Integrating administrative and survey data in the new Italian system for SBS: quality issues

Luzi O., Di Zio M., Oropallo F., Puggioni A., Sanzo R.

1. Introduction

The Italian National Statistical Institute (Istat) is developing a new system for the production of structural business statistics (*SBS*) on economic accounts of small and medium enterprises (enterprises with less than 100 persons employed, *SME*s in the following): the system is based on the use of administrative and fiscal data as primary source of information (at firm-level), integrated with direct survey data as complementary information on specific economic issues or businesses' sub-populations. Based on the new system, it will be possible to obtain a multidimensional set of estimates (*frame* in the following) for a set of key SBS at an extremely refined level of detail, starting from the 2011 reference year. The frame is expected to overcome some limitations of the current SBS production system, which is at present essentially based on direct sample surveys, like low response rates, high sampling errors, costs and statistical burden, cross-section inconsistencies between SBS and National Accounts (NA).

At present, information on profit-and-loss accounts of SMEs is obtained from a sampling survey which annually investigates a sample of about 105,000 enterprises in the industrial, construction, trade and non-financial services sectors. The target variables are turnover, intermediate costs, production value, value added, labor cost. The target population is identified based on the Italian Business Register (BR). The frame design and implementation has required a deep analysis of the information contents of the available administrative (*admin* in the following) sources: as known (among recent references, see for example Norbotten, 2010; Zhang, 2012) the statistical use of admin data gives several advantages but it also poses additional problems w.r.t. direct surveys, e.g. initial costs to get the sources, harmonizing concepts and definitions w.r.t. the target units and statistics, matching classifications, and assessing quality. In effect, as admin data are collected for administrative purposes, the data collection, data coding and data validation are not under the control of the Statistical Offices, and additional data analysis and data processing are needed to ensure the statistical usability of these data. In this paper we provide an overview of this kind of problems in the context of the frame design and implementation. Also taking into account the outcomes of European Projects like the Essnet Admin Data (2013) and the BLUE-ETS project (2012), some key quality dimensions have been considered in order to evaluate and ensure the statistical usability of information on SMEs supplied by the available admin and fiscal databases.

The paper is structured as follows. Section 2 deals with the quality issues faced in the *frame* context: in section 2.1 the problem of the consistency of definitions (for target units and target variables) in the different archives w.r.t. the SBS concepts is discussed; in section 2.2. the quality indicators and the analyses performed to evaluate the sources suitability in terms of coverage and accuracy are summarized; section 2.3 discusses some aspects concerning editing and imputation problems encountered, with special attention to the issue of influential data detection and treatment.

2. Quality issues in the design of the frame

Based on the current agreements with the Italian Authorities belonging to National Statistical System, Istat periodically acquires some sources containing information on SMEs useful for SBS purposes:

- **Financial statements (FS)**: contains profit and loss statements of limited liability companies, drawn up periodically, usually once a year as required by law;
- **Fiscal Sector Studies Data (SS)**: it is a survey introduced by the tax authorities to verify the correctness of the declarations of small businesses;

- **Tax returns form (Unico)**, from the Ministry of Economy and Finance, based on a unified model of tax declarations by legal form, containing economic information for different legal forms;
- **Social Security Register (EMENS)**: contains detailed information on occupation and labor cost.

As the sources are partially overlapping, covering some common SMEs sub-populations, a quality assessment process was needed, in order to prioritize them. A preliminary evaluation study (Casciano *et al*., 2011) provided information about priority to be assigned to the considered sources, together with a first evaluation of their completeness w.r.t. the SBS purposes. The quality assessment process has been based on the following criteria: *relevance* (consistency of administrative and statistical definitions for both the target population and target parameters); *coverage* (completeness of the source in terms of target population units); *completeness* (degree to which a data source includes information to estimate the target statistics); *accuracy* (statistical adequacy of administrative items for estimating target parameters); timeliness (or periodicity, relates to the delivery of input data along time); *integrability* (extent to which the data source is capable of undergoing integration or of being integrated). Concerning the latter, all the sources include a common identification code at enterprise level which ensure the possibility of exact linking between them (provided that possible doublets and units splits/fusions over time have been preliminarily resolved for each single source). In next sections, the other quality dimensions are discussed. All the analyses presented refer to year 2010.

2.1. Consistency: variables definitions and variables harmonization

As already mentioned, Administrative Agencies generally use different concepts, definitions and classifications than those required for the specific statistical purposes. As a consequence, the first problem to face in the case of the frame consisted in analyzing the information content of each source in order to evaluate and ensure their consistency with respect to the SBS and SEC definitions (the frame should represent the "regular" CN estimates from the side of production). For each variable of interest, detailed analyses were conducted to assess differences between the definition of administrative items, and the corresponding concepts as defined by SBS, NA and Emens Regulations. Based on these analyses, information from admin sources was subject to a thorough process of harmonization w.r.t. the target statistical concepts. Furthermore, as each single admin source does not contain all of the information which is required to produce the target statistics, it was needed to integrate information from more sources. As a consequence, besides consistency, possible inconsistencies among target statistics need to be maintained. As an example, the case of labor cost (LC) is shortly discussed. Even if the LC definitions in admin data (FS, SS, Unico, Emens) theoretically refer to the same principle, inconsistencies were highlighted by comparing their elementary data. In principle, the most reliable source of information about LC is the Emens database: in effect, this source is the outcome of processing of individual data about costs for both permanent and temporary employers; even if it does not perfectly meet the SBS LC definition, it provides a complete information on LC by types of employees and it represents an optimal proxy for LC SBS estimation. Inconsistencies among the LC measurement in Emens and in the other sources are mainly due to a different “interpretation” of the concept by the enterprises. In fact, the LC definition allows the enterprises to indicate certain costs either in an item or in another one (especially about costs of contract workers) depending on their “convenience”. Based on the more detailed information on LC coming from Emens, LC measurements in the other sources have been adjusted and harmonized w.r.t. the SBS definition. Accordingly, to ensure the consistency with the other economic accounts items, all the corrections made on LC items had to be compensated by appropriate adjustments on related enterprise items in order to preserve the key economic aggregates (value added, above all) for each enterprise.

Concerning units consistency, it is known that administrative units may not correspond to the target statistical units: in these situations, data of administrative units are to be properly treated to derive the

information on actual statistical units. In the frame context, the same LC example can be used: while FS, SS and Unico contain information on LC at enterprise level, in the Emens source the information on LC is available at employee level: as a consequence, in order to perform the above described treatment of LC items, Emens data had to be preliminarily combined to derive the needed information ant enterprise level.

2.2. Coverage and Accuracy

The coverage of the considered admin sources w.r.t. the SMEs target population/economic activities is reported in Table 1. Among units without coverage (which represent about 5%), main issues concerned some domains with a high presence of foreign companies. To guarantee complete coverage, we adopted two main remedies only for *Special Purpose Entities* with more than 4-5 employees: the first one was to recover the financial statements of the controlled unit from an external source; the second one was to rebuild information by using imputation.

Table 1. Business Register Units by source of Administrative data. Year 2010.

Administrative Source	Units	%
Financial statements	718,239	16.2
Fiscal Purpose Survey	2,931,090	66.0
Tax Return Data	585,863	13.2
No source	*208,688*	*4.6*
Total	**4,443,880**	**100.0**

Concerning accuracy, the first problem was to validate the information directly available from admin sources. To this aim, a comparative evaluation was made, based on the joined analysis of the target variables values from the SME survey with the corresponding values obtained from the admin sources (after the harmonization process). Out of the 37,920 SME responding units, 16,602 were linked with financial statements, 14,363 with SS, 8,906 with Unico for individuals and 3,190 with Unico for partnerships. For each target variable, the percentage differences at elementary data level between survey data and data from each admin source were computed, and a set of quality indicators were considered:

- the *Kolmogorov-Smirnov* index;
- the *proportion of firms in the range of difference of ± 5%;*
- the *proportion of value observed on the SME survey in the range of difference of ± 5%*;
- the *average % difference*;
- the *average value difference* and the *median value difference*;
- the *interquartile range*;
- the *coefficient of variation*.

The main results of the evaluation process can be resumed as follows (see Table 2 as an example on SS):

- *FS*: the indicators calculated between survey data and FS data showed acceptable measures of similarity. Thirteen of seventeen variables had the same empirical distribution (*KS* not significant). The source results highly accurate, confirming its higher priority w.r.t. the other sources for Limited Companies.
- *SS*: the comparative analysis on the covered subset of units (sole proprietorships and family businesses, self-employed persons or partnerships) showed a positive mean difference of *Revenues* (+0.4%) and also on *Purchases of goods and services* (+0.5%), with a % difference for the *Value Added* of -1.6%. More detailed analyses at a 3 digits Nace and size level did not highlight systematic discrepancies. The highest differences were found for cost items and for units involved in service sectors (mainly due to different classifications of costs among the different sources). Other discrepancies were caused by the mismatch between the economic activity in BR and in the source (due to different criteria adopted for identifying the main economic activity).
- *Unico*: in this case, the analysis showed a good level of accuracy for variables on revenue and for *Value Added*, and a lower level for variables on costs, as in the administrative forms a different classification of

costs was required. However, the contribution of such information from Unico can be considered a good proxy of balance sheet data, that were not available for individuals and partnerships.

Table 2 – Quality Indicators for the main SBS variables (SME survey vs. Sector Studies). Year 2010

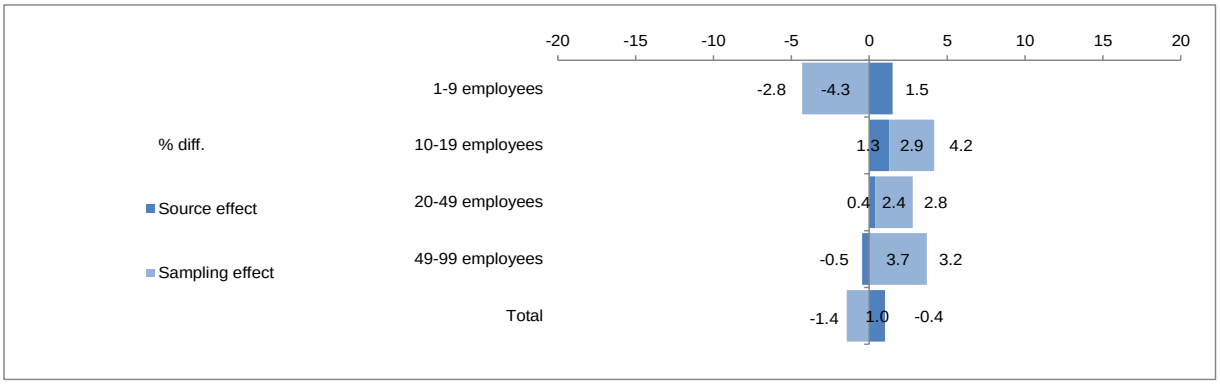
Main economic variables	KS*	± 5% (%units)	± 5% (%value)	%diff.	value diff. (000€)	median diff. (000€)	IQR diff. (000€)	CV diff.
Current earnings excl. VAT, inclusive of indirect taxes	1.0	90.6	94.0	0.4	1.5	0.0	0.0	89.6
Increase in fixed assets for internal works	0.4	99.3	47.0	25.5	0.1	0.0	0.0	339.4
Change in work in progress	0.5	96.6	522.9	-936.1	0.9	0.0	0.0	88.2
Changes in inventories (finished goods, raw materials and goods for resale)	2.1	82.4	31.1	-86.8	-2.5	0.0	0.0	-38.2
Other revenues and income (non-financial, non-overtime)	9.6	61.5	25.1	-19.5	-1.5	0.0	0.0	-40.1
Purchases of goods (a)	10.1	60.7	76.8	-2.4	-5.1	0.0	2.8	-26.4
Purchases of services (b)	5.7	23.0	26.6	10.8	6.5	0.2	7.8	17.8
Purchases of goods and services (a+b)	1.5	52.4	76.9	0.5	1.4	-0.1	5.3	88.0
Tenure Leasehold	4.7	80.5	68.1	2.2	0.3	0.0	0.0	48.2
Other operating expenses	15.1	13.1	7.2	-11.1	-1.2	0.0	3.9	-26.8
Cost of labor	4.8	85.1	81.3	1.6	1.0	0.0	0.0	26.5
Depreciation and amortization	3.6	67.8	60.3	-8.3	-1.2	0.0	0.0	-21.9
Value Added	**1.1**	**52.1**	**48.9**	**-1.6**	**-1.9**	**0.2**	**5.0**	**-46.4**
EBITDA	1.4	45.6	38.2	-4.7	-2.9	0.1	4.5	-29.7
Net Operating Margin	2.1	42.8	36.4	-3.1	-1.5	0.0	5.1	-61.1

* Threshold Value=1.6

Additional analyses on accuracy aimed at measuring the statistical effects on target estimates (*totals*) due to the use of a multi-source database. Actually, the integration of the available sources allows to obtain a unit-level census containing information on the key SBS for about the whole SMEs target population: we are then able to compare survey estimates with administrative-based estimates in order to have an overall evaluation of two main effects: (1) the one due to the replacement of the survey measurements with admin data measurements (*source effect*), and (2) the one due to the use of census data instead of using sampling weights (*sample effect*). Let Y be the target variable: the total estimate of Y based on admin data is $\boldsymbol{Y_{admin}} = \sum_{i=1}^{N} \boldsymbol{y_i^a}$, where *N*~4,300,000 is the number of SMEs in the target population; the Y estimate based on SME survey data is $\boldsymbol{Y_{SME}} = \sum_{i=1}^{n} \boldsymbol{y_i w_i}$, where *n*~105,000 and w_i are the sampling weights. The difference between $\boldsymbol{Y_{admin}}$ and $\boldsymbol{Y_{SME}}$ can expressed as: $\boldsymbol{Y_{admin}} - \boldsymbol{Y_{SME}} = \left(\sum_{i=1}^{N} \boldsymbol{y_i^a} - \sum_{i=1}^{n} \boldsymbol{y_i^a w_i}\right) + \left(\sum_{i=1}^{n} \boldsymbol{y_i^a w_i} - \sum_{i=1}^{n} \boldsymbol{y_i w_i}\right)$. The first expression in brakets represents a measure of the *sampling effect* (having imputed the ~5% of not covered units of each Y via within-cell median imputation); the second component is a measure of the *source effect*. As an example, let consider the variable *Value Added* (see Figure 1): preliminary estimates showed the prevalence of the sampling effect on the source effect. Overall, the first effect was equal to 1.4%, the second one to +1.0%, and both contribute to a total difference of -0.4%. The effect of sample weights has a negative impact (-4.3 %) on the class 1-9 employees and was positive for the upper classes. The source effect was almost always positive: administrative-based estimate is higher than 1% than the survey estimate, and the difference decreases with increasing company size. It goes from +1.5% for micro enterprises to +0.4% for medium enterprises (class 20-49 employees) and becomes negative for those over 49 employees. At the sectorial level there is a prevalence of the sample

effect in service economic activities; this effect is almost always opposite to the effect of replacing the data form admin sources, and it is greater in sectors with a high concentration of micro-enterprises.

Figure 1 – Value added estimate in survey (Y_{SME}), from administrative data (Y_{admin}) and decomposition of the total difference by size. Year 2010



2.3. Improving accuracy: some issues on data editing

As a part of the quality assessment process, a data editing and imputation (e&i) strategy has been designed in the frame context. It is known that admin data are subject to similar errors than statistical data, like missing data, measurement and process errors, whose sources relate to the specific admin data "production" process. In addition, there are other types of error (e.g. identification and aggregation errors, see for example Zhang, 2012) which specifically relate to a multi-source information context like the one characterizing the frame, as they mainly derive from the sources integration process. As for statistical surveys, the objective of e&I in this context in not only to identify and treat errors, but also to investigate their possible sources in order to improve the future use of the admin data. To this aim, the e&i strategy designed for the frame consists of the combined use of different methods (following the general recommendations of EDIMBUS, 2007), adapted to the specific information context. As an example, *selective editing* (EDIMBUS, 2007) is being adopted to identify influential anomalous behaviors (by estimation domains): given the size of the admin data bases, and the fact that the enterprises cannot be re-contacted, the inspection of the influential units has the main purpose of identifying the possible error sources (e.g. lacks in the harmonization process of variables/units definitions, or legal constraints, etc.), better understanding the nature and the characteristics of the admin data themselves, and improving the overall use of the sources for the specific SBS purposes. After a preliminary editing phase, aiming at eliminate formal and obvious errors from the data (e.g. duplicated units, incoherent/unbalanced variable values, unity measure errors), a first application of a simple selective editing principle has allowed to verify the potentials of this approach in terms of in insightful analysis of the admin data. Influential units are identified as follows. Let Y_k the value of variable Y in the source S_k, and Y_{SME} the corresponding value in the SME survey: units are ordered based on the difference $|Y_{jk}-Y_{jSME}|$. Let $D_k = \frac{TY_{SME}-TY_{Sk}}{TY_{SME}}$ be the relative difference between the estimates of Y total in S_k and in the survey, by domain. Given a threshold *k*, the set of influential units consists of the minimum number of units to be corrected so that $D_k<k$ in all domains. In Table 3 the number of units with influential differences requiring a correction at 2 digits Nace and (*k=5%*) for *Revenues, Purchases of goods and services, Value Added* are shown. Based on these preliminary evidences, in the final e&I strategy, a multivariate selective editing approach is being developed, where for each unit, the information available in all the overlapping sources is used to identify the units to be interactively revised. Robust regression-based imputation models are under evaluation to deal with total (due to non-coverage problems) and item (due to sources incompleteness) non responses.

Table 3 – Number and percentage of units with influential discrepancies which need correction, by source. *k=0.05.* Year 2010

Source		Variable		
		Revenues	*Purchases*	*Value Added*
FS	*n. units*	4	42	73
	%units	0.02	0.025	0.44
SS	*n. units*	17	1,002	290
	%units	0.06	4.9	1.07

References

BLUE-ETS (2012). *Deliverable 4.2: Report on methods preferred for the quality indicators of administrative data sources*.

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. *UNECE Work Session on Statistical Data Editing*, Ljubljana, 9-11 May.

EDIMBUS (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Prepared by ISTAT, Statistics Netherlands, and SFSO.

Essnet Admin data (2013). *WP6 - Quality Indicators when using Administrative Data in Statistical Outputs, Deliverable 6.5 / 2011:Final list of quality indicators and associated guidance*. June 2013.

Nordbotten, S. (2010). The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, in Carlson, Nyquist and Villani (eds), *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn,* pp. 205-225.

Zhang, L.C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, Vol. 66, N. 1, pp. 41–63.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics – Administrative data for statistical purposes,* John Wiley and Sons, Chichester.