *Seminar on Using Administrative Data in the Production of Business Statistics-Member States Experiences*, Rome, 16-18 March.

Latouche M., Berthelot J.M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8, n.3, 389- 400

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Templeman C., Hulliger B., Kilchman D. (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. (http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf).

Luzi O., Seri G., De Giorgi V. and Siesto G. (2011). *Development of estimation methods for business statistics variables which cannot be obtained from administrative sources. Variable: Change in stocks of goods and services*. Report for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics.

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

Wallgren A., Wallgren B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes,* John Wiley & Sons.