rivista di statistica ufficiale

n. 1
2016

Temi trattati

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 1
2016

Temi trattati

Istat

Per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione della Rivista di Statistica Ufficiale
Istat – Via Cesare Balbo, 16 – 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 1/2016

Periodico quadrimestrale
ISSN 1828-1982

Registrato presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Dedicato a Gilda Sonetti

The new statistical register "Frame SBS": overview and perspectives[1]

Orietta Luzi,[2] Roberto Monducci[3]

Abstract

The paper summarizes the main features of the new statistical register on economic accounts of Italian small and medium enterprises (SMEs) developed at Istat in 2013-2014. The register, which is called "Frame SBS", allows to annually estimate the main variables of the economic accounts based on the massive use of microdata from integrated administrative sources. The sampling survey on SMEs is used as complementary source of information for estimating those variables which cannot be directly obtained from administrative sources. For methodological and technological details, the paper refers to the other papers published in this Volume of the Statistical Review. The paper highlights the role of the register in the area of business statistics, focusing on its potentials in terms of integrability with other sources for complex economic analyses.

Keywords: Structural Business Statistics, Administrative data, Data integration, Economic analysis.

Sommario

Questo lavoro riassume le principali caratteristiche del processo di produzione del registro statistico sui conti economici delle piccole e medie imprese (PMI) realizzato presso l'Istat nel periodo 2013-2014. Il registro, chiamato "Frame SBS", permette di stimare annualmente le principali variabili di conto economico delle imprese sfruttando massivamente microdati ottenuti da più fonti amministrative. L'indagine campionaria sulle PMI è usata come fonte complementare per la stima delle voci che non possono essere direttamente ottenute dalle fonti. Per i dettagli metodologici e tecnologici, il lavoro rimanda agli altri lavori pubblicati in questo numero della Rivista. Il lavoro evidenzia il ruolo del registro nell'area delle statistiche economiche sulle imprese, concentrandosi sulle sue potenzialità in termini di integrabilità con altre fonti per analisi economiche complesse.

Parole chiave: Statistiche Strutturali sulle Imprese, Dati Amministrativi, Integrazione, Analisi Economica

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

[2] Istat, e-mail: luzi@istat.it.

[3] Istat, e-mail: monducci@istat.it.

1. Introduction

The primary use of administrative data (*admin data* hereafter) for statistical estimation purposes in the business area is proven to produce significant benefits in terms of data quality and costs, and to greatly widen the possibilities of statistical and economic analyses. Concerning the latter, very detailed information across different dimensions (spatial-demographic and longitudinal) becomes usually available, and more reliable analyses can be performed; on the other hand, since admin data are commonly stable (over the time and in terms of information contents) they are also very suited for detailed longitudinal studies. Concerning data quality and costs, the increased completeness of statistical databases obtained on the basis of admin sources commonly produces a number of benefits as a result of the availability of census-like information on (sub-sets of) economic variables. This information also represents useful auxiliary information which can be used to increase the efficiency of samples (e.g. by focusing them on specific uncovered sub-populations or variables). In addition, in multi-source statistical databases relationships between phenomena can be estimated more reliably.

From a methodological point of view, when admin data are used as primary source of information for estimation purposes, possibly in combination with other sources, new and challenging issues need to be handled (see among others Wallgren *et al.* 2007, and Li-Chun Zhang 2012, for an overview and a discussion). Actually, since admin data are gathered for other purposes than statistical ones, additional data analysis and data processing are needed to ensure their statistical usability, such as harmonizing concepts and definitions with respect to the target statistical units and variables, matching classifications, data editing/data validation, data modelling and estimation, etc. In addition, as usually any admin source does entirely cover the variables/population of interest, information from different sources (admin archives and/or direct survey data) needs to be integrated together. In this case, additional methodological issues need to be handled in order to ensure data consistency, i.e. coherent data at micro and aggregate level. At micro-data level, micro-integration methods have been introduced, consisting in a set of approaches aiming at "*improving the data quality in combined sources by searching and correcting for the errors on unit level*" (Bakker 2010). Different micro-integration approaches can be adopted depending on the specific nature of the errors (see Li-Chun Zhang 2012, and Di Zio *et al.* 2014, for additional discussion of these issues). Moreover, as not all the target information is usually covered by the integrated sources, non-responses (either item or total) may result. Not covered information can be treated by either mass imputation (Pannekoek 2011), consisting in the prediction of microdata based on suitable statistical models ensuring consistent estimates at any level of detail, or by using appropriate estimation strategies, which exploit as much as possible the available auxiliary information (see among others Kloek *et al.* 2013).

In this paper we focus the attention on the integrated use of admin data for estimating structural business statistics (*SBS* hereafter). In this context, Costanzo (2011) provides an overview of existing practices in the European Countries, highlighting that, despite the availability of relevant, high quality and suitable admin data in most Countries, their potential is not yet fully exploited by the large majority of Member States. Portugal is the only European Country which has adopted an SBS estimation strategy where surveys have been completely replaced by admin sources directly provided by the businesses data owners

(Chumbau *et al.* 2010). Only France (Chami 2010) and Nordic Countries like Sweden, Denmark, the Netherlands and Norway (Wallgren *et al.* 2007) have implemented integrated systems for estimating SBS which are based on the primary use of admin sources, possibly complemented by sampling surveys to investigate either specific sub-populations or peculiar variables.

In Italy, in the SBS area a number of admin sources nowadays provide high quality information on businesses at high level of detail (Monducci 2010; Luzi *et al.* 2013). This fact has allowed the Italian National Statistical Institute (Istat) to gradually revise its estimation strategy in this context, moving from a production model essentially based on direct survey data complemented by admin information, to a new model where admin sources are extensively used, complemented by direct survey data. In 2013, Istat has actually developed a new statistical register for the annual production of profit-and-loss accounts of small and medium enterprises (SMEs hereafter). This register (called *Frame SBS*) massively uses firm-level admin and fiscal data as primary sources of information to estimate key SBS, while sample data on SMEs are used for estimating those items which are not available in the admin archives. Based on the *Frame SBS*, starting from the 2011 reference year, estimates of key SBS are available at an extremely refined level of detail, as they are obtained from a complete micro-data set. As expected, the design and implementation of the register has required an initial high investment in terms of methodological, technological and operational innovation, nevertheless it has determined substantial gains in terms of accuracy (as estimates of the key SBS are free of sampling errors) and consistency of estimates over time and among statistical domains (including National Accounts). The detailed and comprehensive information which is at the moment available in the *Frame SBS*, possibly integrated with other sources of information, represents a key factor in order to better analyze the characteristics and behavior of the Italian economic system (Monducci 2015).

This paper provides and overview of the main aspects which have been handled for the development of the *Frame SBS*, and illustrates the potentials benefits related to its usage in the Istat system of business statistics. Specific methodological and technological aspects are treated more in depth in the other papers included in this volume. The paper is structured as follows. Section 2 outlines the *Frame SBS*, with an overview of the main methodological and operational innovations which have been implemented for its development. In Section 3 the potentials of the new register in terms of further integration and economic analysis are discussed. Main conclusions and future work are reported in Section 4.

2. The statistical register Frame SBS

In Italy, traditionally SBS are estimated based on direct surveys. Concerning SMEs, a sample survey annually investigates about 105,000 SMEs (enterprises with less than 100 persons employed) in the industrial, construction, trade and non-financial services sectors (about 4.3 million of units as target population). In this context, admin archives are essentially used as complementary sources of information to compensate for nonresponse (see Curatolo *et al.* 2016 for a detailed description of the survey). The main issues in this

context relate to the high burden on enterprises which determines exceedingly low response rates and high sampling errors on parameters' estimates (variables' totals).

The increased stability, timeliness and quality of external admin sources on enterprises profit-and-loss accounts, such as the *Financial Statements (FS)*, the *Sector Studies Survey (SS)*, the *Tax returns forms* (UNICO) and the *Social Security data* (*SSD*) (for more details, see Curatolo *et al.* 2016 in this volume) opened the floor to a complete revision of the Istat approach to the SBS estimation on SMEs: in the new estimation strategy, admin data cover the *core* SBS information and the sample survey is used to investigate variables which are not available in the admin sources. Actually, by combining the admin sources, about 95% of the SMEs' target population is covered each year.

As the considered sources are partially overlapping (some of them provide information on the same variables on common SMEs sub-populations) a quality assessment process on each candidate data source has been initially performed, aiming at selecting the "best source" for each sub-population and variable. The evaluation process (for more details, see Curatolo *et al.* 2016 in this volume) has been based on a set of quality criteria such as *relevance*, *coverage* (in terms of target population units), *completeness* (in terms of covered information on target statistics), *accuracy, timeliness*, *integration* (extent to which the data source is capable of undergoing integration or of being integrated) (Istat 2015). The quality assessment process included a "harmonization" step aiming at reconciling the admin and the statistical definitions as described by the SBS regulation. Based on this process, "priorities" have been assigned to each source: for each source only some variables for some sub-populations of businesses have been considered reliable enough, and a specific priority among the considered sources has been established.

In Figure 1 the coverage and the priorities among sources are graphically represented: for each source *p* (*p= FS, SS, UNICO/IRAP*), the non-dotted areas correspond to the covered SME sub-populations, identified based on the Business Register[4]. The variables Y_j^p (j=1,…,k) represent the SBS variables covered by the *p* considered sources at microdata level. The SSD archive contains information on employment costs for all the SMEs with at least 1 employee[5]. The variables Y_j^{SME} relate to the SBS information collected by the annual sample survey on SMEs.

Based on this framework, the annual production process of the statistical register has been implemented. It consists of three macro-phases:

1. *Sources' acquisition and standardization;*
2. *Sources' integration;*
3. *Estimation.*

2.1 Sources' acquisition and standardization

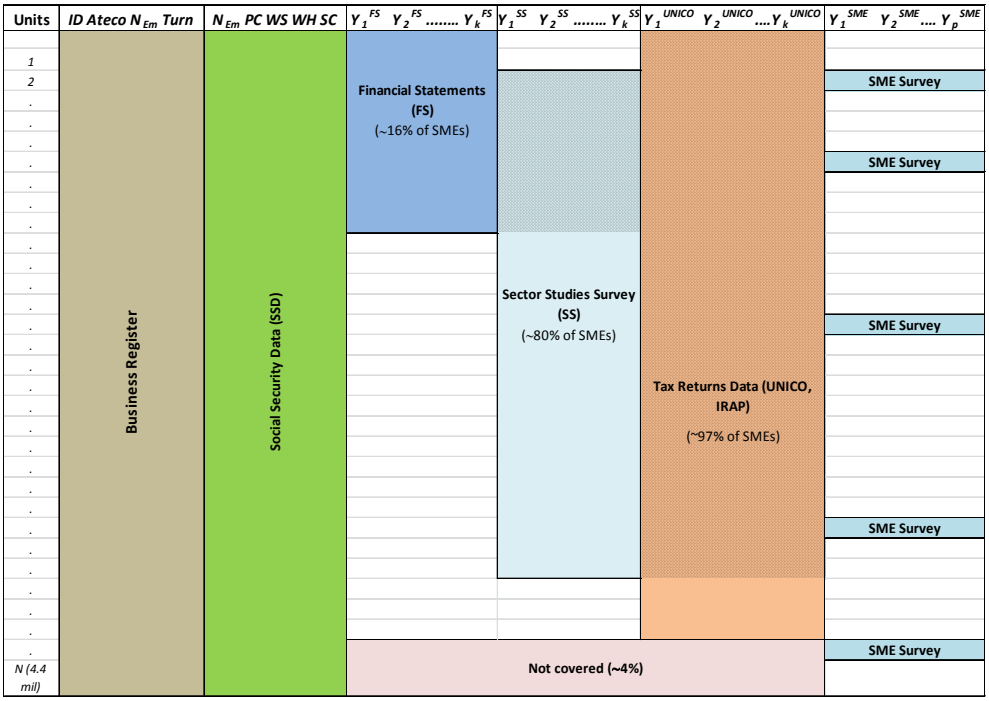
In this macro-phase (for more details, see Altarocca *et al.* 2016 in this volume) of the production process, the admin sources are acquired from the Istat centralized Business

[4] Containing structural information on enterprises such as: Economic Activity (Ateco), Number of Employees (Nem), Turnover (Turn).

[5] Personnel Costs (PC), Wages and Salaries (WS), Worked Hours (WH), Social Contributions (SC).

Registers sector, which is in charge of performing first data treatments and of uniquely identifying statistical units in each source. The sub-set of items required for the SBS estimation purposes is drawn from each source. Selected data are then subject to a structured set of activities aiming mainly at harmonizing variables and deriving SBS items, verifying that information relating to the same unit is unique, identify and eliminate possible inconsistent information reported within each unit, eliminate unusable information (for more details, see Sanzo *et al*. 2016 in this volume).

Figure 1 – The Frame SBS information framework



2.2 Sources' integration

At this stage of the production process (for more details, see Altarocca *et al.* 2016 in this volume), the sources' data are combined together, and a data validation process on the integrated data is performed. The aim is to ensure the consistency of related information coming from different sources, and to identify possible measurement errors in the data. The data validation process includes the harmonization of information on employment and labour cost coming from admin and fiscal data with respect to the corresponding one coming from the *Social Security Data* (for more details, see Arnaldi *et al.* 2016 in this volume). Outliers and influential data are selected on the distributions of economic parameters like per-capita labor cost and (labour) productivity, in order to identify possible measurement errors in variables.

At the end of these processes, two sub-groups of items are identified: the first group (referred to as *main economic aggregates*) is represented by the key SBS variables[6] which are extensively covered at microdata level by the integrated sources; the second sub-group relates to the *components of the main economic aggregates,* which are characterized by inadequate coverage rates and/or quality in the admin sources.

2.3 Estimation

According to the above variables' classification, a *mixed* estimation procedure has been adopted.

For the *main economic aggregates*, a predictive approach based on *mass imputation* has naturally allowed to build a complete micro-data file: in this phase, non- available information is predicted on the basis of admin data using a combination of different imputation techniques, which are applied to separate groups of variables taking into account their distributional characteristics and their relations with other variables (for more details, see Di Zio *et al.* 2016 in this volume).

For the *components of the main economic variables*, domain estimates at the detail level required by the SBS Regulation are obtained based on a design based/model assisted approach consisting in the use of a *projection estimator* (for more details, see Righi *et al.* 2016 in this volume) which exploits the SME survey data while ensuring consistency with respect to the main economic aggregates taken as auxiliary information.

It has to be underlined that the *Frame SBS* statistical production process exploits innovative IT solutions based on the development of a new data warehouse of integrated SBS information which is well-suited for supporting the management of modules in generic workflows (for more details, see Sanzo *et al.* 2016 in this volume).

The *Frame SBS,* when combined with the data from the annual survey on profit-and-loss accounts of Large Enterprises (enterprises with more than 100 employees) currently represents the reference framework for the convergence and consistency of many surveys on specific economic topics. Moreover, its integrated use in combination with data from other statistical registers (referring to both structural and short-term trends), has opened new perspectives for Istat data users, as illustrated in the next section.

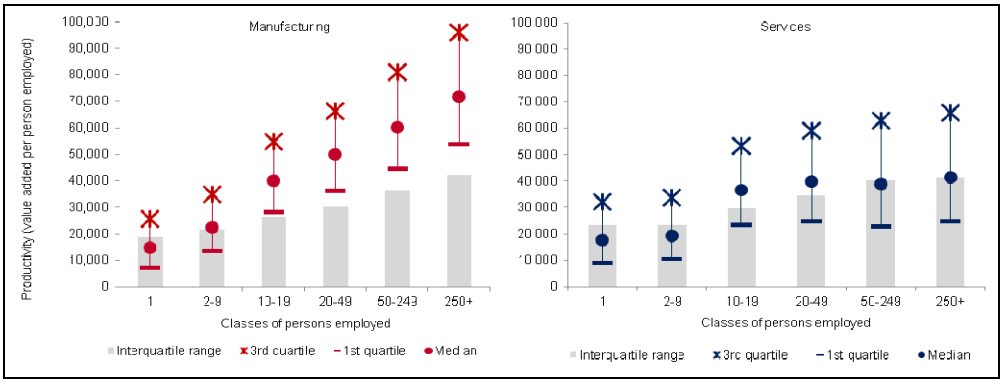
3. Potentials of the statistical register Frame SBS for economic analysis

The availability, on an annual basis, of main profit-and-loss accounts data on all companies active in Italy allows to carry out insightful analyses on both business structure and dynamics. As for the former, it is possible to assess the degree of heterogeneity within the business system, identifying the better- and worse- performing segments (e.g. sectors, clusters, etc.).

[6] Income from Sales and Services (Turnover), Changes in stocks of finished and semi-finished products, Changes in contract work in progress, Changes in internal work capitalized under fixed assets, Other income and earnings (neither inancial, nor extraordinary), Purchases of goods, Purchases of services, Use of third party assets, Changes in stocks of raw materials and for resale, Other operating charges, Personnel Costs.

In order to better illustrate the information potential of the *Frame SBS* register, Figure 2 reports some statistics about the distribution of the labour productivity by firms' size classes in 2013, in the manufacturing and services sectors. Besides confirming the well-known positive correlation between firm's size and productivity, the data show the heterogeneity within all size classes, revealing for instance that with the exception of the micro enterprises segment, in any other size class the most productive firms (i.e. the ones belonging to the fourth quartile of the productive distribution) perform better than the median firm of the next higher size class.

Figure 2 – Value added per person employed, by size classes – Year 2013 (euros)



Source: Istat

With regard to the dynamic analysis, the statistical register *Frame SBS* allows to longitudinally evaluate the performance of single production units, pointing out for example the firm- and sector-level developments underlying the aggregate dynamics. The latter element is particularly important for an assessment of the resilience and vulnerability of the Italian business system, as the *Frame SBS* makes it possible to monitor on an annual basis the relative competitive position of all the Italian firms within their own sector or across the entire business system, in terms of profitability, productivity and other economic performance indicators.

The register makes it possible to evaluate whether (and how) the Italian productive system that is coming out from the crisis differ from the one that entered it, for example in terms of number and size of the units, employment, and (labour) productivity. In 2010-2013 about 21% of the "persistent" firms increased the number of persons employed. From a sector perspective, the share of firms with a net job creation is higher in manufacturing (30%) than in the service sector (19,7%). These changes have partially modified the structure of Italian firms by size. In the same period, over 50% of firms increased their value added, and 15% showed a simultaneous increase in terms of value added and employment. On the other side, 43% of firms have experienced a fall both in value added and employment.

Finally, it has to be underlined that the statistical register *Frame SBS* provides a "structure information cornerstone" for further integrations with other firm-level statistical registers, referring both to structural and short-term economic events. This feature allows to identify the developments underlying some important recent trends, also taking account, in

a multidimensional way, the structural features and the strategic choices used by firms to cope with those trends.

4. Conclusions

The statistical register *Frame SBS* can be considered an advanced example in Istat of statistical production based on the direct use of multi-source, administrative data. The adoption of a mixed estimation strategy, exploiting as much as possible the available information, and the use of innovative methodological approaches for data validation, data prediction and estimation, ensure high levels of quality for the final outputs.

The benefits associated to the use of the *Frame SBS* mainly relate to the increased accuracy of cross-sectional estimates of the main SBS aggregates, and better coherence over time of the SBS estimates. At present, much more information is available not only for external data users, but also for Istat internal users, since the *Frame SBS* represents a source of auxiliary information for statistical production processes in other economic domains.

In the long term, further methodological developments are expected to produce additional benefits. In particular, a reduction of survey costs and statistical burden, associated to a further increase of data quality, will result from: the extension of the direct use of administrative data to other SBS variables and sub-populations (e.g. enterprises with more than 100 persons employed); the revision of the design of the overall SBS estimation strategy (with particular reference to the sample design for small enterprises); the adoption of innovative estimation approaches (like Small Area Estimation models, see Luzi *et al.* 2015) for specific, complex business accounts variables. A further area of development will concern the problem of identifying appropriate indicators for measuring the quality of the *Frame SBS* outputs, since they are obtained based on the direct use of multi-source administrative data. In this context, the results of European projects like the Essnet AdminData (2013) and the Essnet BLUE-ETS (2012) could be profitably exploited.

From an analytical point of view, the *Frame SBS* represents a powerful source of micro level business information for economic analysis and to support to policy advice. Its peculiar characteristics are: fully consistency with official figures, harmonization of all variables across different data sources, and totally scalable data from the micro to the meso up to the macro level of economic analysis. In addition, given its register-based nature, the *Frame SBS* represents the natural hub of an open information system that can be integrated with other variables coming from short terms survey data, dedicated qualitative surveys and policy oriented indicators. According to the growing need for micro-integrated registers designed for economic analysis, a short-term implementation plan of *Frame SBS* for economic analysis has been established, with the development of a set of further indicators aimed at assessing the competitiveness and the growth potential of Italian firms according to three relevant dimensions of enterprise's activity: employment and wages, engagement in foreign trade, business location. The output of this project will be an open system accessible through a dedicated research lab inside ISTAT where all relevant stakeholders and independent researchers will actively participate to expand and deepen our knowledge about the resilience features and evolutionary patterns of the Italian economy.

Riferimenti bibliografici

Arnaldi S., C. Baldi, R. Filippello, L. Mastrantonio, S. Pacini, P. Sassaroli e F. Tartamella. 2016. The labour cost variables in the building of the Frame. *Rivista di Statistica Ufficiale Istat,* n. 1/2016.

Altarocca F., D. Bellisai, A. Laureti Palma e R. Sanzo 2016. New experiences in the production of business statistics: the construction of the 'Frame' and the SBS-datawarehouse. *Rivista di Statistica Ufficiale Istat,* n. 1/2016.

Bakker B. F. M. 2010. Micro-Integration: State of the art. In *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, http://www.cros-portal.eu/content/wp1-state-art.

Chami S. 2010. Reengineering French structural business statistics - an extended use of administrative data. *European Conference on Quality in Official Statistics* (Q2010), Helsinki.

Chumbau A., H.J. Pereira e S. Rodrigues S. 2010. *Simplified Business Information (IES): Impact of Admin Data in the production of Business Statistics*. Presented at the Admin Data ESSnet Seminar "Using administrative data in the production of business statistics. Member states experiences", Rome, March.

Costanzo L. 2011. *An Overview of the Use of Administrative Data for Business Statistics in Europe*. Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session CPS035).

Curatolo S., V. De Giorgi, F. Oropallo, A. Puggioni e G. Siesto. 2016. Quality analysis and harmonization issues in the context of the SBS frame. *Rivista di Statistica Ufficiale Istat,* n. 1/2016.

Di Zio M., u. Guarnera e R. Varriale. 2016. The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources. *Rivista di Statistica Ufficiale Istat,* n. 1/2016.

Essnet AdminData. 2013. *WP6 - Quality Indicators when using Administrative Data in Statistical Outputs, Deliverable 6.5 / 2011:Final list of quality indicators and associated guidance*, available from http://essnet.admindata.eu/WorkPackage?objectId=4257.

Essnet BLUE-ETS. 2012. *Deliverable 4.2: Report on methods preferred for the quality indicators of administrative data sources*. http://www.blue-ets.istat.it/index.php?id=7.

Kloek W. e S. Vâju. 2013. The use of administrative data in integrated statistics. *NTTS - Conferences on New Techniques and Technologies for Statistics*. Brussels, 5-7 March.

Istat. 2015. *Linee guida per la qualità dei processi statistici di fonte amministrativa*. http://www.istat.it/it/files/2010/09/LineeGuida_v.1.0_Luglio_2015.pdf.

Luzi O., F. Solari e R. Monducci. 2015. Small area estimation for business statistics: new perspectives at Istat. Invited paper. *ITACOSM 2015 - 4th ITAlian Conference on Survey Methodology*. Rome, 24-26 June.

Luzi O., U. Guarnera e P. Righi. 2014. The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. *European Conference on Quality in Official Statistics* (Q2014). Vienna.

Luzi O. e M. Di Zio. 2014. Editing administrative data. In *Memobust Handbook on Methodology of Modern Business Statistics.* http://www.cros-portal.eu/content/handbook-methodology-modern-business-statistics.

Luzi O., M. Di Zio, F. Oropallo, A. Puggioni e R. Sanzo. 2013. Integrating administrative and survey data in the new Italian system for SBS: quality issues, *3rd European Establishment Statistics Workshop (EESW 2013)*, 9-11 September. Nuremberg, Germany.

Monducci R. 2015. Measuring economic, social and environmental resilience. *Joint IEA/ISI Strategic Forum 2015 and High-Level Expert Group on the Measurement of Economic Performance and Social Progress Workshop.* http://www.oecd.org/statistics/measuring-economic-social-progress/IEA-ISI%20Strategic%20Forum%20and%20Resilience%20Workshop%20agenda.pdf. Rome, 25-26 November.

Monducci R. 2010. Statistiche ufficiali e analisi della competitività del sistema delle imprese: aspetti concettuali, problemi di misurazione, strategie di miglioramento della qualità. *Atti della X Conferenza nazionale di statistica*, Roma, dicembre.

Pannekoek J. 2011. Models and algorithms for micro-integration. In *Report on WP2: Methodological developments. Essnet on Data Integration.* http://www.cros-portal.eu/content/wp2-development-methods.

Rao J. N. K. 2003. *Small Area Estimation*. New York: John Wiley and Sons.

Righi P. 2016. Estimation procedure and inference for component totals of the economic aggregates in the new Italian Business frame. *Rivista di Statistica Ufficiale Istat,* n. 1/2016.

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

Zhang L.-C. 2012. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*. 66; 41-63.

Quality analysis and harmonization issues in the context of "Frame SBS"[1]

Silvana Curatolo, Viviana De Giorgi,
Filippo Oropallo, Augusto Puggioni, Giampiero Siesto[2]

Abstract

The paper describes the results of the quality analysis behind the process of integration and harmonization of administrative data with sample survey of small and medium-sized enterprises (SME). The integration process involves both Structural Business Statistics (SBS) and National Accounts (NA) to reduce the distance from the different estimates. The quality analysis considers a study of the coverage of administrative data, the harmonization of definitions and comparison analysis among administrative and survey data with a subsequent analysis of the distribution of differences in critical domains to verify the absence of systematic errors. The final quality analysis explores separately the impact of the new sources on the SME estimates by decomposing the difference in two parts: (1) the source effect due to the use of administrative data and (2) the sampling effect due to the passage from the sample to census estimations.

Keywords: Micro integration, Administrative sources, Quality indicators

Sommario

Il documento descrive i risultati delle analisi di qualità alla base del processo di integrazione e armonizzazione di dati amministrativi con l'indagine sulle piccole e medie imprese (PMI). Il processo di integrazione riguarda sia le statistiche strutturali sulle imprese e sia le stima effettuate nell'ambito dei conti nazionali al fine di ridurre la distanza tra le diverse stime. L'analisi di qualità riguarda lo studio sulla copertura delle fonti amministrative, l'armonizzazione delle definizioni delle variabili e con il confronto tra i dati amministrativi e quelli di indagine con analisi sulle distribuzioni delle differenze in domini critici al fine di verificare l'assenza di errori sistematici. Si mostra, infine, la scomposizione della differenza tra stima campionaria e la nuova stima censuaria in due componenti: (1) "effetto fonte" dovuto dell'utilizzo di dati amministrativi ed (2) "effetto campionario" dovuto al passaggio da stime campionarie a stime censuarie.

Parole chiave: Micro integrazione, dati amministrativi, indicatori di qualità

[1] Although the paper is the result of the work of all the authors, are to be given to: S. Curatolo paragraphs 5.1 and 5.3.1; V. De Giorgi paragraphs 1, 3, 3.1, 3.2, 3.3 and Annex A; F. Oropallo paragraphs 4, 5.2, 6 and 7; A. Puggioni paragraphs 2, 5.3 and Annex B; G. Siesto paragraphs 1, 5.3 and Annex B. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

[2] Istat: curatolo@istat.it, degiorgi@istat.it, oropallo@istat.it, puggioni@istat.it, siesto@istat.it

Introduction

In recent years Istat has intensified the use of administrative sources to support the processes of production of business statistics (Casciano et al. 2012, Luzi et al. 2013). In particular, the use of financial statement data and economic data from Sector studies and Tax returns, made it possible to estimate the major items of the income statement (turnover, costs, value added, Change in stocks, etc.) for almost all of the statistical units of the Italian Statistical Business Register (Asia).

Thanks to the integration process of all sources, a coverage and a comparison analysis has been done in the context of the sub-working group relating to the quality analysis and preliminary estimation of the frame of SBS (Structural Business Statistics).

Preliminary estimates and comparisons with survey estimates has been done for the main SBS variables and in particular for the variable value added per worker has been evaluated the "sampling effect" (sum of all observations with respect to the weighted sum) respect to the "source effect" (replacing survey data with harmonized administrative data).

1. The survey on small and medium enterprises: from sampling design to final estimations

This paragraph describes the production process of the Small and Medium-sized Enterprises survey (SME), by explaining the sample design, the rules for collecting and processing of data, the integration with administrative sources and the estimation procedure used in the SME survey. The paper outlines the domains of estimation for the SBS regulation, the main variables to be transmitted to Eurostat, a short description of SME data production process, the analysis of the available administrative sources having economic information on Italian enterprises and the problems encountered with their use.

The reference population of the SME sample survey is the enterprises with less than 100 persons employed running an industrial, commercial or services activity. This survey, together with the total survey on enterprises with 100 person employed and more (SCI) responds to the requirement of EU Council Regulation on SBS no. 295/2008.

The sampling design is a one stage stratified sampling, with an equal probability of selection for units; the strata are defined by the combination of the modality of the classification of the economic activity Nace Rev.2 (4 digit), size class and administrative region. The sample of the survey referred to the year t is extracted from Asia register in the year t-1, in accordance with the domains requested by the SBS regulation. For the reference year 2010, the sample consisted of 100,703 enterprises (about 2.3% of the number of enterprises of Asia).

The data collection of SME survey starts sending a letter to the sampled enterprises (in June of the year t+1), which contains the internal code (Asia enterprise code) and the password to access the website of the survey, where they can download the electronic questionnaire and, after compiling it, transmit it via the same website. Since the reference year 2010, the method for collecting questionnaires is totally electronic. The response rate of SME2010 was low, accounting for 38.4%. One of the reasons for this is linked to the complexity of the questionnaire. It consists, in fact, of more than 200 variables (regarding value and cost of production, employment by job category and sex, personnel costs broken

down into different items, external staff with related costs, investments during the year of reference by type of goods, expenditure on environmental protection and other information both quantitative and qualitative) due to the fact it has to be useful not only for complying the SBS regulation but also for national accounts.

The questionnaires, received through the website, are loaded into a RDBMS (Relational Database Management System) and then subjected to an E&I (Editing and Imputation) process that indicates errors and warnings for approximately 23% of the enterprises. Such cases are then solved interactively by data revisers, by comparing survey data with administrative ones or in some cases by contacting the enterprises. The validation process continues until the month of March of the year t+2.

The phase of integration of non-response begins after the revision, by using the administrative data from financial statements and Sector Studies surveys[3]. The available administrative data represent the macro items of the income statement (according to the scheme of the Fourth EU Directive). The integration of total non-response of SME survey is based on the combined use of administrative data and the donor technique. The phases of the integration process are the following:

1) in order to identify the set of non-response that has to be integrated by administrative data, we compare the validated data file of the SME survey with the files of the theoretical sample and that one used for monitoring arrivals;

2) then, we link the enterprise code, and extract information on the main economic activity of the enterprise and its occupational structure (number of persons employed and employees);

3) afterwards, we use data from financial statements and Sector studies survey to integrate non-response. Obviously, the sources overlap and therefore we need to assign a priority to them, on the basis of information on data quality, consistency between administrative and statistical definitions and the number of available items. The integrated file (no duplication of enterprise codes) will contain all the variables available from the administrative sources together with the structural information from Asia (Nace and employment);

4) total non-response is integrated by identifying for each non-respondent a donor with a similar economic profile (economic activity, size, administrative region). In case we cannot identify the donor enterprise for some strata, the procedure collapse the strata characteristics. The integration process is launched in two steps: first, for the enterprises with employees by selecting donors with personnel costs and, second, for enterprises without employees. Once the group of donors is identified for each stratum, a casual extraction is used to integrate the data of the non-respondent unit with the data of the donor, weighted on the relationship between the persons employed of the non-respondent and the donor persons employed.

The integration process ends with the replacement of the data calculated as described above with those actually declared by the enterprises in the administrative sources for the items of the income statement, while the subheadings are calculated through relations of composition or on the basis of indicators.

At the end of the non-response integration process, position indicators on the whole data

[3] See par. 3.2.

set (respondent and integrated enterprises) are calculated in order to identify potential outliers in the different domains, which could have a significant influence on the final estimates. The coverage of the sample through theoretical integration is 80.3%.

On the respondent units the calibration estimates are calculated to obtain final weights that, under certain assumptions, are corrective of residual non-response and error list. They ensure the respect of equality between the totals of the population (number of enterprises and persons employed) and the sample estimates.

The estimator of the total $Y_{(D)}$ referred to the domain D is

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

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

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

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

where:

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

The SBS data transmission to Eurostat (SME sample survey and SCI survey) takes place in June of year t+2 (18 months from the end of the reference year), as showed in Table 1, by defining the statistical confidentiality through the software Tau-Argus.[4]

The SBS Regulation requests the transmission of all the information estimated by the sample survey on the small and medium enterprise and the total survey on the larger enterprises according to the following domains:

Table 1 - Domains of estimation of the Structural Business Statistics (SBS) n° 295/2008

1 - Nace Rev.2 at 4 digits without distinction for size classes of persons employed (SME+SCI)

2 - Nace Rev.2 at 3 digits by for size classes of persons employed:

	size classes of persons employed						
	0-1	**2-9**	**0-9**	**10-19**	**20-49**	**50-249**	**250+**
Industry and construction *– sections: B,C,D,E,F*			SME	SME	SME	SME+SCI	SCI
Trade and services *– sections: G,H,I,J,K(division 66),L, M,N,P, Q,R,S (divisions 95 and 96)*	SME	SME		SME	SME	SME+SCI	SCI

3 - Nace Rev.2 at 2 digits by administrative regions without distinction for size classes of persons employed (SME+SCI; industry, construction and service

4 - Nace Rev.2 at 3 digits by administrative regions without distinction for size classes of persons employed (SME+SCI; trade)

[4] Software designed to protect statistical tables.

The SBS variables are required for the annexes 1 (services), 2 (industry), 3 (trade), 4 (construction) and 8 (business services that cover enterprises with 20 and more persons employed running in specific economic activities); the most important variables requested are the number of enterprises, turnover, value of production, gross margin on goods for resale, value added at factor cost, gross operating surplus, total purchases of goods and services, purchases of goods for resale, personnel costs, gross salary, hours worked, gross investment in tangible goods, number of persons employed and number of employees.

2. Analysis of SBS and National Accounts definitions and other constraints required for the compilation of National Accounts

National Accounts (NA) are the framework for the quantitative description of the economic and financial situation of an economic system, its components and the relationships between them established in a given period of time. The transactions carried out by economic agents (so called "institutional units") in relations with other resident in the economic territory units or with non-residents, as well as the changes that these transactions determine on the level and composition of the stock of assets and real or financial liabilities held by them are the matter of the measure.

The elements that characterize the system, in fact, are:

- Systematic and extensive detail in the description of the economy
- Harmonization of methodologies (comparability over time and space)
- Accuracy and precision of concepts, definitions, classifications and accounting rules
- Consistent, reliable and comparable description in terms of quantity of the economies of the EU countries.

According to the criteria of the European Union only a comprehensive measure of GDP makes such aggregate comparable between different Countries and can be used as an indicator for the calculation of contributions from the Member States to the Union, for the control of Maastricht parameters and the allocation of structural funds.

The search for exhaustiveness is guaranteed in the process of construction of the NA by:

- the compliance of definitions and methods with the rules stated by European system of national and regional accounts in the European Union (ESA);

- the measurement of the different components of the non-observed economy[5] and other forms of under-coverage connected to the quality and reliability of information sources.

The gaps between the SBS data and the final NA estimates are traceable and associated to differences in definitions set by the two different regulations, it is the above mentioned concerning SBS and Regulation (EU) No 549/2013 of 21 May 2013 on the European system of national and regional accounts in the European Union (ESA 2010). The two statistical domains serve different, but not conflicting, purposes: in particular ESA is an integrated representation of the economic system, seen as a system of flows between all the different actors, and then a number of strict rules and constraints must be met to assure the consistency.

[5] See for further details "Measuring the Non-Observed Economy - A Handbook" (OECD, 2002).

The main differences between NA, business accounting and SBS are:

1) the definition of production and value added does not totally correspond, since in NA some items need to be reclassified to property income (costs for financial leasing, interest, rent on land, and so on) some to current transfers (insurance premiums and claims);

2) NA estimates do not include capital gains and losses, that have to be eliminated from the values reported in business accounting;

3) a different valuation is required for the economic aggregates: value of production, value added, intermediate costs: producer prices in NA (i.e. net of subsidies on production and products received and including taxes on production and products paid), versus factor cost in SBS (i.e. including subsidies on products and production received net of taxes on products and production paid);

4) NA require that the accounts of enterprises are consistent with those of other enterprises and units in other institutional sectors. So, taxes and subsidies valued on the basis of accounts of enterprises must be consistent with those received by or paid by general government. In practice, this is not observed and a rule is needed in order to achieve consistency. Normally information from general government is more reliable than that from enterprises, and the data from business accounts is adjusted;

5) NA estimates need to be exhaustive, so that a number of specific adjustments are necessary to take into account underreporting of value added, income in kind supplied free of charge to employees, tips;

6) NA estimate domains are more detailed than in SBS: industries in NA are currently defined by Nace 4-digit and size classes (1-5, 6-9, 10-19, 20-99, 100-249, 250-499, 500 and beyond). Therefore the SBS Frame data need to guarantee the joint significance of data for 4-digits Nace Rev.2 domains and for 3-digits Nace Rev.2 and size class; as the data are also used to compile institutional sectors accounts, the legal form of the unit (“corporation”, “not corporation”) have to be considered, too.

The choice of the methodology for the construction of the SBS Frame has taken into account the needs of both SBS and National Accounts domains, since, when fully operational, the two systems should be integrated: in other words, the flow of data into the new system must provide an intermediate output that constitutes the common input for the compilation of SBS data on the one side, and of NA on the other. It is important to note, though, that to comply with ESA requirements, the valuation and/or classification of some entries of income statement needs to be modified: this generally involves not the "key-variables", but the elementary items that detail the main entries (e.g. the value of insurance premiums paid, rents paid on land, payments for financial leasing, and so on) and that are not always reported in administrative data. As a consequence, all such details have to be estimated in the SBS Frame, by means of a statistical approach grounded on the results of the sample survey.

In the past different approaches were followed. In the NA context a composite estimator was based on SME survey integrated with the population of companies (financial statements). In the SBS context the SME estimates were based on a calibrator estimator with the imputation of missing responses made by integration of financial statements and Sector studies survey (from Italian Revenue Agency) for unincorporated units and sole proprietorships (80% of the Statistical Business Register).

The availability of the new database has totally overhauled the previous approach in the general revision of NA for the transition to ESA 2010[6]: the total amount of the economic aggregates, and value added in particular, as reported in the SBS Frame has been incorporated as such in NA, to represent the observed part of the economic flows. As a consequence, in principle NA and SBS data for SMEs only diverge for some conceptual differences[7]. In general SBS definitions are consistent with business accounting, while national accounting differs on a number of points, because it has a different objective: national accounting aims to describe within a coherent framework all the activities of a country and not just those of an enterprise or group of enterprises. This objective of a consistent picture across all entities in an economy and their relationship with the rest of the world brings constraints which do not affect business accounting.

The revised approach to adjust value added in NA for the under–declaration of income of SMEs because of fraud[8] calls for supplementary information, not needed for SBS (enterprise income as well as other items of the balance sheet), that have to be included in the SBS Frame. The individual information provided by the new data base has also allowed to revise the procedure to adjust value added for the under–declaration of income of SMEs.

On the occasion of the benchmarking revision, the formula to define production and value added from have been modified [9], for three main reasons:

a) to achieve a better definition of production and intermediate costs according to the ESA 2010 definitions;
b) introducing some corrections to the aggregates in accordance with some reservations set by the European Commission on the compilation of Italian Gross national income (GNI);
c) moving from a definition of production at producer prices to one at basic prices.

A relevant innovation in the calculation of value added is the inclusion in the value of production of a share of revenues not related to the characteristic activity of the enterprise, and registered under the accounting item "Other revenues". Actually, only some of the transactions registered under the item have to be registered as production (for example, insurance claims, capital gains, financial revenues must be excluded), so that a detail of the elementary entries is needed, which is not, however, available nor in administrative nor in the survey data. As a consequence, a statistical procedure have been realized, based on the information available in the Notes to the financial statements of corporations, to elaborate an appropriate structure of the composition of the item "Other revenues" to be used in the SBS Frame to properly sub-divide its value. The development of the SBS Frame represents a substantial innovation in the information system on which NA estimates are grounded and

---

[6] The transmission to Eurostat of the new series is scheduled by September 2014. The first compilation of the SBS Frame is set for 2011, the year chosen as the benchmarking year for NA. The SBS Frame together with the census-like survey on big enterprises from 100 workers up (SCI) constitute the SBS system data base on business which NA estimates rely on.

[7] In the previous NA methodology per capita values stemming from business surveys are grossed up using estimates of labour input. The new procedure is grounded on an additive approach that distinguishes between observed and non-observed economy. Accounting data as supplied by the SBS, appropriately treated and integrated with other data sources to comply with ESA definitions, provide the total amount the aggregates, limited to the "observed" part of the economy; the results of underground (both concealed to tax authorities and generated by non-registered labour input) and illegal activities are then added.

[8] See "Economia non osservata nei conti nazionali" www.istat.it/it/archivio/175791.

[9] See "I nuovi conti nazionali in Sec 2010" www.istat.it/it/archivio/133556.

allows to substantially improves the quality of the accounts.

3. Coverage analysis

This part of the paper describes the results of the analyses - reached in the early stages of the working group (January 2013) - on the coverage of the population of the small and medium-sized enterprises sample survey for the reference year 2010[10], taking into account the administrative data available within Istat. In this context, the coverage is given in terms of number of enterprises, persons employed, employees, and variable VAT-Turnover as they come from Asia in the same reference year. The reference population consists of the enterprises with less than 100 persons employed.

The preliminary analysis has involved the identification of the field of observation in terms of economic activities to be covered (enterprises of industry and non-financial services private sectors) and the study of the available sources of data:

The sources of data identified for the study are the following:

- the Business Register (Asia register);
- the Financial statements of corporate enterprises;
- the Fiscal data (Sector studies and tax return data);
- the Social Security data.

As shown in Table 1, units without coverage represents about 4%. Main issues concerns some domains with a high presence of foreign companies. To guarantee coverage we have adopted two remedies only for *special purpose entities* with more than 3 employees: the first one is to recover the financial statements from other external source; the second one is to reconstruct information by using imputation methods.

Table 1 - Business Register units by source of administrative data. Year 2010.

ADMINISTRATIVE SOURCES	Units	%
Financial statements	718,538	16.2
Sector studies	2,930,988	66.0
Tax Return Data	627,671	14,1
No source	*166,686*	*3,8*
Total	**4,443,883**	**100.0**

3.1 Sources used and domains of coverage analysis

The target population is the same of the reference population used for the extraction of the sample of the survey on SMEs. For the year 2010 the total number of enterprises is 4,443,883 SMEs. For the 2010 SME survey sample, 100,703 unit are extracted: 37,922 are respondents and 42,986 are integrated by administrative sources (income statements and fiscal data).

[10] The year 2010 is the last year available at the date of the analyses.

The administrative data used for these analyses are the financial statements of corporate enterprises, the data of the fiscal survey named Sector studies; the fiscal statements called Modello Unico for corporate enterprises, unincorporated firms and sole proprietorships. In order to complete the analysis of coverage, both administrative and statistical sources (including the SME survey data) have been used. Another fiscal source, the regional tax on productive activities (Irap), has not been used in the early stage.

The coverage analysis is carried not considering the case of errors or data inconsistency, which means no editing has been done on data. By exploiting the results of Casciano at al. we use a hierarchy of source priority[11]. The analysis is conducted for all domains of reference for SBS, in terms of both the total number of enterprises and the sub-population of firms with employees. The statistical classification of economic activities used is Nace Rev. 2.0, and the domains are the following:

- SBS reference domain:
  - class of economic activity (4-digit Nace Rev.2)
  - group of economic activity (3-digit Nace Rev.2) by size-class of persons employed
  - division of economic activity (2-digit Nace Rev.2) by size-class of persons employed
  - division of economic activity (2-digit Nace Rev.2) by administrative region
- other (non-SBS) domains:
  - division of economic activity (2-digit Nace Rev.2) by legal form
  - division of economic activity (2-digit Nace Rev.2) by class of volume of business[12]

The information related to the number of persons employed, region, VAT-Turnover and legal form are derived from Asia of the reference year, not taking into account the information reported by the SME survey respondent[13].

3.2 Description of the sources

a) The Italian statistical business register (Asia)

Asia is the statistical business register of Italian enterprises, in accordance with the EEC Regulation no. 2186/93 of 22 July 1993 on "Community coordination of the development of business registers for statistical purposes". Asia represents an official source on the structure of the business population and demography that identifies the enterprises, and their statistical variables. As regards the SME sample survey, Asia has the role of the frame list for extracting the SME samples, as for all Istat business survey. It is also a reference to

[11] 1. financial statements of corporate enterprises, income statement; 2. Sector studies, form F; 3. Sector studies, form G; 4. Modello Unico, form PF-RG; 5. Modello Unico, form PF-RE; 6 Modello Unico, form SP-RG; 7 Modello Unico, form SP-RE; 8. Modello Unico, form SC-RS; 9. Modello Unico, form PF-CM; 10. Modello Unico, form PF-RF; 11. Modello Unico, form SP-RF; 11. Modello Unico, form SC-RF. Priorities are subject to continuous revision and updating, according to different selection criteria. One of them is according to the number of usable variables.

[12] Volume of business is the proxy value for turnover stored in Asia.

[13] The size class is calculated according to the Asia "old" computation of persons employed, in this case preferred to the "new" classification since the present analysis is on the reference year 2010, whose calculation method of the persons employed is the "old" one. We believe, however, that the differences are relatively negligible.

update structural information on enterprises (economic activity, persons employed, employees, etc.) and link through the fiscal code all administrative sources available (Istat, 2011, Nota metodologica). In this context the role of Asia is to extract the structural information and link the administrative data.

b) The financial statements of corporate enterprises

They are registered at the Italian Chambers of Commerce. The profit and loss account of the financial statement sets the costs against the revenues for the reporting period, and gives the result of economic operation management.

c) The Sector studies survey

This survey is conducted since 1990s by the Ministry of Economy and Finance and represents the tool the Fiscal Authority uses to detect the economic and fiscal parameters for professional activities, self-employed workers and enterprises, in terms of structural as well as economic characteristics, in order to assess their ability to produce income. Sector Studies aim to identify and quantify the voluntary concealment of business performance by comparing the declared economic results with some threshold defined by several direct and indirect indicators. There are some exclusion an non-enforceability principles, the most important one is the threshold for turnover (around 5,2 million euros). Nowadays Sector Studies cover almost all "market" enterprises.

The Sector studies questionnaire is organized into several forms, including those ones relating to economic and financial data used for this study:

- Form F - for enterprises in manufacturing, trade and services sectors;
- Form G - for professional activities.

d) The Modello Unico

It is a unified form of tax statement by which the natural and legal persons may submit tax returns, and VAT, Irap and withholding agents tax; the taxpayer obliged either to the compilation of Sector studies survey or its parameters, are also required to fill in the form of Sector studies together with the tax statement.

All income earners in a specific year are obliged to fill in the related form of the Modello Unico, and in general the form of Modello Unico to be filled in is different in accordance with the legal form of the taxpayer. For the purposes of this study, only the forms on the income statement of enterprises and self-employed are usable.

d.1) The Modello Unico for sole proprietorships (PF)

Individuals who fill this model are: (i) the natural persons obliged to keep accounting records, (ii) employees, and (iii) generally those who have earned income during the tax period. With regard to income statement and self-employment, the forms used in this analysis are the following:

- form PF-RE - self-employment income from the exercise of arts and professions;
- form PF-RF - income statement of ordinary accounting;
- form PF-RG - income statement of simplified accounting and flat rate schemes;
- form PF-CM - minimum taxpayers.

d.2) The Modello Unico for unincorporated enterprises (SP)

Unincorporated firms resident in Italy are required to fill in this form: partnerships; general partnerships and limited partnerships; shipping company; society of fact or irregular (treated as partnerships or partnerships depending on whether or not to run commercial

activity); unincorporated associations among natural persons; marital companies; European economic groups of interest.

With regard to income statement, the related forms used for this analysis are:

- form SP-RE - self-employment income from the exercise of arts and professions;
- form SP-RF - income statement of ordinary accounting;
- form SP-RG - income statement of simplified accounting and flat rate schemes.

d.3) The Modello Unico for corporate enterprises (SC)

All corporate enterprises resident in Italy are obliged to fill in this form. For this work, we consider:

- form SC-RF – income statement of ordinary accounting;
- form SC-RS - various statements[14].

d.4) The Irap form

The Irap form is used to declare the regional tax on productive activities carried out by enterprises. It is filled regardless of the accounting system adopted and is composed of several sub-forms (IQ, IP and IC) in accordance with the different type of the enterprises. This source has been used in the frame process only since the year of reference 2011.

e) The Istat Oros survey

The Istat short term survey named Oros, which stands for Occupazione (Employment), Retribuzioni (Wages), Oneri Sociali (Other labour cost), aims to produce quarterly information on the evolution (and levels) of gross wage, other labour cost and employment. The Oros survey uses administrative data coming from the Italian Social Security Institute (Inps) for estimating employment, wages and other Personnel costs. The Oros survey is used in this context as a further source to complete the coverage analysis for retrieving information on the cost of labour.

Before moving on coverage and comparisons analyses, it is interesting to examine the Table 2 which contains the information from the SME survey respondents by size classes of persons employed regarding the profit and loss account to which they refer for filling in the questionnaire of the survey[15].

Table 2 – Number of respondent enterprises in the SME survey by size classes and profit and loss account used for SME questionnaire. Year 2010

SIZE CLASSES	Total number of respondents	Chart of accounts	Financial statement (IV directive)	Financial statement (IAS)	Modello Unico	Sector studies	Other	Total
0 - 9	26,348	8,342	9,456	799	8,635	3,415	3,836	34,483
10 - 19	5,636	2,204	3,992	242	698	744	1,145	9,025
20 - 49	3,903	1,562	3,143	175	367	342	974	6,563
50 - 99	1,769	777	1,503	95	162	51	487	3,075
Total	**37,656**	**12,885**	**18,094**	**1,311**	**9,862**	**4,552**	**6,442**	**53,146**

Most of enterprises uses financial statement according to the IV Directive or the chart of

[14] At first this source was not available, then the tables in the appendix do not report numbers related to it.

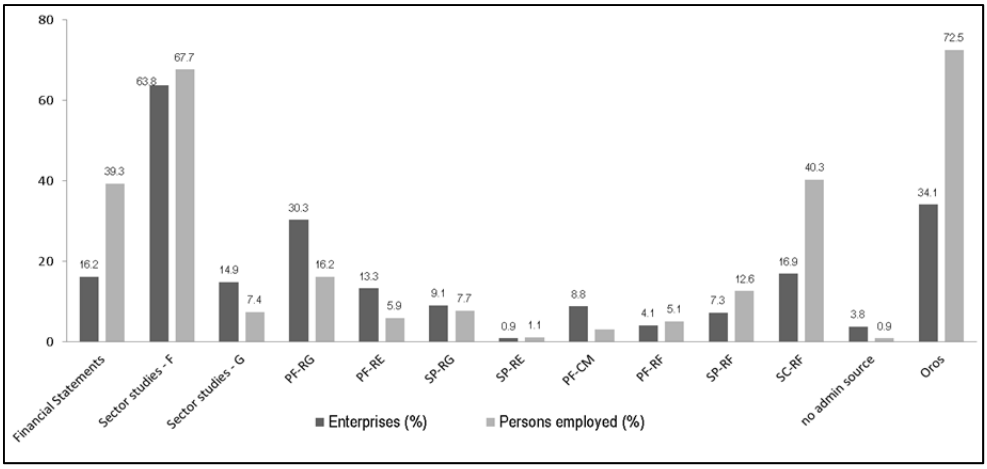
[15] In the questionnaire are allowed multiple responses.

accounts but there are many enterprises that use of Modello Unico (9,862) or Sector Studies (4,552). In the size class 0-9 financial statements in IV Directive, the chart of accounts and fiscal form Modello Unico are used in an almost equivalent manner.

3.3 Coverage analysis of the SMEs population

The administrative data of the year 2010 have been preliminarily linked to Asia2010, through the fiscal code, thus identifying the number of enterprises active in the year 2010 for each source. A summary is shown in Figure 1, where if an enterprise is present in more than one source it is counted for each of them.

Figure 1: Percentages of Asia enterprises and persons employed out of the SMEs reference population by single administrative source. Year 2010.



As we read in Figure 1, enterprises with financial statements represent 16.2% of the population and 39.3% in terms of employment. These percentage are very different by size classes: they increase according to the size and in the class 50-99 reach percentages of about 87,5% in terms of enterprises and employment.

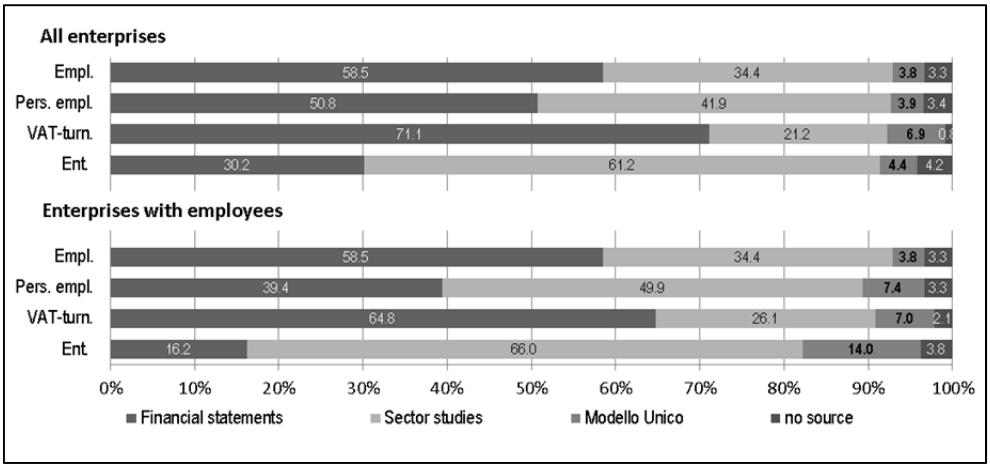
Enterprises that fill in the form F of the Sector studies are more than half and account for more than two-thirds of the employment. The best represented class is that one with 06-09 persons employed, where coverage of enterprises and employment are close to 83%, while the worst represented is the class 50-99 (22-23% of enterprises and employment), mainly due to the turnover threshold.

The form G of the Sector studies represents 14.9% of the enterprises and 7.4% of the employment, percentages in line with the form PF-RE of the Modello Unico.

Very high percentages are covered by the source Oros: 34.1% of the total number of enterprises and 98.8% of those ones with employees[16].

[16] See Annex A, Table A.1.for more details.

Figure 2: Percentage of Asia enterprises and persons employed out of the SMEs reference population by overlapped administrative source. Year 2010.



The sources that covers most of the reference population are the financial statements and the Sector studies. A further significant contribution is given by the form PF-CM of the Modello Unico (8.8% of enterprises) and by the form SC-RF (4.5% of VAT-Turnover)[17].

The lack of covering is 3.8% in terms of enterprises[18], 2.1% for VAT-turnover[19] and 3.3% for persons employed. These percentages are more or less the same if calculated for enterprises with employees.

For SBS domains no division, group or class of economic activity by size class and no division by region are uncovered[20]. Poorly covered domains[21] are small or very small ones, and. Also uncovered non-SBS domains are few. In many of these cases data are retrievable from Oros[22].

[17] See Annex A, Table A.2.for more details.

[18] As mentioned before, the Irap form has been included later in the analysis.

[19] It is worth mentioning that the volume of business, valid for the purposes of VAT, is only a proxy of the turnover, as there are some economic activities exempt from that statement.

[20] Except for a very few cells, in which there is in Asia only one enterprise whose data we cannot retrieve from any source. Similar considerations can be made for enterprises with employees, whose uncovered domains are always represented by 1-2 enterprises in the reference population.

[21] With less than 50% of enterprises of the reference domain.

[22] Uncovered domains involve 131 enterprises, and enterprises in poorly covered domains concern mostly specific legal forms, such as public-economic entities, special entities and public services companies, and fiscal representative entities.

4. Harmonization of variables and indicators of comparison

Administrative data has been processed and harmonised in order to replicate the income statement as faithfully as defined in SBS and the SEC regulations. The proposed scheme in Table 3 shows the main variables estimated from administrative sources and then a consistency analysis has been done with the data matched with SMEs survey and the relative theoretical coverage.

Table 3 – Variables of the profit and loss account re-classified according the value of production

VARIABLES OF THE PROFITS AND LOSSES ACCOUNT	SMEs Survey	Financial statements	Sector studies	Tax returns data	Coverage
Value of production at costs of factors (PRCF)	**PRCF (a)**	**X**	**X***	**X***	**ABC***
+ Turnover	11100	x	*x*	x	ABC
+ Changes in inventories of finished goods, work in progress and semi-finished	11200	x	*x**	x*	ABC*
+ Changes in work in progress	11300	x	x	x	ABC
+ Increase in fixed assets for internal work	11400	x	x		AB
+ Other income (non-financial, non-overtime)	11500	x	x	x	ABC
- *Purchases of goods for resale*	*12103*				*S*
- *Change in stockss of goods for resale (opening balance - final amount)*	*12602*				*S*
Intermediate costs (COSTI_SBS)	**COSTI_SBS (b)**	**X**	**X***	**X***	**ABC***
+ Purchases of goods and services	12100+12200	x	x	x	ABC
Purchases - Total	12100	x	x	x	ABC
Services - Total	12200	x	x	x	ABC
+ Charges for the use of third party assets: total	12300	x	x	x	ABC
+ Changes in inventories of raw materials and goods - total	12600	x	*x**	x*	ABC*
+ Other operating expenses - total	12900	x	x	x	ABC
- *Purchases of goods for resale*	*12103*				*S*
- *Change in stockss of goods for resale (opening balance - final amount)*	*12602*				*S*
Value Added (VA_SBS= PRCF - COSTI_SBS)	**VACF**	**X**	**X**	**X***	**ABC**
Labour cost (CL)	**44000**	**X**	**X**	**X***	**ABCD**
workers independent	indip				D
workers employees	dip				D
hours worked	orelav				D
Gross Operating Surplus (MOL_SBS=VA_SBS-CL_SBS)	**MOL_SBS**	**X**	**X**	**X**	**ABC**
Cost of structure (CS_SBS)	12500+12700	X	X	X	ABC
Net Operating Surplus (MON_SBS=MOL_SBS-CS_SBS)	**MON_SBS**	**X**	**X**	**X**	**ABC**

* The variable cannot be calculated directly from the administrative data, but can be achieved through a process of estimation.

Sources and coverage:
A → Source: Financial statements (700k/4500k) ~ 16%
B → Source: Sector studies (3000k/4500k) ~ 67%
AB → Financial statements + Sector studies (3400k/4500k) ~ 77%
ABC → Financial statements + Sector studies + Tax return ~ 95%
D → Source: Social Security Data (all enterprises with at least 1 employee ~ 1.300k)
S → Source: Small Medium Enterprises Survey

The calculation of the target variables show, however, problems related to the correct definition of the Value of production, Intermediate costs (which are not subtracted from revenues and cost of goods for resale) and Personnel costs of atypical contractual forms (collaboration etc..) that should be excluded and included in the intermediate costs. The value added is then calculated as the difference between proxies of production value and intermediate costs. The variables turnover, purchases of goods and services, value added, labour cost, gross operating surplus are therefore those directly obtainable from administrative sources.

To validate the information contained in administrative data and to evaluate the consistency of the harmonization process of variable definitions for the main SBS a comparison analysis has been made.

For that purpose, the distributions of the variables in SME survey are compared with the corresponding variable reconstructed from administrative data, calculating the percentage differences compared with the survey estimates.

For each variable, the quality indicators used for comparison are the following:

1. the Kolmogorov-Smirnov index (KS) (similarity of two distributions);
2. the proportion of firms in the range of difference (± 0%, ± 2%, ± 5%);
3. the proportion of value observed in the range of difference (±0%,±2%, ±5%);
4) the average % difference;
5) the average value difference;
6) the median value difference;
7) the interquartile range;
8) the coefficient of variation.

5. Comparisons between different sources

The aim of this paragraph is to evaluate the consistency of the variables that can be derived from the administrative sources described above with those derived from the SME survey for the reference year 2010[23] through the computation of quality indicators.

5.1 Comparison between Financial statements and SME survey

For this analysis the information from 37,920 respondent units to the survey and 742,359 units from Financial statements of corporate enterprises have been linked obtaining 16,602 units useful for the analysis. Therefore, it has been possible to compare the distributions of the variables from SME survey with the corresponding ones reconstructed from the income statement, thus calculating the percentage differences (Table 4).

The comparison between the SME survey data with income statements shows a good fit for many variables. Out of 17 variables comparable, more than half (13) show a distribution more or less similar: this is linked to the characteristics of the survey, which is based on a questionnaire in accordance with the IV accounting directive and to the fact that the editing and imputation phase use the balance sheets to detect and correct the outliers. It should

[23] Only for the Irap form, the reference year is 2011.

however be emphasized that the data of about 20% of the respondent enterprises is subject to a further investigation while 80% of firms are not affected by the process of editing and imputation.

In details, concentrating the analysis on the variables linked to the value of production, the comparison shows a good fit for "Revenue from sales and services", "Change in stocks for work in progress", " Work performed by entity and capitalised "; while the comparisons is definitely not good for the variables "Change in stocks of finished goods" and "Other operating income". Regarding the latter variable, as it can be seen from the table, despite the test is not significant, more than 53% of enterprises have the same value in both sources that represent about 51% of the total in value. However, despite it is a variable that in according to the accounting principles include extraordinary items, such as capital gains, not requested by the SME questionnaire, it is residual and does not affect significantly the estimation of the value added.

Table 4 - Main indicators of the comparison between the SME survey and income statements variables (16,602 enterprises)

VARIABLES	KS*	Median diff.	First quartile	Third quartile	*% units*			*% value*		
					±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	0.70	0.0	0.0	0.0	70.5	90.5	93.2	72.3	89.4	91.9
Change in stocks of finished goods (a)	5.75	0.0	0.0	0.0	82.6	84.2	84.5	-	-	-
Change in contract work in progress	0.57	0.0	0.0	0.0	95.2	95.4	95.4	-	-	-
Work performed by entity and capitalised	0.10	0.0	0.0	0.0	98.4	98.6	98.6	78.3	84.6	84.7
Other operating income	7.66	0.0	-0.9	0.0	53.8	58.7	60.4	50.9	61.5	64.2
Purchases of goods	0.95	0.0	0.0	0.0	65,5	79,7	84,6	64,4	84,3	87,9
Purchases of services	0.99	0.0	0.0	0.0	47,3	63,7	70,5	50,4	76,1	82,9
Purchases of goods and services	0.25	0.0	0.0	0.0	47,6	74,8	82,8	51,0	85,6	89,8
Use of third party assets	2.10	0.0	0.0	0.0	74,9	85,0	87,3	73,2	88,3	91,2
Personnel costs	1.67	0.0	0.0	0.0	65,8	82,6	87,3	60,5	82,5	88,6
Wages and salaries	0.29	0.0	-0,1	0.0	46,2	68,0	79,2	34,6	65,3	78,7
Depreciation and amortization	1.49	0.0	0.0	2,2	60,7	71,7	77,8	36,9	57,1	74,2
Change in stocks of raw materials and consumables (b)	3.75	0.0	0.0	0.0	82,1	85,1	85,3	-	-	-
Change in stocks (a-b)	0.66	0.0	0.0	0.0	79,1	83,7	84,1	-	-	-
Other operating expenses	4.34	0.0	0.0	0.0	54,9	61,8	63,6	44,4	51,0	53,2
Value added	1.04	0.0	-1,1	0,3	33,0	58,7	68,0	28,2	63,9	74,4
Gross operating surplus	1.85	0.0	-5,2	0.0	34,2	52,7	58,6	31,5	62,5	69,5

* Threshold Value 1.6

In terms of costs of production, the only variables that are not significant to test for equal distributions are "Change in stocks of raw materials and consumables" and "Other operating expenses". The KS indicator is slightly negative for the variable "Personnel costs". The variable "Other operating expenses" shows an empirical distribution similar to the corresponding distribution in SME survey. However, even in this case, it could be considered as a residual part of the value added. It represents only 1,7% of the negative components of the value added. Regarding the change in stocks and analysing the single

components there is not a good fit. That depends on errors in filling in the questionnaire for the difficulty of some enterprises to distinguish change in stocks on the side of production from the side of costs. By considering, instead, the variable as a whole "Change in stocks (a-b)", the fit improves so as to ensure that the test is significant.

5.2 Comparison between the Sector studies and SME survey

In the Survey we have 37,920 respondent units, of which 14,363 have been matched with Sector studies data. The main results of the comparison analysis on the subset of individuals (sole proprietorships and family businesses, self-employed persons or partnerships) show a positive mean difference of revenues (+0.4%) and also on purchases of goods and services (+0.5%), with a percentage difference for the value added of -1.6%. More detailed analyses at a 3-digit of Nace and company size level of disaggregation do not highlight systematic discrepancies. The highest differences have been reported for the cost items and for units involved in service sectors. It is due to different classifications of costs among different sources, in particular for the item other operating expenses. Other discrepancies have been caused by the mismatch between the economic activity in Asia and in the administrative data (due to the different criteria for identifying the main economic activity). (Table 5).

Table 5. - Main indicators of the comparison between the SME survey and Sector studies (form F) variables (14,363 enterprises)

MAIN ECONOMIC VARIABLES	KS*	± 5% (% units)	± 5% (% value)	%diff.	value diff. (000€)	median diff. (000€)	IQR diff. (000€)	CV diff.
Revenues from sales and services	1.0	90.6	94.0	0.4	1.5	0.0	0.0	89.6
Work performed by entity and capitalised	0.4	99.3	47.0	25.5	0.1	0.0	0.0	339.4
Change in contract work in progress	0.5	96.6	522.9	-936.1	0.9	0.0	0.0	88.2
Change in stocks (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 operating income	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
Use of third party assets	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
Personnel costs	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
Gross operating surplus	1.4	45.6	38.2	-4.7	-2.9	0.1	4.5	-29.7
Net operating surplus	2.1	42.8	36.4	-3.1	-1.5	0.0	5.1	-61.1

* Threshold Value 1.6

5.3 Comparison between "Modello Unico" and SME survey

From the fiscal source minimum taxpayers (Modello Unico PF with CM form), sole proprietorships (Modello Unico PF with RE and RG forms) and partnerships (Modello Unico SP with RE and RG forms) have been considered. Tax return data has been aggregated according to a criterion of proximity to the definitions of the SBS Regulation.

The analysis are conducted at the micro level by comparing distributions and calculating the KS indicator. The indicator measures the similarity of distributions for each couple of variables: the closer it gets to zero the more likely is the hypothesis of similarity. The comparison analysis is carried out without a preliminary editing and imputation process. Some results, then, could be influenced by the presence of outliers.

For the purposes of the analysis, only those enterprises responding to the SME survey (37,656) have been selected, while in the case of the tax data enterprises that fill in the form of the CM minimum taxpayers (393,048 units) or the Modello Unico PF, for the forms RE "income from self-employment" (592,131 units) and RG "determining incomes of enterprise in simplified accounting and flat rate" (1,347,005 units), or the Modello Unico SP (partnerships), for the forms RE (40,229 units) and RG (413,963 units), have been taken into consideration.

Then, it is carried out the linkage by fiscal code present in Asia in order to extract information on economic activity, geographical location and the number of persons employed on which to calculate size classes of persons employed (0-1, 2-3, 4-5, 6-9, 10-19, 20-49 and 50-99). Subsequently, SME survey is linked with the above mentioned sources, obtaining the following results in terms of number of enterprises:

- panel SME - Modello Unico PF (CM): 1,406 enterprises;
- panel SME - Modello Unico PF (RE): 1,835 enterprises;
- panel SME - Modello Unico PF (RG): 6,025 enterprises;
- panel SME - Modello Unico SP (RE): 410 enterprises;
- panel SME - Modello Unico SP (RG): 2,780 enterprises.

The comparison between SME survey and minimum taxpayers data shows a good fit between the "Total of positive components" (CM) and "Revenues from sales and services" (SME) and also the consistency of the variable "Value added" in the two sources is quite satisfactory. For more details, see Annex B, Table B.1.

The comparison between SME survey and tax return data on sole proprietorships (Modello Unico PF), highlights a good fit for "Revenues from sales and services" while for the other items do not seem to be a satisfactory consistency; however, it does not discharge completely on the "Value added". In fact, the value added, according to the KS index, is sufficiently consistent in the two sources, despite the possible presence of costs for project and temporary workers in personnel costs rather than in service costs. That is valid, if considering the tax definitions, but it might not be true that firms matched have actually made use of external staff. For more details, see Annex B, Table B.2 e B.3.

The comparison between SME survey and tax data from Modello Unico SP, RE, and RG highlights the great fit for "Revenues from sales and services" while for the other items, especially for costs, consistency is not always satisfactory, but the "Value added" is sufficiently consistent in the two sources. For more details, see Annex B, Table B.4 e B.5.

The comparison between survey data with Tax return data shows, then, a good level of comparison for the revenue variables and the value added and less for expenses variables, because in the administrative forms a different classification of costs is required. However,

the contribution of such information from Tax return data have to be considered a good proxy of income statements data, that are not available for individuals and partnerships.

In the assessment of administrative sources on the basis of the contribution of information and its consistency with the statistical definitions, a residual role has been assigned to Modello Unico. That finds justification in the fact that tax data contributes to the overall estimate of turnover to about 0.4% (minimum taxpayers CM form) to 0.1% (sole proprietorships RE form) to 0.3% (sole proprietorships RG form) to 0.1% (partnerships RE form) and 0.2% (partnerships RG form). In terms of value added the contributes are 1,2% (minimum taxpayers CM form), 0.3% (sole proprietorships RE form), 0.5% (sole proprietorships RG form), 0.2% (partnerships RE form) and 0.2% (partnerships RG form).

5.3.1 Comparison between Irap form and SME survey

The source Irap has been introduced in the process of frame from the year of reference 2011, so in this case the comparisons between the main variables available from the different sources have been based on the data of the reference year 2011.

Table 6 - Main indicators of the comparison between Irap and SME survey variables (14,461 enterprises)

VARIABLES	KS*	Median diff.	*% units*			*% value*		
			±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	1.64	0.0	65.2	88.4	91.6	73.0	91.6	93.4
Change in stocks of finished goods (a)	40.48	0.0	81.0	82.6	82.7	-	-	-
Changes in contract work in progress	28.39	0.0	94.5	94.6	94.6	-	-	-
Work performed by entity and capitalised	24.66	0.0	98.2	98.3	98.4	76.6	78.1	78.3
Other operating income	19.49	0.0	56.2	62.9	64.3	57.7	72.6	75.0
Purchases of goods	10.13	0.0	61.8	78.6	83.4	67.6	89.5	91.7
Purchases of services	0.63	0.0	38.8	58.4	65.7	44.1	69.6	77.7
Purchases of goods and services	0.43	0.0	38.0	68.6	77.6	51.3	86.3	91.3
Use of third party assets	14.95	0.0	70.3	81.6	84.1	70.8	86.9	89.6
Depreciation and amortization	4.27	0.0	76.6	89.4	91.0	77.6	91.6	93.7
Change stock of raw materials and consumables (b)	33.00	0.0	78.8	81.7	81.9	-	-	-
Other operating expenses	4.88	0.0	45.3	53.9	56.2	43.7	52.2	54.3
Change in stocks (a-b)	28.49	0.0	78.2	82.9	83.3	-	-	-
Value added	0.57	0.0	29.4	56.4	65.9	31.1	63.0	73.2

* Threshold Value 1.6

The tax is calculated as the difference between the amount of revenues and the amount of the costs. Some costs are not deductible and among these, personnel costs, costs for other workers assimilated to employees (project workers), taxes on the buildings are the most significant ones. The value added calculated from this source should be higher than that one from the statistical and the other administrative sources (SME, income statements, Sector studies and so on). In particular, the Irap source should estimate a lower cost for services (due to the presence of external workers in the other sources) and lower other operating

costs (due to the presence of taxes on buildings in the other sources).

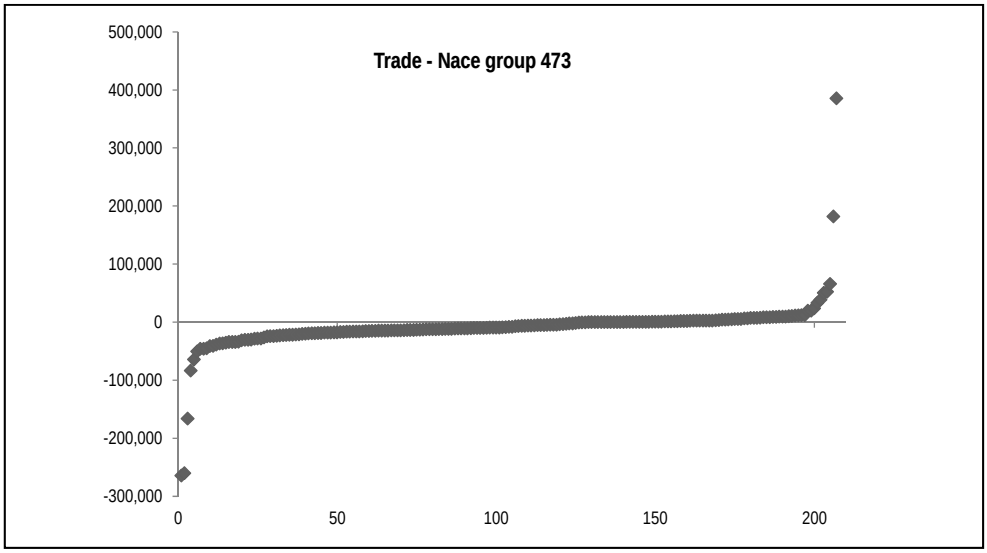
The analysis regards the comparison between the Irap data for the corporate companies (form IC section I) and SME survey data. The comparison involves 14,461 units in both sources and shows a good fit for most variables, except for those ones that include non-deductible costs for Irap such as services costs and other operating costs (Table 6). Several enterprises show the same value in both the sources, although the results improve significantly considering the ranges of variation ± 2%. With regard to the value added in the comparison is better in the range ±5% that includes 65.9% of enterprises, representing 73.2% of the value.

6. Analysis of critical domains

When domains showed relevant and significant differences a deeper analysis has been made. This analysis has been carried out at a group level of Nace by size class and confirms the absence of any systematic discrepancies. The highest differences have been found for the cost variables and relate to the different classification of costs between the different sources examined, in particular for the item other operating costs. Other discrepancies are related to the mismatch between the main economic activity from Asia and the economic activity from the Sector studies source (due to the different criteria for identifying the main economic activity).

As regards the problem of the domains not covered, that is, with a high presence of foreign enterprises, the solution is to exclude from Frame (and Asia) the legal form "1900-special purpose entities without a stable plant" (with less than 3 employees), but only for those with more than 3 employees will need to recover the financial statement data via other sources: Telemaco, Irap Fiscal data.

Figure 3 – Absolute differences of the value added in Sector studies and survey – Nace group 473 year 2010 (euro)



The most critical economic activity of the Trade sector (Figure 3) do not show a clear sign of systematic differences. In all cases, the gains or losses depend on a few influential observations that mark in a positive or negative difference in the final evaluation. The test KS for comparing the distributions do not report a significant difference between the distributions of the Value added observed in the two sources. Moreover, the share of the observed value for the companies belonging to the range difference of ± 10% is always greater than 50%.

There is a different classification of revenues for some companies, particularly those that make buying and selling of real estate, so the differences with SMEs relate to the different classification of revenues from sales in other revenues. For firms in the Nace sector 681 (Buying and selling real estate) there are 4 outliers that affect the negative difference. For these companies, revenues should be adjusted for the purchase of buildings to be sold and change in inventories. This information should be retrieved from the survey (the same way as is done for trade firms for goods to resale) or specific information about sales by kind of activity by using additional fiscal data (i.e. form D of the Sector studies).

The transport sector (Nace 494) has a prevalent tail of negative differences (Sector studies<SMEs) and the difference in value added is equal to -8.1% compared to the SMEs (the KS indicators is not significant). There seems to be a problem of non-deductibility of costs from the side of the Sector studies. The cause of the discrepancy concerns the indication of higher operating expenses in the Sector studies (+ 41%), while for the other revenues (-3.5%) prevails SMEs.

The service sectors with the highest differences are social services, but the low number of observation at the group level (Nace 3 digits) do not allow to obtain robust indicators for the comparisons. Indicators are affected by the presence of higher values in revenues in the Sector studies compared with SMEs.

For free-lances and professional activities there is a problem relating to the application of the accounting cash method compared to the general accounting accrual method to determine the financial year. This issue makes it difficult to compare the differences between Sector studies and SMEs it depends on different interpretations of the survey questionnaires which does not present an element of systematic errors, but may generate measurement errors.

In manufacturing sectors there are no systematic differences for the following variables: Revenue, Purchases of goods and services and Value added. In some domain there are differences between SMEs and Sector studies due to anomalous cases that affected the indicators used for comparison. In all cases, however, is predominant the share of enterprises with percentage differences of ± 10%.

In the construction activities differences in Value added do not show a regularity neither by number of persons employed or by Nace class. There might be a different interpretation of the classification of items in the value of production (income and work in progress) between SMEs and Sector studies, but this issue do not significantly affect the comparison of revenues.

7. Decomposition of differences in the final estimations

At the end of the quality analysis, some variables have been directly calculated from the administrative sources and assumed as auxiliary variables X, for which differences among sources are not imputed to a different definition. They are the following:

- Turnover;
- Proxy of Value of production (including purchase of goods for resale, changes in raw materials and goods for resale);
- Proxy of Intermediate costs (including purchase of goods for resale and excluding changes in raw materials);
- Purchases of goods and services;
- Value added (difference between the proxy of the Value of production and the proxy of Intermediate costs);
- Personnel costs;
- Gross Operating Surplus (GOS).

For these variables the following five quality criteria have been satisfied:

1. total difference below 5%;
2. KS (Kolmogorov-Smirnoff indicator) not significant ;
3. Share of enterprises and the value of the variable at least 70% for the percentage difference within 5%;
4. Good consistency and no significant distortion of these indicators at the level of NACE 2-digit;
5. Good consistency and no significant distortion of these indicators at the level of the class of employees.

The other variables from administrative sources have to be considered as a proxy X^* to be used for the estimation of the Y variables coherent with SBS and SEC regulations at domain level.

The model $Y = f(X^*)$ needed survey data to correct discrepancies in administrative data due to sectorial regulations, no perfect alignment of the definitions and linkage errors. In these case the use of auxiliary information will improve the predictive power of estimators and will open the way to two alternative research developments: mass imputation (Chipperfield et al. 2012) or aggregate combined estimators (Pfeffermann, 2003).

The process of integration of the multi-source database and the reconstruction for all the population of the main economic variables allowed us to evaluate two component in the differences in survey estimates compared with administrative-based estimates: (1) the first one is the replacement (source effect) of the survey with administrative data and (2) the second one is the sample weights (sample effect) compared to census data. The exercise has been made on the variable value added and is coherent with the previous analysis conducted within the integration of the sample of SMEs (Casciano et al., 2012).

The final dataset used in this evaluation integrated missing records or incorrect data (approximately 6%) with a cold deck imputation procedure of the per capita value of the median by stratum (Ateco three-digit group, size class and region). The imputed value is then equal to per capita median of the stratum multiplied the number of employees of the unit with missing data. The resulting values preserved the distributions of each variable in the strata. The coherence at unit level is guaranteed by the fact that the value added is recalculated after the single imputation of the value of production and of the intermediate costs.

The estimate of the Value Added based on Administrative data is the following:

$$Y_{admin} = \sum_{i=1}^{N} y_i^a$$

The estimate of the Value Added based on Survey data is the following:

$$Y_{sme} = \sum_{i=1}^{n} y_i w_i$$

The difference is decomposed in two parts, adding and subtracting the survey estimate based on administrative data ($\sum_{i=1}^{n} y_i^a w_i$):

$$Y_{admin} - Y_{sme} = \left(\sum_{i=1}^{N} y_i^a - \sum_{i=1}^{n} y_i^a w_i\right) + \left(\sum_{i=1}^{n} y_i^a w_i - \sum_{i=1}^{n} y_i w_i\right)$$

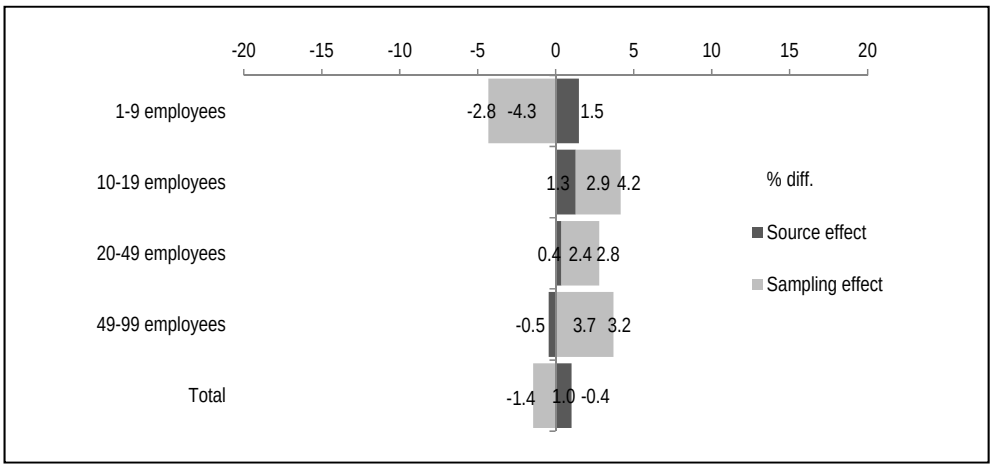
The first part represents the sampling effect in which the weights w_i are substituted with all the observations of the Business Register with complete information for the estimated variable (the missing values, about 5%, are imputed through a median imputation by stratum):

$$sampling\ effect = \sum_{i=1}^{N} y_i^a - \sum_{i=1}^{n} y_i^a w_i$$

The second part represented the source effect (with the same weight w_i):

$$source\ effect = \sum_{i=1}^{n} y_i^a w_i - \sum_{i=1}^{n} y_i w_i$$

Figure 4. Value added estimate in survey (Y_{sme}), in administrative data (Y_{admin}) and decomposition of the total difference by size, Year 2010



The decomposition of the differences between census estimation from administrative data (Y_{admin}) and survey calibrated estimation (Y_{sme}), for the variable Value added (Figure 4), showed a prevalence of the so called "sampling effect" (sum of all observations with respect to the weighted sum), related to sampling error, on the "source effect" related to measurement error . The first is equal to -1.4 percentage points and the second to +1.0 percentage points, and both contribute to a difference between the estimated administrative and survey data from SMEs -0.4%. In fact the effect of sample weights have a negative impact (-4.3 percentage points) on the class 1-9 employees and is positive for the upper classes. The source effect is almost always positive: administrative-based estimate is higher than 1% and decreases with increasing company size. It went from +1.5% for micro enterprises to +0.4% for medium enterprises with 20-49 employees and became negative for those over 49 employees. At the sectorial level there is a prevalence of the sample effect in service activities and this effect is almost always opposite to the effect of replacing the data from administrative sources and is greater in sectors with a high concentration of micro-enterprises.

Conclusions

The quality analysis has regarded the following aspects of the integration process: (1) the coverage of the population of reference; (2) the harmonization of variables of the income statement from multiple sources; (3) comparison indicators and distribution of differences; (4) decomposition of differences.

The coverage analysis was carried out respecting the importance of different sources. Not all sources have the same informative contents and not all enterprises have the same organization. Financial Statements had a good fit because it covers companies with an ordinary accounting system and this is demonstrated by the comparison indicators with survey data: the distributions of all main variables were similar. They cover about the 16% of the reference population but they represent more than half of its value added.

The sole proprietorships and unincorporated enterprises are covered almost all with Sector Studies (about 66%). In this case the harmonization process and the reclassification of the income statement had permitted to estimate the main variables including the value added. Enterprises with a simplified account system, that play a minor role (about 14%), were estimated through Fiscal data. In this case the harmonization process has been more burdensome and the comparison with survey data was less satisfactory.

In general the analysis of the comparisons between administrative data and survey data reveals a good fit for the most important variables and the presence of errors on the side of the survey that affect quality indicators. For the main variables has been confirmed a random distribution of the differences, moreover the analysis of critical domains confirms the absence of systematic errors.

The decomposition of the differences showed the prevalence of the sampling effect on the source effect. The first is equal to -1.4 percentage points and the second to +1.0 percentage points, and both contribute to a total difference of -0.4%. This analysis confirm the good fit of administrative data to SBS requirements, with a small measurement error as confirmed by the source effect. So the source effect represents a minor part respect to the sampling error that derived only from the SME survey, as a result of the misrepresentation

of the final respondents (as a result of MNAR in the initial sample[24]) compared to the reference population.

[24] Missing Not at Random (MNAR): missing observations related to a specific subset of the population. It has been verified that the estimated variable depend on the response rate (F. Oropallo 2011).

References

Casciano M.C., A. Cirianni, V. De Giorgi, T. Di Francescantonio, A. Mazzilli, O. Luzi, F. Oropallo, M. Rinaldi, E. Santi, G. Seri, G. Siesto. 2011. *Utilizzo delle fonti amministrative nella rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni*. Istat Working Papers n.7/2011.

Casciano M.C., V. De Giorgi, F. Oropallo, G. Siesto. 2012. "Estimation of Structural Business Statistics for Small Firms by Using Administrative Data*". Rivista di statistica ufficiale* n.2-3/2012.

Chipperfield J., J. Chessman, R. Lim. 2012. "Combining Household Surveys Using Mass Imputation to Estimate Population Totals". *Australian & New Zealand Journal of Statistics*. Vol. 54, Issue 2, June 2012: 223-238.

Commission Regulation (EC) No 97/2009 of 2 February 2009 implementing Regulation (EC) No 295/2008 of the European Parliament and of the Council concerning structural business statistics, as regards the use of the flexible module.

Commission Regulation (EC) No 250/2009 of 11 March 2009 implementing Regulation (EC) No 295/2008 of the European Parliament and of the Council as regards the definitions of characteristics, the technical format for the transmission of data, the double reporting requirements for Nace Rev.1.1 and Nace Rev.2 and derogations for structural business statistics.

Commission Regulation (EC) No 251/2009 of 11 March 2009 implementing and amending Regulation (EC) No 295/2008 of the European Parliament and of the Council as regards the series of data to be produced for structural business statistics and the adaptations necessary after the revision of the statistical classification of products by activity (CPA).

Commission Regulation (EU) No 275/2010 of 30 March 2010 implementing Regulation (EC) No 295/2008 of the European Parliament and of the Council, as regards the criteria for the evaluation of the quality of structural business statistics.

Commission Regulation (EU) No 549/2013 of 21 May 2013 on the European system of national and regional accounts in the European Union.

Istat. 2012. *Struttura e competitività del sistema delle imprese industriali e dei servizi.* Statistica Report del 29 ottobre 2012.

Istat. 2014. *I nuovi conti nazionali in Sec 2010.* Statistica Report del 6 ottobre 2014.

Istat. 2015. *Economia non osservata nei conti nazionali.* Statistica Report del 4 dicembre 2015.

Luzi O., G. Seri, V. De Giorgi, G. Siesto. 2013. "*Estimating Business Statistics by integrating administrative and survey data: an experimental study on small and medium enterprises". Rivista di statistica ufficiale*, n. 2-3/2013.

OECD. 2002. *"Measuring the Non-Observed Economy - A Handbook"* (OECD, 2002).

Oropallo F. 2011. *Analisi delle differenze strutturali nella performance economica tra unità rispondenti e unità non rispondenti nella rilevazione dei risultati economici delle piccole e medie imprese (PMI) – Contributi Istat* – n. 7/2011.

Pfeffermann D. 2013. "New important Developments in Small Area Estimation". *Statistical Science of Institute of Mathematical Statistics.* Vol. 28, No. 1: 40-68.

Regulation (EC) No 295/2008 of the European Parliament and of the Council of 11 March 2008 concerning structural business statistics (recast).

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

Annex A

Table A.1 - Number of Asia enterprises and persons employed of the SMEs reference population by single administrative source. Year 2010.

SOURCE	Enterprises		Persons employed (no.)	Enterprises (%)	Persons employed (%)
	Total (no.)	*with employees (no.)*			
Financial Statements	718,538	*462,650*	4,940,295	16.2	39.3
SS - F	2,834,584	*1,172,856*	8,513,828	63.8	67.7
SS - G	661,340	*116,483*	930,808	14.9	7.4
PF-RG	1,347,005	*331,649*	2,043,266	30.3	16.2
PF-RE	592,131	*88,097*	742,431	13.3	5.9
SP-RG	402,471	*154,826*	965,249	9.1	7.7
SP-RE	39,367	*18,416*	143,231	0.9	1.1
SC-RS	793,130	*472,712*	n.d.	17.8	n.d.
PF-CM	393,048	*1,113*	394,580	8.8	3.1
PF-RF	183,003	*106,972*	641,296	4.1	5.1
SP-RF	326,217	*205,838*	1,584,309	7.3	12.6
SC-RF	753,135	*482,232*	5,070,668	16.9	40.3
at least one admin source	4,277,197	*1,466,425*	12,463,874	96.2	99.1
no admin source	166,686	*64,685*	112,582	3.8	0.9
Oros	1,516,169	*1,512,825*	9,119,873	34.1	72.5
SME survey	37,656	*23,502*	412,466	0.8	3.3
SMEs population (from Asia)	**4,443,883**	***1,531,110***	**12,576,456**	**100.0**	**100.0**

Table A.2 - Number of Asia enterprise and persons employed of the SMEs reference population by overlapped administrative source. Year 2010.

SOURCES	*Enterprises (no.)*	*VAT Turnover (th.)*	*Persons employed (no.)*	*Employees (no.)*	*Enterprises (%)*	*VAT Turnover (%)*	*Persons employed (%)*	*Employees (%)*
All enterprises								
Financial statements	*718,538*	*1,130,303,208*	*4,940,295.21*	*4,102,499.97*	*16.2*	*64.8*	*39.4*	*58.5*
SS-F	2,275,422	398,015,809	5,353,790.07	2,194,976.91	51.2	22.9	42.6	31.3
SS-G	655,566	55,185,613	920,566.62	215,064.18	14.8	3.2	7.3	3.1
PF-RG	115,605	5,685,656	151,522.47	34,211.36	2.6	0.3	1.2	0.5
PF-RE	36,731	1,391,467	37,869.66	1,412.24	0.8	0.1	0.3	0.0
SP-RG	37,299	2,851,284	79,319.23	21,201.25	0.8	0.2	0.6	0.3
SP-RE	2,045	1,368,307	7,156.51	2,916.74	0.0	0.1	0.1	0.0
PF-CM	392,979	6,832,844	394,507.19	531.57	8.8	0.4	3.1	0.0
PF-RF	7,512	3,820,944	24,284.31	14,953.19	0.2	0.2	0.2	0.2
SP-RF	17,803	21,420,833	103,302.27	72,185.60	0.4	1.2	0.8	1.0
SC-RF	17,697	78,057,218	143,394.60	123,449.28	0.4	4.5	1.1	1.8
no source	166,686	36,478,938	420,447.52	233,630.14	3.8	2.1	3.3	3.3
TOTAL	**4,443,883**	**1,741,412,122**	**12,576,455.66**	**7,017,032.43**	**100.0**	**100.0**	**100.0**	**100.0**
Enterprises with employees								
Financial statements	462,650	1,042,144,197	4,660,389.09	4,102,499.97	30.2	71.1	50.8	58.5
SS-F	820,538	284,973,622	3,480,048.87	2,194,976.91	53.6	19.5	38.0	31.3
SS-G	114,762	25,147,913	353,972.75	215,064.18	7.5	1.7	3.9	3.1
PFRG	26,179	2,204,135	61,520.34	34,211.36	1.7	0.2	0.7	0.5
PFRE	1,171	143,560	2,608.83	1,412.24	0.1	0.0	0.0	0.0
SPRG	13,496	1,488,630	42,759.33	21,201.25	0.9	0.1	0.5	0.3
SPRE	722	1,215,400	4,452.95	2,916.74	0.0	0.1	0.0	0.0
PFCM	1,108	23,207	1,673.27	531.57	0.1	0.0	0.0	0.0
PFRF	3,798	3,226,327	19,968.05	14,953.19	0.2	0.2	0.2	0.2
SPRF	10,518	20,053,044	91,578.83	72,185.60	0.7	1.4	1.0	1.0
SCRF	11,483	71,413,642	136,706.16	123,449.28	0.7	4.9	1.5	1.8
no source	64,685	11,234,629	310,649.78	233,630.14	4.2	0.8	3.4	3.3
TOTALE	**1,531,110**	**1,463,268,306**	**9,166,328.25**	**7,017,032.43**	**100.0**	**100.0**	**100.0**	**100.0**

Annex B

Table B.1 - Main indicators of the comparison between SME survey and minimum taxpayers (CM) variables (1,406 enterprises)

VARIABLES	KS	Median difference	First quartile	Third quartile	*% units*			*% value*		
					±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	0.57	0.0	0.0	0.0	75.7	85.2	87.8	74.4	85.1	88.1
Value added at factor cost	0.45	0.0	-0.4	1.8	37.3	53.9	59.7	34.7	53.1	60.2

Table B.2 - Main indicators of the comparison between SME survey and sole proprietorships (Modello Unico PF, RE form) variables (1,835 enterprises)

VARIABLES	KS	Median difference	First quartile	Third quartile	*% units*			*% value*		
					±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	0.30	0.0	0.0	0.0	69.9	82.6	86.3	66.0	86.2	89.6
Other revenues and income	13.27	0.0	0.0	0.0	82.3	82.6	82.9	14.0	14.3	16.1
Purchases of goods	18,16	-26,7	-68,2	66,2	1.9	2.5	3.1	0.0	1.0	1.3
Purchases of services	4.97	1.2	-20.0	93.4	0.7	3.1	6.5	0.0	3.2	6.9
Purchases of goods and services	4.09	5.6	-5.0	64.4	7.8	16.8	22.9	2.1	28.3	36.0
Use of third party assets	16.66	0.0	0.0	0.0	62.2	64.6	66.2	26.8	41.7	48.9
Personnel costs	14.61	0.0	0.0	0.0	89.9	92.8	93.5	59.7	79.4	83.8
Value added at factor cost	0.38	0.0	-0.8	3.2	20.2	51.7	64.4	11.5	53.3	69.7
Gross operating surplus	0.50	0.0	-1-0	3.1	19.0	50.6	62.9	11.3	53.0	67.8

Table B.3 - Main indicators of the comparison between the SME survey and sole proprietorships (Modello Unico PF, RG form) variables (6,018 enterprises)

VARIABLES	KS	Median difference	First quartile	Third quartile	*% units*			*% value*		
					±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	0.66	0.0	0.0	0.0	75.6	89.8	91.9	71.0	91.2	93.3
Other revenues and income	18.51	0.0	0.0	0.0	80.0	80.3	80.5	6.9	8.1	9.5
Purchases of goods	9.16	0.0	-2.8	0.4	35.0	51.0	57.6	30.0	76.5	82.3
Purchases of services	12.60	49.3	1.3	176.5	4.2	7.6	11.8	2.8	7.8	14.1
Purchases of goods and services	5.0	10.6	0.9	43.7	4.8	18.3	29.0	1.8	41.6	58.1
Use of third party assets	20.8	0.0	-100	0	64.5	65.2	65.3	10.4	13.2	14.6
Personnel costs	31.2	0.0	0.0	0.0	68.7	73.0	75.1	45.0	64.8	73.8
Value added at factor cost	0.84	0.0	-6.0	1.4	14.7	42.7	56.1	9.1	44.1	58.9
Gross operating surplus	1.95	-0.2	-8.3	0.0	18.1	46.0	57.7	12.5	44.8	58.2

Table B.4 - Main indicators of the comparison between SME survey and partnerships (Modello Unico SP, RE form) variables (410 enterprises)

VARIABLES	KS	Median difference	First quartile	Third quartile	*% units*			*% value*		
					±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	0.17	0.0	0.0	0.0	66.1	85.6	90.2	58.6	93.8	95.2
Other revenues and income	6.95	0.0	0.0	0.0	10.3	79.7	80.5	10.3	79.7	80.5
Purchases of goods	6.84	-67.8	-94.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Purchases of services	2.73	29.6	-0.1	111.7	1.0	5.9	14.1	0.0	5.6	11.8
Purchases of goods and services	1.39	8.2	-1.9	42.7	2.9	14.1	27.3	0.2	11.4	20.6
Use of third party assets	3.92	0.0	0.0	11.2	41.0	47.6	51.5	32.0	39.4	42.5
Personnel costs	6.24	0.0	0.0	0.0	67.1	75.9	79.3	47.1	76.8	81.2
Value added at factor cost	0.42	0.2	-0.8	4.9	9.8	46.3	62.4	1.5	60.4	72.9
Gross operating surplus	0.45	0.1	-0.9	4.5	10.2	46.8	62.2	1.8	61.6	70.5

Table B.5 - Main indicators of the comparison between SME survey and partnerships (Modello Unico SP, RG form) variables (2,780 enterprises)

Variables	KS	Median difference	First quartile	Third quartile	*% units*			*% value*		
					±0%	±2%	±5%	±0%	±2%	±5%
Revenues from sales and services	1.24	0.0	0.0	0.0	68.6	85.7	87.7	63.3	90.7	92.7
Other revenues and income	14.53	0.0	-18.6	0.0	70.1	70.7	71.2	5.6	8.1	10.1
Purchases of goods	9.09	0.0	-3.0	0.2	36.7	52.4	58.5	29.9	79.2	85.3
Purchases of services	8.92	52.1	3.9	182.5	2.4	6.4	10.0	1.9	9.5	16.8
Purchases of goods and services	3.80	14.8	1.6	56.4	2.5	13.1	22.4	0.7	34.9	58.9
Use of third party assets	14.73	0.0	-100.0	0.0	53.3	53.8	54.1	6.5	7.4	8.5
Personnel costs	20.48	0.0	0.0	0.0	59.1	63.8	66.9	40.8	56.0	66.2
Value added at factor cost	0.75	-0.1	-8.5	3.1	8.9	34.3	48.2	5.9	34.5	49.9
Gross operating surplus	1.80	-0.98	-13.7	0.0	12.0	38.5	50.6	9.1	38.3	52.1

The labour cost variables in the building of the "Frame SBS"[1]

Stefania Arnaldi,[2] Ciro Baldi,[3] Rosalba Filippello,[4] Livia Mastrantonio,[5] Silvia Pacini,[6] Paolo Sassaroli,[7] Francesca Tartamella[8]

Abstract

The building of the new Structural Business Statistics (SBS) Register (so called Frame) exploits company accounts and enterprise level fiscal data as main sources concerning the variables on revenues and costs. When dealing with labour cost variables an additional source is available: social security data. There are advantages and disadvantages in the use of social security data to feed the variables on wages and labour costs into the frame. In fact, while being heterodox with the main sources, their use has the advantage to introduce consistency with employment data entering the Frame via the business register. This work reports the findings obtained trying to answer questions like: which is the most suitable source to comply with SBS definitions on wages and labour costs? What are the magnitudes and characteristics of discrepancies between sources? What are the causes of these differences? The main reasons of differences have been identified through an in-depth study of the definitions of administrative variables and comparative analyses between sources, at both the macro and micro level. The solution proposed for the Frame involves an innovative correction procedure using social security data to reduce the definitional bias in company accounts due to the inclusion of costs related to external workers into the labour cost.

Keywords: labour cost, multiple administrative sources, harmonization, integration

---

[1] This work is the results of the common effort of the authors. However the paragraphs can be attributed as follows: par. 1 C. Baldi, S. Pacini and F. Tartamella; par. 2 S. Pacini; par. 3 S. Arnaldi; par. 4 R. Filippello and P. Sassaroli; par. 5 S. Pacini; par 6.1 C. Baldi; par. 6.2 C. Baldi and S. Pacini; par. 7.1 L. Mastrantonio and F. Tartamella; par. 7.2 C. Baldi; par. 8 C. Baldi, S. Pacini and F. Tartamella. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat. We express our gratitude to Fabiana Rocci and Diego Bellisai, participants in the study group that analyzed the topic, who contributed to build the body of knowledge behind this paper.

[2] Istat, e-mail: arnaldi@istat.it
[3] Istat, e-mail: baldi@istat.it
[4] Istat, e-mail: filippel@istat.it
[5] Istat, e-mail: mastrant@istat.it
[6] Istat, e-mail: pacini@istat.it
[7] Istat, e-mail: sassarol@istat.it
[8] Istat, e-mail: tartamel@istat.it

1. Introduction

In the building of the new Structural Business Statistics (SBS) Register (hereafter Frame), the treatment of the labour costs variables has represented a new challenge. In fact, regarding these variables, in addition to the company accounts and the fiscal sources, social security data could be used to compile the Frame (on the input sources of the Frame see Curatolo et al, 2015). The data of social security declarations, containing information on wages and social contributions, while having the disadvantage of being heterodox with respect to the other sources used for the revenues and other labour costs and thus of carrying the risk of introducing inconsistencies with the rest of the accounting framework, has the merit of being consistent with the data on employment that enters in the Frame via the business register. Moreover, the data of social security in Istat are the cornerstone of the business based labour market statistics, that is the statistics produced to comply with the regulations on Labour Cost Index, Structure of Earning Surveys and more recently, Labour Cost Survey. Consequently, the use of this data for the frame and the Structural Business Statistics would have carried out the extra advantage of producing horizontal consistency among statistical domains, that is the final aim of the incoming Frame Regulation Integrating Business Statistics (FRIBS).

However, since the first analysis, it has been clear that the two kind of sources have some non-negligible differences. To decide which source was more suitable for the Frame has, thus, requested a study to respond to some basic questions: why the comparison between these sources show such discrepancies? Which are the differences in the definitions of the wages between sources? Which is the most appropriate source for the SBS purposes? An in-depth analysis of the empirical and theoretical differences between the sources has been necessary and has required the involvement of many competencies: national account expertise, knowledge on social security data and on accounting theory and practices, statistical and data mining skills. The analysis has been rewarded by a far more clearer understanding of the contents of the company accounts and social security data, and, since the administrative obligations from which they stem are at the base of the enterprises informative systems, of the sources from which the enterprises draw information to fill in the statistical questionnaires. This led to a better use of all available sources to produce the Frame labour cost variables.

This paper reviews this statistical exploration and reports the main findings and the implications for the building of the frame. Since the SBS regulation is referred to the Industries and Services private sectors excluding Financial activities (NACE rev 2. Sections B to S excluding K) what follows is only referred to this population of enterprises. The structure of the document is as follows. Paragraph 2 describes the main administrative obligations of the enterprises and how they shape the accounting bookkeeping. Paragraph 3 illustrates the main empirical differences in the labour cost variables between these sources. Paragraph 4 analyzes the definitions of the different sources at a very detailed level and contrasts them with the SBS and European System of Accounts (ESA) regulation. Paragraph 5 deepen the analysis of Par. 3 to check whether some of the theoretical discrepancies may explain the differences found in the data. Paragraph 6 reports the first attempt of correcting the source to reduce the causes of bias with respect to the statistical definitions.

2. Sources and details of the labour cost variable

The input sources that feed the Frame all have information on total labour cost (TLC), with different level of detail but also different consistency with the desired content (Casciano et al, 2011; Curatolo et al, 2015).

Besides the Profit and Loss part of the company accounts (BIL) and the fiscal sources (Tax returns – UNICO, and statistical studies for the estimation of taxes due by firms in specific industries - SDS), the Frame can benefit from a supplementary source, that have only information on employee TLC, so it is not useful for the other economic variables required by the SBS Regulation, but has valuable and sound information on employees and their corresponding costs. This source is the new Employee Wage Register (whose Italian acronym is RACLI) with the individual level data linked to the employee ones. This latter register has been produced, since 2011, mainly using the new archive of social security declarations of enterprises with at least one employee. Starting from January 2010, due to an important change in administrative obligations, firms have to present, within 30 days from the reference month, a new administrative declaration, the "Uniemens" concerning both individual and firm data to the main social security institute (INPS). As a consequence, very detailed and timely administrative data for each employee, the characteristics of his jobs, and the associated firm are now available. The exploitation of this new administrative data source was accelerated thanks to the 2011 Industry and Services Census, so that the new (virtual) Census estimates on employment has been based on the new individual-level Employment Register (that is the base of the Business Register aggregated for enterprises). To this linked employer-employee job register, wages and salaries have been coherently added after procedures of check, editing and integration. The labour costs variables other than wages are at the moment not available in the individual declarations used at Istat. They are added to the RACLI register in the following manner. The amounts retained by the enterprises each year for financing the severance payment (TFR) are estimated at individual level applying the usual rate; the social contributions either paid to INPS or to other institutes (e.g. contributions to insurance schemes for occupational accidents) are are obtained, for the time being, only at firm level Using the data processed by Istat Oros Survey[9] (Baldi et al., 2008). In the near future they will be calculated at individual level starting from the Uniemens declarations. An interesting pioneering attempt to estimate the labour costs at individual level, even if with different sources available, is contained in Grand and Quaranta (2014). The new RACLI register so obtained could contribute to the Frame only for the TLC and its components, that is for a little part of the SBS target variables which are filled in using other administrative information such as profit and loss account and other fiscal data.

Regarding personnel costs, one of the main characteristic of RACLI is that wages and total labour cost are coherent with employment estimations of the Employment/Business Register itself, while in the other administrative sources of the Frame the information on the number of employees is not available at all (see table 1). On the other hand all the other Frame sources have (more or less) information on TLC, but only the company account

[9] The calculation of the labour costs other than wages was done mainly through the virtual DM10 form reconstructed for each firm by the social security institute (INPS) and elaborated by the Istat Oros Survey; nevertheless in the near future it will hopefully calculated at individual level

(BIL) has the desired detail to estimate not only the complete labour cost, but also its components.

Table 1 - The total labour cost variable and its details in the different sources of the FRAME

VARIABLES	BIL	SDS	UNICO	RACLI
Total Labour Cost (TCL)	X	X	X	X
- wages and salaries	X	-	-	X
- social contributions	X	-	-	X
- severance payments (TFR)	X	-	-	X
- other costs	X	-	-	-
Employees	-	-	-	X

Besides differences in the details of the variables the sources have quite different coverage. The compilation (and dissemination) of profit and loss accounts is compulsory by law only for incorporated enterprises, while it is optional of unincorporated enterprise, so BIL covers all Incorporated enterprises and few unincorporated enterprises, thus being concentrated especially on medium large firms. SDS covers all enterprises in most market sector of economic activities with turnover below (about) 5millions thus representing especially small firms. UNICO is the fiscal statement that all unincorporated partnership, professionals and owner of individual enterprises have to fill in, therefore it covers mainly small-medium unincorporated enterprises; and RACLI covers all enterprises with at least 1 employee.

In order to define the better and correct use of each source for the statistical aim at hand, a comparison on employee costs among the different sources to understand their pros and cons has been done. Before going into the analysis is helpful to clarify where the different sources originates from, that is looking into the enterprises accounting and information systems.

3. Information systems within the company: payroll accounting and general accounting

The evaluation of labour cost is mainly the result of fiscal, social security and labour duties arising from the payroll accounting and the general accounting obligations for the enterprises.

Payroll accounting is a subfield of the general accounting bookkeeping, and along with the other subject areas, is an integral part of the company management information system. It includes all the financial records related to employee salaries, withholdings and deductions, the relationships with the social security institutions and the full documentation related to employment.

Establishing, conducting and terminating employment relationships of a "subordinate" or "pseudo-subordinate" nature requires many administrative and accounting operations, as well as obligations towards social security institutions. Some of these duties are generated only at the beginning of the employment, at its termination or whenever changes to the

employment relationship occur. Other duties are instead periodical and regularly required and repeated throughout the duration of the employment for each employee: such as payslips and payroll processing, the issuance of the “Libro unico del Lavoro” (a Single Employment Ledger containing the payroll records and schedule), tax and social security withholdings calculations and payments, and the monthly declarations of the social contributions “Uniemens”.At the end of each year, employers are required to meet the compliance for the tax and national insurance obligations.

Since the firms have the responsibility to keep record of all the events that have economic significance during the year, all obligations end up being recorded by the general accounting. General accounting entries are prepared using a chart of accounts, which is a structure containing all the items covered in the general ledger. By using special dedicated accounts, all the administrative transactions between the company and the external environment are recorded on the specific general ledger accounts with each of these accounts representing a type of expense (or income), debit (or credit), etc. The accounting journals record therefore the liquidation and payment of monthly salaries, the contributions paid by the company, withholding taxes, social security and welfare and their payment, the provision for employee severance indemnities, and any liquidation of it (TFR) in case of termination of the employment relationship. It is important to note that these records are prepared in accordance with general accounting principles, and the financial statements are prepared on the basis of the accounting entries related to a particular financial year. These principles are at the base of the differences between the concepts recorded in the payroll accounting and the one recorded in general accounting.

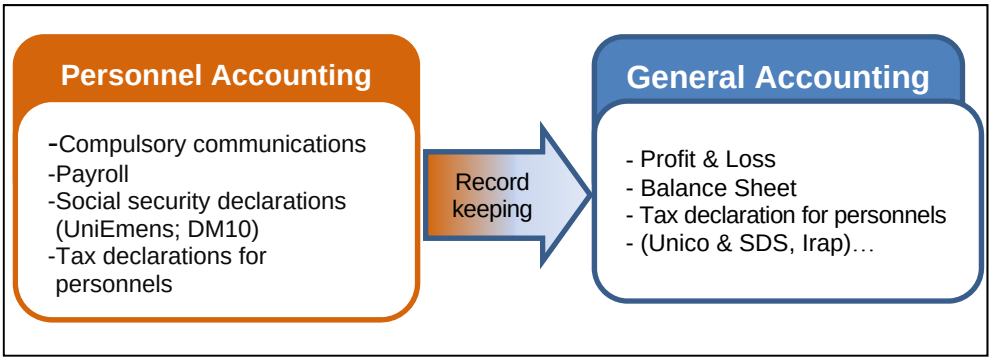
In accordance with the general accounting principles, specifically n. 12 of the Interpretative Document N.1 of the accounting standard, for each period there should be recognized "...the wages and salaries...including the portion relating to accrued and unpaid bonuses and leaves accrued but not taken before withholding taxes and social security contributions paid by the employee ... ", in addition to "...charges relating to the shares of the bonuses and holiday pay accrued and unpaid..". Therefore, in the yearly business closure, firms register the adjustment necessary to show the overall cost of employment based on the specific financial year’s accruals, and the liabilities to the employees accrued but unpaid, in accordance with the principle of accrual regardless of the date of collection or payment.

Payroll Accounting data is summarized in some specific items of the budget, especially in voice B9) personnel costs.

Those firms that adopt the system of ordinary accounting are required to prepare financial statements, and for corporations, there is an obligation to register their financial statements with the Companies Register kept by the Chamber of Commerce (C.C.I.A.A.). For individual companies, professional partnerships and collaborations, a trial balance is prepared, which involves the juxtaposition of costs and revenues, which indicates the cost of personnel. The latter however is not registered and therefore remains an internal document.

The fiscal sources UNICO SDS and IRAP (declarations for regional taxes on production activity) are then prepared based on the data from trial balance and financial statements, however, since each of these returns has different purposes, the Payroll data entered in each of them is valued differently.

Figure 1 – The different accountancy within each firm



One of the consequence of the all these different accounting duties each firms has to implement is that the use of administrative source for statistical purposes is influenced by the firms accounting system they refer to.

4. An overall picture of the differences among sources on labour cost

This section aims to highlight empirically the differences between the variables related to wages and salaries and labour costs among the available sources. In particular, the comparison of the wages and salaries is possible only between RACLI and BIL sources, while that on total labour costs can be carried out on the three sources RACLI , BIL and SDS. The results shown in the tables following are referred to the year 2010 but the analyses on the previous year data basically confirm the same evidences.

To get a first measure on the impact on aggregate data of choosing one source instead of another we use the percentage difference measure, calculated on the matched firms between the two sources as follows:

$$d1 = 100 * \frac{\sum_i v_{ai} - \sum_i v_{bi}}{\sum_i v_{bi}}$$

where $v_{.i}$ is the value of the examined variable (wages and salaries or labour cost), for firm i in the data source a or b (RACLI or BIL or SDS). The sum is over the matched firms in a given NACE section.

Table 2 shows that this difference is systematically negative in each section of economic activity for both wages and labour cost variables when the RACLI source is in the comparison. That is RACLI values are lower than those of the other two fiscal sources. At the same time, the differences between SDS and BIL, are quite small with no systematic clear sign.

Table 2 - Comparison measures, based on d1 indicator, on wages and salaries (RACLI-BIL) and on labour cost (RACLI-BIL, RACLI-SDS, SDS-BIL), with or without extreme values (unweighted)

SECTION OF ECONOMIC ACTIVITY	Wages and salaries			Total Labour Cost								
	RACLI(a) vs BIL(b)			RACLI(a) vs BIL(b)			RACLI(a) vs SDS(b)			SDS(a) vs-BIL(b)		
	d1[1]	*d1_t5*[2]	*d1_t10*[3]	*d1*[1]	*d1_t5*[2]	*d1_t10*[3]	*d1*[1]	*d1_t5*[2]	*d1_t10*[3]	*d1*[1]	*d1_t5*[2]	*d1_t10*[3]
B - Mining and quarrying	-4.3	-3.7	-3.5	-4.5	-4.3	-4.0	-2.8	-2.9	-2.8	-0.7	-0.3	0.0
C - Manufacturing	-4.6	-4.1	-3.7	-5.2	-4.8	-4.4	-3.3	-3.5	-3.2	-0.8	-0.3	-0.1
D - Electricity, gas, steam and cond. supply	-4.7	-4.8	-3.8	-6.4	-6.8	-6.4	-9.8	-10.8	-9.8	1.2	1.2	0.0
E - Water supply, sewerage, waste, activities	-6.4	-5.9	-5.5	-8.1	-7.6	-7.1	-6.5	-6.0	-5.6	-1.3	-0.7	-0.4
F - Construction	-7.2	-6.3	-5.7	-7.3	-6.9	-6.4	-4.5	-4.3	-4.0	-0.3	0.0	0.2
G - Wholesale and retail trade	-5.2	-4.1	-3.6	-6.2	-5.2	-4.7	-3.5	-3.2	-2.8	-1.2	-0.3	0.0
H - Transportation and storage	-15.9	-12.4	-11.8	-15.3	-12.0	-11.5	-13.2	-13.1	-12.8	-4.7	0.2	0.3
I – Accomodation, food service activities	-6.4	-2.7	-2.3	-6.7	-3.0	-2.7	-0.8	-0.9	-0.8	-1.6	-0.4	-0.1
J – Information, communication	-8.0	-7.0	-6.2	-9.3	-8.3	-7.6	-7.4	-7.1	-7.0	-1.1	-0.3	0.1
K - Financial and insurance activities	-8.0	-5.9	-5.3	-9.3	-7.8	-7.2	-6.8	-6.3	-5.5	-2.0	-1.2	-0.3
L - Real estate activities	-6.7	-4.8	-4.1	-9.2	-7.0	-6.0	-4.7	-3.9	-3.6	-2.2	-1.7	-0.7
M - Professional, scientific technical activities	-8.1	-6.5	-5.8	-9.6	-8.1	-7.4	-5.3	-5.0	-4.7	-0.8	-0.1	0.3
N - Administrative support activity	-11.9	-9.5	-8.0	-12.4	-10.4	-9.0	-7.9	-6.6	-6.1	-0.5	-0.8	-0.3
P - Education	-8.2	-7.1	-6.2	-9.1	-8.0	-7.2	-4.5	-4.4	-4.0	-2.3	0.6	0.7
Q - Human health, social work activities	-6.6	-5.7	-5.1	-7.1	-6.3	-5.8	-3.3	-2.4	-2.1	0.5	-0.2	-0.1
R - Arts, entertainment and recreation	-46.3	-12.5	-9.5	-41.9	-11.9	-9.1	-7.2	-6.1	-5.1	0.0	-0.6	-0.3
S - Other service activity	-7.6	-5.9	-4.7	-8.0	-6.7	-5.6	-2.1	-1.5	-1.1	-1.6	-0.7	-0.3
B - S Total	**-7.2**	**-5.5**	**-4.9**	**-7.7**	**-6.3**	**-5.7**	**-4.3**	**-4.2**	**-3.9**	**-1.1**	**-0.3**	**0.0**

1 - d1: calculated on the original distribution with extreme values

2 - d1_t5: calculated excluding, within each Nace, the lower and upper 2.5% of the distribution of the differences

3 - d1_t10: calculated excluding, within each Nace, the lower and upper 5% of the distribution of the differences

Since the d1 indicator is sensitive to extreme values the same indicator has been computed without extreme values namely excluding, within each NACE, the lower and the upper 2.5% and 5% of the distribution of the differences (respectively d1_t5 and d1_t10

indicators). The values of these last indicators decrease appreciably (table 2). Moreover the reduction of the difference due to the exclusion of extreme values is less between RACLI and SDS and between SDS and BIL than in the comparison RACLI–BIL. These results signal that part of the problem may be due to outlying observations, especially between these last two sources. However the RACLI source continues to measure lower labour costs compared with the other sources suggesting that these differences are structural.

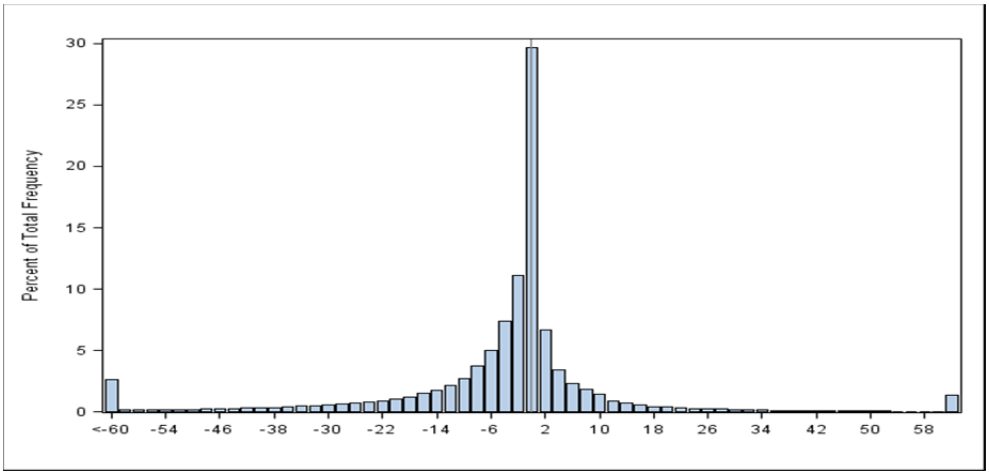
To deepen the analysis in this direction, we move to a micro level comparison, that is examining the distribution of the differences between matched firms. Since d1 is an asymmetrical measure, the following analysis is performed with the measure below:

$$d2_i = 100 * \frac{v_{ai} - v_{bi}}{(v_{ai} - v_{bi})/2}$$

where d2 is a symmetrical measure varying between -200 and +200.

Figure 2 show the distribution of d2 for the variable wages and salaries between RACLI and BIL.

Figure 2 – Distribution of d2 differences on wages and salaries between RACLI and BIL.



In order to measure a bias, not influenced by outliers possibly due to definitions differences, the main location measure we use is the median. To provide a magnitude of the similarity of the measures we use the following two indicators calculated on d2:

$$freq2=100*(F_{-2<=d2<=+2}/F)$$

$$freq5=100*(F_{-5<=d2<=+5}/F)$$

where F is the number of firms and freq2 and freq5 represent the percentage of firms, on total matched firms, having a d2 distance respectively between ∓ 2 or ∓ 5. Empirical evidence on the wages and salaries differences between RACLI and BIL is shown in the following table broken down by legal form, size classes and economic activity sections of the firms. The d2 median distance is always negative. On average 40% is the percentage of firms with a differences within 2% and this percentage rise to almost 60% for differences

within 5%. Among economic activities, the transportation sector scores the highest differences and one of the lowest percentage of firms with a difference lower than +/- 2%. In the transportation sector freq2 and freq5 are significantly lower than the average ones, implying a greater structural difference among the data compared. Regarding the firm size classes, based on the number of persons employed, the negative d2 median grows with size. Considering the legal form, mutual co-operative societies show the lowest freq5.

Table 3 - Comparison measures, based on d2 indicator, on wages and salaries between RACLI and BIL matched units, by legal form, size classes and sections of economic activity.

	N. Firms	d2_Me	freq2	freq5
Section of economic activity				
B - Mining and quarrying	1,145	-0.7	43.2	64.5
C - Manufacturing	96,956	-1.1	43.1	63.3
D - Electricity, gas, steam and conditioning supply	1,239	-2.0	33.6	53.4
E - Water supply, sewerage, waste management and remediation activities	3,629	-2.3	36.3	54.5
F - Construction	70,205	-1.6	27.3	47.9
G - Wholesale and retail trade; repair of motor vehicles and motorcycles	115,525	-0.2	45.4	62.7
H - Transportation and storage	20,982	-6.3	26.7	39.3
I - Accomodation and food service activities	31,379	0.0	48.9	65.1
J - Information and communication	22,044	-1.6	36.5	54.1
K - Financial and insurance activities	3,903	-1.2	40.4	57.8
L - Real estate activities	13,916	0.0	42.2	59.0
M - Professional, scientific and technical activities	26,629	-1.5	37.2	55.1
N - Administrative and support service activities	24,517	-1.7	34.6	51.3
P - Education	4,148	-0.3	39.6	55.5
Q - Human health and social work activities	11,242	-0.6	45.0	61.6
R - Arts, entertainment and recreation	7,341	-0.8	37.7	51.4
S - Other service activities	8,508	-0.1	39.6	59.0
Size class				
00-01	17,058	0.0	41.7	47.9
02-03	132,249	0.0	41.5	56.5
04-05	81,399	-0.5	39.8	58.1
06-09	89,958	-1.1	39.0	59.1
10-19	85,972	-1.6	38.5	59.0
20-49	44,375	-2.1	36.8	58.5
50-99	12,297	-2.2	36.2	59.2
Legal form				
LF1 - Sole proprietorship, individual entrepreneur, self employed and own account worker	1	0.0	100.0	100.0
LF2 - Partnership	493	-0.9	41.0	60.6
LF3 - Joint-stock companies	21,546	-1.6	38.3	61.7
LF4 - Limited liability companies	406,374	-0.8	39.9	57.9
LF5 - Prevalently mutual co-operative societies	21,093	-1.2	35.1	49.4
LF6 - Other co-operative societies	8,494	-1.2	39.7	56.7
LF7 - Consortium	5,244	-1.1	37.1	55.7
LF8 - Municipal companies	29	-1.3	44.8	58.6
LF9 - Other legal form	34	-8.1	23.5	35.3
Total	463,308	-0.9	39.6	57.7

The same indicators have been calculated for the labour cost variable including in the comparison the other source (Table 4). The comparisons are performed within pair of sources and include a different number of firms, since each source cover different portions of the economy.Looking at the total distribution, negative d2 median, equal to -2.7, for RACLI–BIL is higher than that for wages and salary so for this latter there is a best matching data between this two sources. On the other hand the RACLI distance from SDS on labour cost is in general lower than that from BIL.

Table 4: Comparison measures, based on d2 indicator, on Labour Cost (RACLI-BIL, SDS-BIL, RACLI-SDS) by legal form, size classes and sections of economic activity.

	RACLI - BIL			RACLI - SDS			SDS - BIL		
	N. Firms	*d2_Me*	*freq2*	*N. Firms*	*d2_Me*	*freq2*	*N. Firms*	*d2_Me*	*freq2*
				Section of economic activity					
B - Mining and quarrying	1,146	-2.4	28.8	1,526	-1.6	27.8	956	0.0	85.1
C - Manufacturing	96,954	-2.4	33.3	203,847	-1.3	34.0	73,807	0.0	79.0
D - Electricity, gas, steam and cond. supply	1,239	-3.5	26.1	41	-6.2	17.1	35	0.0	65.7
E - Water supply, sewerage, waste, activities	3,630	-5.0	21.7	1,873	-3.6	24.8	995	0.0	77.6
F - Construction	70,197	-3.9	18.3	209,516	-1.7	17.0	59,111	0.0	80.9
G - Wholesale and retail trade	115,527	-2.2	35.0	325,800	-0.7	38.6	94,155	0.0	81.9
H - Transportation and storage	20,986	-8.3	17.2	39,003	-9.8	13.2	13,881	0.0	78.6
I – Accomodation, food service activities	31,377	-0.6	42.6	165,536	0.2	40.6	26,489	0.0	88.1
J – Information, communication	22,047	-3.1	29.7	27,902	-2.7	29.5	17,916	0.0	69.6
K - Financial and insurance activities	3,902	-3.4	30.2	19,163	-1.8	34.8	3,350	0.0	71.6
L - Real estate activities	13,918	-1.7	34.6	23,700	-0.8	34.2	12,359	0.0	82.2
M - Professional, scientific technical activities	26,626	-3.4	28.9	96,166	-1.0	36.1	18,771	0.0	69.4

Table 4 continue: Comparison measures, based on d2 indicator, on Labour Cost (RACLI-BIL, SDS-BIL, RACLI-SDS) by legal form, size classes and sections of economic activity.

	RACLI - BIL			RACLI - SDS			SDS - BIL		
	N. Firms	*d2_Me*	*freq2*	*N. Firms*	*d2_Me*	*freq2*	*N. Firms*	*d2_Me*	*freq2*
				Section of economic activity					
N - Administrative support activity	24,516	-3.8	26.0	34,684	-1.7	28.7	14,966	0.0	74.9
P - Education	4,150	-2.4	31.6	2,921	-2.0	30.9	548	0.0	71.4
Q - Human health, social work activities	11,237	-2.5	35.6	52,738	-0.4	41.7	4,924	0.0	79.0
R - Arts, entertainment and recreation	7,344	-2.3	26.7	9,582	-0.3	33.5	3,344	0.0	78.0
S - Other service activity	8,507	-1.9	31.5	66,802	0.6	33.6	7,010	0.0	79.0
				Size class					
00-01	17,060	-3.3	19.4	69,403	0.8	19.0	11,321	0.0	80.0
02-03	132,251	-2.4	29.8	586,560	-0.3	32.2	103,246	0.0	80.8
04-05	81,391	-2.4	30.9	273,271	-1.0	34.4	66,774	0.0	80.3
06-09	89,961	-2.6	31.4	198,398	-1.5	36.0	74,176	0.0	79.7
10-19	85,967	-2.9	31.4	116,433	-2.1	35.0	67,754	0.0	78.1
20-49	44,377	-3.2	30.7	33,410	-2.7	32.2	26,412	0.0	76.8
50-99	12,296	-3.3	30.4	3,325	-3.2	29.5	2,934	0.0	76.4
				Legal form					
Sole proprietorship, individual entrepreneur	1	-27	00	566,478	0.1	33.8			
Partnership	493	-21	314	333,240	-1.1	35.0	388	0.0	82.2
Joint-stock companies	21,544	-33	291	8,860	-3.0	27.5	8,404	0.0	74.7
Limited liability companies	406,368	-26	307	356,077	-2.4	29.6	330,150	0.0	79.8
Prevalently mutual co-operative societies	21,101	-31	256	11,611	-2.5	25.0	10,118	0.0	79.9
Other co-operative societies	8,490	-34	289	2,217	-4.0	21.9	2,035	0.0	75.3
Consortium	5,243	-28	296	1,742	-3.1	26.8	1,490	0.0	73.6
Municipal companies	29	-161	103	95	-5.1	18.9	11	0.0	72.7
Other legal form	34	-86	265	480	-5.6	19.8	21	0.0	47.6
Total	463,303	-2.7	30.3	1,280,800	-0.8	32.8	352,617	0.0	79.6

Finally it is important to stress that the d2 median for BIL-SDS is zero (that is, BIL values almost equal to those of SDS for total labour cost) and the percentage of firms with a difference not higher than +/- 2% is much higher compared to that of RACLI with the other sources. These findings are expected since BIL and SDS are both fruit of the general accounting, while RACLI descend from the personnel accounting. RACLI shows higher similarity with SDS and this might be partially due to the fact that SDS covers also smaller enterprises, with respect to BIL, for which lower differences have been recorded. It is likely that smaller firms have a general accounting bookkeeping less structured than larger firms and then less different from personnel accounting.

The results of these analyses together with the consideration on the different system of account within the firms (§3) lead to explore more in deep the theoretical definition of the same variable in different administrative data and in particular in data drown by the social security system (RACLI source) and those based on fiscal and general account data (BIL and SDS).

5. Main theoretical hints on differences on TLC definition: social security vs profit and loss account vs SBS Regulation

According to the definitional content of the different sources on the TLC, there are some relevant differences among the data we are taking into consideration.

Table 5 briefly summarizes these main differences among the definitions of profit and loss and fiscal sources in general (BIL/SDS/UNICO), social security data (RACLI) and the SBS Regulation which is our theoretical benchmark.

On one hand this regulation only states general contents of labour costs variable suggesting that only the remuneration of work done should be included. On the other hand,the SBS explicit reference to company accounts means that the SBS definitions are delegated to the company accounts rules and intepretations and they leave space or in some cases the official interpretations explicitly indicates, to include in labour costs also items that could be more properly intermediate costs and viceversa to not include items that should be remuneration of the work done.

Going more in depth on the actual content of each financial statement item and on the accounting practices the pros and cons of each source to measure the statistical definitions can be highlighted. The inconsistency of the input sources with the desired definitions may derive from a different definition of the underlying employment, or different content of wages and salaries or differences imputable to social contributions. Some of the items may have a negligible impact, while others can lead to biased estimates of labour costs.

As concerns employment, in company account (BIL), the definition of the personnel costs may include not only costs for employees but also for agency and project workers. The official interpretations of the company accounts, in fact states that, for the principle of prevalence of substance to form. These external workers may be "assimilated" to employees[10]. In the statistical definition, of course, agency workers and "external workers"

[10] This principle descends from fiscal rules, for which all external workers (agency workers, outworkers) costs are defined as "assimilated" to employees costs.

or "outworkers" have to be excluded from personnel costs which should be referred only to employees.

Table 5 - Difference among sources and between each source and SBS Regulation

CAUSES OF DIFFERENCE	BIL - SDS/UNICO - (Accounting and fiscal data)	RACLI (social security data)	Theory (SBS Regulation[11])
Employment			
Agency workers	Included	Not included	Not included
Outworkers	May be included, according to accounting rules	Not included	Not included
Working associates	May be included, according to accounting rules	Not included	Not included
Workers on secondment Host firm	May be included, according to accounting rules	Not included	Not included
Workers on secondment: Home firm	Included	Included	Included
Employees hired abroad	Included	Non included	Not included
Registration principle	Accrual	Cash	Accrual
Wages and salaries			
Meal voucher	Not included	Included only the part above 5,29 euro per voucher	Included
Fringe benefits	Not included	Totally included if above 258,23 Euro (otherwise excl.)	Included
Exceptional payment for leaving the enterprise	Included (in "other costs")	Excluded	Included
Stock options	Included (often in "other costs")	Excluded	Included
Travel allowances	Included	Partially included (amounts above fixed limits)	May be included
Other labour costs			
Yearly provisions/payment to funds for severance pay (TFR)	Included	Estimated	Included
Supplementary contribution	Included	Excluded	Included
Providences to employees	Included	Excluded	Included

About SDS source, the variable defined as personnel cost includes some items that cannot be statistically considered as employee costs. But this elements can be deduced from the total since they are detailed with separate sub-codes:

- expenditures that compensate self-employed services;
- expenditures for use of outworkers or agency workers;
- remuneration to associates for activity as administrators.

So for the statistical aims the TLC variable is obtained by subtracting these sub-items[12].

In UNICO, the item personnel costs include by definition expenditure for employees

[11] The definitions in the SBS Regulation refer to company accounting practice without details of specific items that have to be included/excluded. On the other hand SBS data are the input for national accounts that instead have to refer to ESA 2010 highly detailed definitions.

[12] Since the completeness and quality of these sub-items is questionable because they are not relevant for administrative purposes doubts remains on the accuracy of the final figures on TLC when the firms report these kind of items.

and "assimilated" workers and in some cases also costs for self-employed services. Moreover in this source separate details about the costs of this "assimilated personnel" is not available but for statistical purposes they should be estimated and excluded from personnel costs.

Another relevant inconsistency derived from different boundaries of employment that can lead to significant discrepancies among the examined sources is related to workers on secondment. According to the regulation their cost should be recorded in the labour costs of the organization for which they are at payroll (home organization) and in intermediate costs for the host firm. However the accounting principle establishes that they can be included in host firm labour costs in BIL so there is a potential risk of duplication of TLC aggregate values due to this kind of firms.

Focusing on inconsistencies derived from different boundaries of wages and salaries the main aspects are related to the following items.

- Meals vouchers and fringe benefits: they should be included in wages and salaries according to Regulation, while they are recorded as intermediate consumption in BIL due to the principle for which costs are classified according to their nature. In RACLI only the amount above fixed limits is included in wages and salaries.
- Exceptional payment for leaving the enterprise and stock options: they should be recorded in wages and salaries according to SBS definition while they are included in BIL labour costs but under other personnel costs. They are not at all in RACLI figures because not subjected to social contributions.
- Travel allowances: SBS definition does not mentions them specifically so it is hard to determine whether they should be accounted in intermediate costs or in labour costs[13]. They are included in wages and salaries of BIL-SDS-UNICO, while they are only partially included (amounts above fixed limits) for RACLI since only the part considered compensation for the employee is subject to social contributions.

Differences can be referred also to social contributions. About severance pay, for example, that should be included in social contribution definition, it is included in BIL-SDS-UNICO labour costs while RACLI provides only an estimate. More in general, some components of social contributions are only estimated or missing in the RACLI Register. Although these should account for tiny amounts of the social contributions this raises doubts on their use for the SBS Regulation requirements.

The comparison analysis on the TLC variable put on evidence that none of the administrative data have exactly the information requested for the statistical aim: each source has both pros and cons. In other terms, there is not a unique source with the desired content of employees wages, other labour costs and total labour cost, each source should be adjusted.

In particular the use of fiscal sources on TLC implies necessarily some corrections for the costs not pertaining employees. On the other hand, in RACLI social security based data the labour cost is referred only to employees but there are some other aspects to remember. First of all this information do not include costs for other benefits to employees (they are not collected for social security data neither estimated afterwards). Secondly, social contributions are partly directly derived from social security data, while the part due to

[13] For ESA 2010 travel allowances are intermediate costs.

other institutions (like contributions to insurance schemes for occupational accidents and diseases) is estimated.

So we had two main alternative possibilities about the use of the different data source in the Frame: the integration of Personnel Costs from the register based on social security data or the use of these latter data to correct the fiscal source on the same variables. The choice was basically between using the same sources with the same priority as those supplying the other economic variables, hence disposing of a coherent accounting framework or using RACLI for labour costs, with the advantage of having the same sources that is the base of the estimate of employment, thus keeping the maximum consistency with the labour input. To enhance the coherence among all economic flows and balancing items it has been preferred to gather all information from the same sources. So the final decision was to use the same source for all SBS variables including personnel costs but using social security information to apply some corrections according to the statistical definitions.

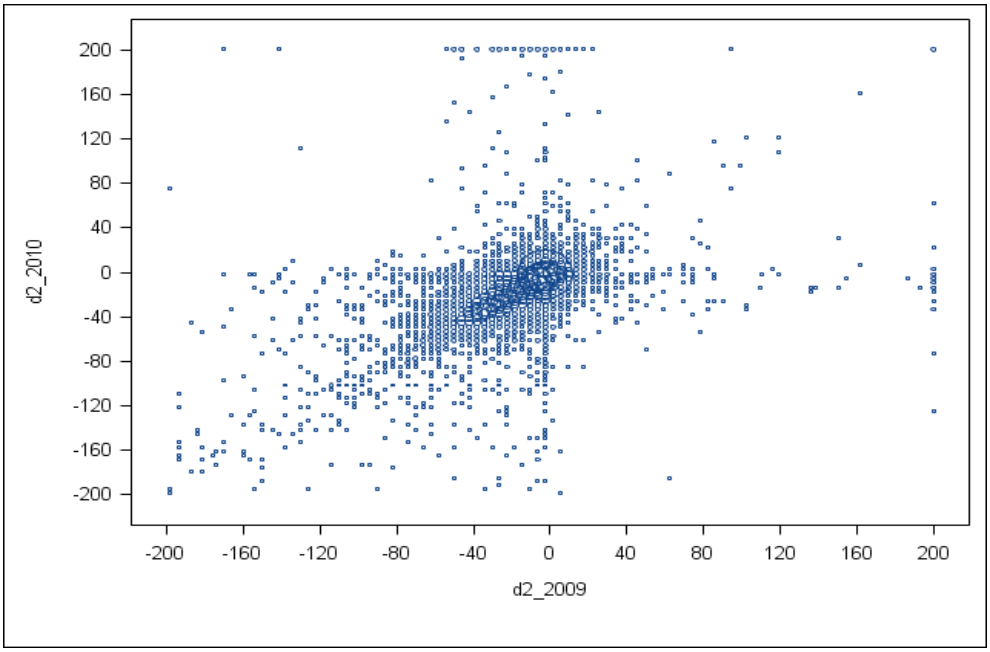
6. Empirical hints on differences among sources

6.1 The causes of the differences

The analysis of paragraph 4 has provided a descriptive overview of the basic measures of the differences between company accounts and social security data. The differences are quite widespread among economic sectors, size classes and legal forms even if in some cells, notably in transport sector (section H), their size is more marked. In this paragraph the analysis is deepened trying to understand which are the main drivers of these discrepancies.

A first question to which is necessary to respond is whether the differences are persistent over time. As stated above the size of the differences are rouglhly the same between 2009 and 2010. To investigate further, figure 3 plots the differences on wages and salaries between BIL and RACLI for each enterprise observed in 2009 against those observed for 2010 in section H. The data are binned in classes of differences to avoid overplotting and the size of the bubble indicates the number of enterprise in that bin. Since the data are concentrated on the diagonal of the graph it is possible to conclude that enterprises with the larger difference in one year are also those with the larger differences in the subsequent year. This hints at the facts that the differences are explained by structural characteristics of the units and are not due to contingencies.

Figure 3. D2 distances on wages and salaries BIL-RACLI: 2009 versus 2010.



Once proved the persistency over time of the difference between the sources on the same variable, to derive some insights on the causes of the differences a look in the labour cost structure may be of help even if it is fully available only for BIL. One such insight come from the analysis of the severance payment (TFR) that is the amount of wages retained by the enterprises (or versed to an external Fund) each year to be provided to the workers at the termination of the labour contract. For most enterprises its size is about 7.4% of all non-occasional elements of the total wages plus a part due the appreciation of the fund already accumulated. Typically the distribution of the TFR should concentrate on the range of 6-9% of the total wages and salaries as table 6 confirms (63.9% of firms in all sections). The percentage of firms in the entire economy for which this ratio is below what expected is 25.1% (14.4% between 1 and 6 and 9.7% below 1). It interesting to stress that in section H, this shares rise to 40.5% (29.3% between 1 and 6 and 11.2 below 1).

Table 6 – Firms by classes of the percentage ratio of TFR on wages in the BIL source

FIRMS	% Ratio of TFR on wages					
	missing	<1	1-6	6-9	>9	all
	All sections					
number	2,667	44,761	66,822	296,343	53,072	463,665
%	0.6	9.7	14.4	63.9	11.4	100
	section H: Transportation and storage					
number	137	2,360	6,152	10,940	1,409	20,998
%	0.7	11.2	29.3	52.1	6.7	100

Tables 7 suggests that this feature is associated with the differences in wages between the two sources, RACLI and BIL. In section H, while the d2 median of the difference is only -2.2% for firms that have a ratio of the TFR over wages in the normal range 6-9%, it jumps to -25.3% for the firm with a TFR in the range 1-6%.

Table 7 – Distribution of the difference RACLI-BIL in wages and salaries of transportation and storage section by classes of the percentage ratio of TFR on wages

% Ratio of TFR on wages	D2 RACLI-BIL wages				
	N	Mean	Q1	Median	Q3
<1	2,359	-20.9	-32.3	-10.3	0
1-6	6,147	-26.9	-35.7	-25.3	-12.7
6-9	10,924	-5.7	-10.2	-2.2	0
>9	1,409	3.7	-4	0	5.3
All	20,839	-13.1	-22.5	-6.5	0

This evidence suggests that the firms that show larger negative differences between the two sources may have part of their total wages not subject to the TFR provision. This may occur in either of two situations. The first is when part of the wages indicated in the profit and loss account are paid to workers that are not employees (as mentioned in paragraph 5 and measured in paragraph 6.2) and for whom the legislation does not envisage the TFR provision. The second is when company accounts wages include some components that are not subject to the TFR. The base for TFR computation is composed by all the non-occasional elements of wages. According to the Italian legislation, however not all amount paid to the employees are considered wage from the point of view of tax and social security contributions. The general principle is that only elements that are income for the worker are considered taxable and subject to social contributions. One element that is only partially subject are travel allowance (which weight in transport sector is relevant) since part of it is a refund of expenses. In the accounting practice travel allowances are totally included in wages and salaries while they are only partially included for Racli (see table 5), since for the social security contribution the amount of transfer allowance above fixed limits has to be considered a refund for the employee so no contributions has to be paid on it. At the same time, the exceeded part is considered remuneration so it has to be subjected to the payment of the contribution rate.

6.2 The outworkers

Table 8 confirms that profit and loss account may include in the voice “Personnel cost” also workers not properly employees, like staff workers and project workers. It compares the distribution of the d2 difference of wages between RACLI and BIL separately for firms which have only employees and firms with at least one outworker. This information is drawn from the newly available Business register that provide, for each enterprise, the number and types of employees and of outworkers attached to it and in some cases their costs.

The evidence is striking: while the median difference of firms with only employees waver around zero, that of firms with outworkers is systematically negative. This is even more clear when looking at the first quantile of the distribution. What is probably occurring is that, for many firms, the wages in the accounts data include the compensation for

outworkers while, by definition, the social security source, used in this comparison does not. The table, however, shows also that in Transportation, the sector with the highest negative difference, passing from one type of enterprises to the other, the distribution does not shift much to the left, leaving the parameters of the distribution roughly unchanged.

Table 8 – Distribution of the differences on wages between RACLI and BIL separated for firms with/without at least one outworker.

Nace sections	without outworker				with at least one outworker			
	N. firms	Q1	Median	Q3	N. firms	Q1	Median	Q3
Total (B-S)	316,529	-5.0	-0.1	0.8	146,938	-10.9	-2.5	0.1
Industry (B-E)	60,998	-3.9	-0.3	0.6	42,067	-7.8	-2.1	0.0
Industry (B-F)	117,843	-5.0	-0.4	1.3	55,403	-9.5	-2.5	0.0
Services (G-S)	198,686	-5.0	0.0	0.6	91,535	-12.1	-2.5	0.1
B - Mining and quarrying	802	-4.2	-0.3	0.5	344	-6.8	-1.3	0.4
C - Manufacturing	57,545	-3.8	-0.2	0.6	39,499	-7.7	-2.1	0.0
D - Electricity, gas, steam and air conditioning supply	593	-6.2	-1.1	0.2	650	-10.0	-2.5	0.1
E - Water supply; waste managment	2,058	-7.7	-1.1	0.1	1,574	-11.3	-3.3	0.0
F - Construction	56,845	-6.4	-0.7	2.9	13,336	-15.3	-4.8	0.0
G - Wholesale and retail trade	82,533	-3.9	0.0	0.4	33,037	-9.2	-1.9	0.2
H - Transporting and storage	14,970	-21.9	-5.6	0.0	6,012	-21.6	-6.2	-0.2
I - Accommodation and food service activities	24,770	-2.4	0.0	1.4	6,614	-7.6	-1.0	0.6
J - Information and communication	13,014	-5.7	-0.3	0.5	9,044	-13.9	-3.2	0.0
K - Financial and insurance activities	2,502	-6.4	-0.3	0.1	1,393	-11.7	-2.7	0.0
L - Real estate activities	10,490	-4.0	0.0	1.1	3,435	-10.6	-1.3	0.7
M - Professional, scientific and technical activities	15,137	-5.9	-0.4	0.3	11,496	-13.9	-3.1	0.0
N - Administrative and support service activities	15,850	-6.9	-0.2	0.8	8,669	-16.6	-3.9	0.0
P - Education	1,935	-2.0	0.0	1.7	2,221	-17.4	-2.4	0.7
Q - Human health and social work activities	6,666	-3.3	0.0	0.5	4,584	-13.0	-2.5	0.0
R - Arts, entertainment and recreation	4,968	-5.3	0.0	1.6	2,381	-18.4	-3.8	0.4
S - Other services activities	5,851	-3.9	0.0	1.8	2,649	-13.3	-2.6	0.3

The analysis confirms that most of the differences between social security and company account/fiscal data are due to two effects: differences in underlying employment and items included or not in wages and salaries. The first element can be measured and must be corrected to estimate the desired SBS target variable. About the second one, it is instead uncertain if the inclusion of some items (i.e. travel allowances) should be corrected since the SBS is vague on the specific contents[14] and moreover they cannot be directly measured though the sources available.

[14] The general principle underlying the classification of costs suggest that labour costs should include only the remuneration of work done while other expenses for the production process should be classified in intermediate consumption. On the other hand the SBS explicit reference to company account items leaves space for interpretation since Italian practice allow to include in labour costs also items that could be more properly intermediate costs.

7. The correction: theory and practice

7.1 The theoretical frame

The need to edit the value of employees costs derived from the company accounts and the fiscal sources flows from the requisite of consistency within the whole theoretical accounting framework. It has to be kept in mind, in fact, that the final objective of the Frame is the representation of the whole accounting framework, not only the estimate of the labour cost. This implies that consistency have to be achieved not only with respect to theoretical definition and classification, but also between the estimate of all economic flows and balances and the underlying labour input. This is the condition to be able to determine the correct per capita value (in terms of costs and productivity) and to represent correctly also the distributive flows directed to the remuneration of the different types of labour input underlying production.

All types of labour input, in fact, contribute to production: employees and self-employed but also all kind of "external" employment (i.e. agency worker, outworkers etc). What change is how their cost (or their remuneration) is classified. The cost of "external" employment should be included in intermediate consumption, so that the value added (=production-intermediate consumption) is netted only of labour cost of external workers. The cost of employees is instead recorded in personnel costs and deducted from the value added to obtain the gross operating surplus. So this balancing item is computed when all internal and external employees are remunerated. However the gross operating surplus has yet to remunerate another labour input. Self-employed, in fact, contribute to the production with both labour and capital and have to be remunerated for both. In other terms, self-employed, who also owe the enterprises (or part of it) withdraw part of the enterprise profit to compensate both their labour and capital (also for the risk) input, thus receiving what is called a "mixed income" as part of the gross operating surplus, in the simplified accounting framework here represented.

It is therefore clear that economic flows have to be represented and classified in different cost items, according to the classification of the underlying labour input. Therefore the cost for external worker have to be included in intermediate consumption and not in personnel costs to correctly represent value added, while income flows remunerating self-employed do not have to be included neither in intermediate consumption, nor in personnel costs, since their remuneration have to be included in gross operating surplus[15].

This is the theoretical accounting framework, but when we deal with accounting and fiscal registers we have to consider, as shown in figure 4, that:

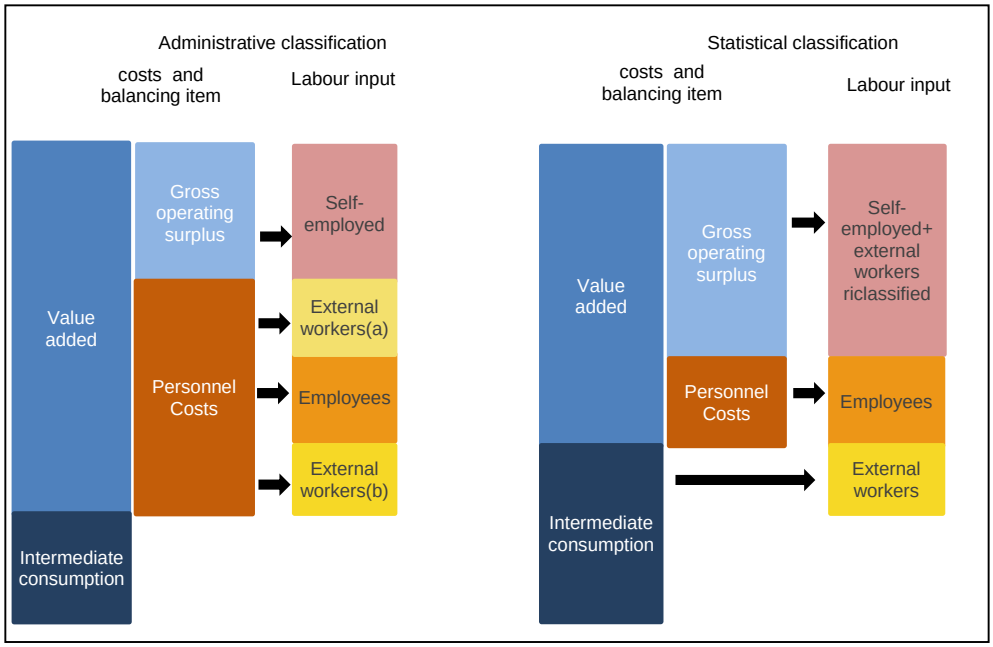
- In fiscal definition external workers receive a compensation that can be assimilated to employee costs (i.e. have the same characteristics in term of tax-deductibility): for this reason in fiscal forms and accounting practices, costs related to external workers (b in figure 4) can be recorded in the personnel costs.
- Fiscal practice may also induce many enterprises to classify the remuneration of self-employed as cost of external workers (a in figure 4).

This impose a re-classification of costs to be coherent both with our accounting

[15] From a National accounts point of view this flow is distributed in the secondary distribution of income account.

framework and with the correct representation of labour input as shown in the right hand side of figure 4.

Figure 4 - Classification of labout input costs: from administrative sources to statistical classification



Only costs related to proper employees have to remain in personnel costs. This imply that:

- costs related to external worker(b) have to be subtracted from labour costs and be classified as intermediate consumption
- costs related to external worker(a) have to be subtracted from labour costs determining a higher gross operating surplus that will remunerate these external workers reclassified as self–employed.

7.2 The correction method

The analysis sketched in paragraph 6 has demonstrated that the labour cost of some firms in accounting and fiscal data includes the costs of external workers and thus it can lead to overestimate the variable requested in the SBS Regulation. This definition/measurement error has to be corrected in order to avoid any bias in the final estimate of the labour cost statistical variable. A correction method has to tackle different points:

- identify which units must be corrected and which must not;
- calculate the size of correction;
- reallocate this correction between intermediate consumption and gross operating surplus.

As discussed previously, the new Employment Register provides a lot of information on the number of external workers, their wage compensation and social contributions for each type of external worker (see table 9).

Table 9 – Details available on external workers in the new Employment Register.

	n. of persons employed	wage compensation	social contributions
External workers reclassified as self-employed	X	X	X
Collaborators	X	X	X
Manager	X	X	X
Vouchers	X	X	X
Associates	X	X	X
Others	X	X	X
Temporary workers	X	*	*

**estimated by the RACLI register*

The only category of personnel assimilated to employees for which no information on costs are present are the temporary workers. For them, however, an estimate of wage and social contribution costs can be estimated by RACLI register. More in deep they have been imputed with the medium cost of blue and withe collars of the firms where temporary workers are employed.

Enterprises that have external workers and whose labour costs measured by the profit and loss are above those measured by RACLI are likely to be in the list of units to be corrected. This condition, however, if it is necessary is not sufficient, since the difference between the two sources may be due to other idiosyncratic definitional differences and not to the allocation of cost of non employees into the personnel costs. As we saw in paragraph 5 there are many of these differences in the two type of accounting. It has been decided that for the purpose of SBS statistics the other definitional differences do not need to be corrected , since the regulation definition, those that the frame aim to measure, are such that the company accounts definition can fit into them. In symbols:

$$w_i^{CA} = w_i^R + n_i + d_i + e_i$$

The equation states that the difference between the wages in company account of enterprise i, w_i^{CA}, and the wages in the social security based register, w_i^R, is equal to a difference, n_i, due to the inclusion of the wages of external workers and a difference, d_i, due to other systematic definitional differences and e_i, an i.i.d. error with zero mean due to random factors.

$$n_i \geq 0$$

It will be equal zero for firms that do not employ external workers or for firms that do not add their wages to those of the employees in the company accounts.

The method aims to obtain an estimate of n_i since it is the only element that should be eliminated by the correction method, to get to a corrected version of the wages of company accounts, $\widetilde{w}_i^{CA}$ equal to:

$$\widetilde{w}_i^{CA} = w_i^R + d_i + e_i$$

This correction is done in two steps: the first step obtain an estimate of d_i , $\widetilde{d_i}$, and, second, obtain an estimate of n_i , $\widetilde{n_i}$ given $\widetilde{d_i}$.

The estimate $\widetilde{d_i}$ is obtained by taking the median of the differences between wages in company accounts and wages in Racli for enterprises that have only employees aggregates in cells according to economic activity, size, legal form. Since by definition in such firms $n_i=0$ the average of differences is an estimate of the systematic definitional discrepancy between the two sources. The choice of taking the median instead of the mean is due to the presence of outliers.

The estimate $\widetilde{n_i}$ can thus be obtained using the information contained in the employment register and synthetized in table 9. What is not straightforward is choosing the costs of which types of workers have to be deducted from the company account figure. In fact different enterprises may have included costs of different kind workers along with those of true employees. Since, in principle, once taken into account the definitional difference, the wages in company accounts should be equal to the one in social security data plus a random error, the choice has been to estimate $\widetilde{n_i}$ as the sum of costs of those categories of workers that minimizes, for each firm, the difference between the wages of company accounts and the wages of RACLI plus the definitional difference. In symbols:

$$\widetilde{n_i} = \min(\text{abs}(\widetilde{w}_i^{CA} - (w_i^R + \widetilde{d_i})))$$

$$\text{Since } \widetilde{w}_i^{CA} = w_i^{CA} - \widetilde{n_i} = w_i^{CA} - \sum_{j\epsilon S_i} w_{ji}^E$$

Where w_{ji}^E is the wage of the *j-th* category of external worker (temporary woker, project worker..) employed at firm *i*. The sum $\sum_{j\epsilon S_i} w_{ji}^E$ thus represent the sum of wage of external workers over the set of categories S_i employed at firm *i*

$$\widetilde{n_i} = min_{S_i}(\text{abs}((w_i^{CA} + \widetilde{n_i}) - (w_i^R + \widetilde{d_i})))$$

Thus the minimum is over all possible sets S_i of categories of workers employed at firm *i*.

8. Final Remarks

The availability of multiple administrative sources on the same variables and the comparison among them let to better understand details and characteristics of the data, their administrative definition/aim, the way they are measured and differences among them. The reconciliation of administrative data and statistical purposes requires a deep knowledge of the sources to correct them or to measure the residual discrepancies. The comparison analysis on the TLC variable pointed out that none of the administrative data sources have exactly the information requested for the statistical aim. Each source has both pros and cons on different aspects due to the administrative purpose they are produced for. For the Frame and the satisfaction of the SBS Regulation, the choice for TLC variables was basically between two possibilities. The first was the use of the same sources (accounting and fiscal data) with the same priority as those supplying the other economic variables, hence disposing of a coherent accounting framework. The alternative was the use of social security data for labour costs with the advantage of having the same sources that is the base of the estimate of employment, thus keeping the maximum consistency with the labour input. Because of coherence among information drown from the same sources and the explicit reference of the SBS Regulation to balance items, the accounting/fiscal personnel cost variables have been included in the Frame but after correcting for wrong inclusions with the support of social security information. So far the evolution on the administrative

data available and used for statistical purposes let us to estimate and correct personnel cost data of Frame sources when not properly employee costs were included. Other aspects should be deepened and possibly corrected like worker on secondment and the remuneration in kinds, and some other may be enhanced like the correction method. Moreover, since the statistical definition of the same variable may be partially different in different regulations, as SBS and ESA, they must be taken into consideration to extend the use of administrative data to satisfy other regulations.

Riferimenti bibliografici

Baldi C., F. Ceccato, E. Cimino, M.C. Congia, S. Pacini, F. Rapiti e D. Tuzi. 2008. *Il controllo e la correzione in una indagine congiunturale su dati amministrativi. Il caso della rilevazione Oros*. Roma: Istat.

Curatolo S., V. De Giorgi, F. Oropallo, A. Puggioni, G. Siesto. 2016. "Quality analysis and harmonization problems in the context of the SBS frame". *Rivista di statistica ufficiale*. N.1/2016.

Garofalo G., I. Rocchetti e C. Viviano. 2012. *A revision of the Italian Business Register: a new methodological and conceptual "backbone" for a new informative system on employment.* Washington, D.C. 17–20 September 2012: 23rd Meeting of the Wiesbaden Group on Business Registers.

Grand E. e R. Quaranta. 2014. "La ricostruzione delle informazioni sugli oneri sociali obbligatori e sul costo del lavoro a partire dai dati individuali e di impresa di fonte Inps". *Politica economica*, n. 1, Il Mulino.

Regulation (EU) no 295/2008 of the European Parliament and of the council of 11 march 2008 Concerning structural business statistics, Official Journal of the Europe of 9.4.2008

Eurostat. 2013. European System of accounts 2010

Vekeman G. 2012. *Confronting various administrative data sources to estimate employment variables* paper presented at the International Conference on Quality 2012. Athens 2012, 29 May-1 Jun.

Casciano M.C., A. Cirianni, V. De Giorgi, T. Di Francescantonio, A. Mazzilli, O. Luzi, F. Oropallo, M. Rinaldi, E. Santi, G. Seri, G. Siesto. 2011. *Utilizzo delle fonti amministrative nella rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni*. Working Papers Istat, N.7/2011.

Estimation of the main variables of the economic account of small and medium enterprises based on administrative sources

Marco Di Zio[1] *Ugo Guarnera*[1] *Roberta Varriale*[1]

Abstract

The paper describes the imputation procedure of the main variables of small and medium-sized enterprise balance sheet. The procedure is used as part of the project aimed at creating an integrated system for the production of detailed estimates on enterprise economic performance. The variables are imputed using mainly administrative sources as Financial Statements, Studi di Settore and Tax return. The proposed procedure represents an integrated set of different imputation techniques: Predictive Mean Matching, nearest neighbor donor, and a two-step procedure for the treatment of variables characterised by a high presence of zeros. A first evaluation of the procedure is carried out by comparing the estimates based on administrative data with those obtained by the use of sampling weights.

Keywords: Imputation, predictive mean matching, nearest neighbor donor

Sommario

Il lavoro descrive la procedura di imputazione delle principali variabili del conto economico delle piccole e medie imprese. La procedura é stata utilizzata nell'ambito del progetto finalizzato alla realizzazione di un sistema integrato per la produzione di stime dettagliate sui risultati economici delle imprese. Le variabili vengono ricostruite utilizzando principalmente le fonti amministrative Bilanci delle Società di Capitale, Studi di Settore e modello Unico. La procedura proposta é un insieme integrato di diverse tecniche di imputazione: Predictive Mean Matching, donatore di minima distanza, ed una procedura a due passi per il trattamento delle variabili caratterizzate da una elevata presenza di zeri. Una prima valutazione della procedura é stata ottenuta confrontando le stime basate su dati amministrativi con quelle ottenute mediante l'utilizzo dei pesi campionari.

Parole Chiave: Imputazione, predictive mean matching, donatore di minima distanza

[1] Istat, Directorate for methodology and statistical process design. email: dizio@istat.it, guarnera@istat.it, varriale@istat.it. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat

1. INTRODUCTION

In Istat, Structural Business Statistics (SBS) for small and medium enterprises (SME) are traditionally based on sample surveys. In the last years, the increasing availability of information from administrative sources made it possible to take into account the possibility of using administrative data to improve the quality of the produced statistics. Until now this information has been generally used as auxiliary information to treat non-response in survey data and to calibrate the estimates on known aggregates.

The level of maturity in the analysis of these kinds of data lead Istat to use administrative data as a primary source for information to produce SBS statistics. In 2011 for the first time, data from administrative sources as *Financial Statement, Studi di settore, Tax Return* are used to build a microdata file composed of the main economic variables. The choice of producing a microdata file follows from the difficulty of providing coherent estimates at different level of aggregation, in this regard we remind that these data are also used by National Accounts to build national economic aggregates (Istat, 2014).

Since not all the variables are available in all the data sources, and the sources cover only subsets of the target population, the microdata file is a result of an imputation process. The imputation procedure is based on a combination of different techniques that are introduced to comply with requirements given by constraints, such as statistical relationships among main variables, balance edits, and presence of zero-inflated variables.

Given such a complexity, the assessment of the procedure is not an easy task. A comparison with official estimates based on the SME sample survey data is carried out. The differences are decomposed in terms of sampling and measurement errors. The analysis of the impact of the different error sources may be useful to validate the results and to improve the process of production of statistics in this context.

The paper is structured as follows. Section 2 describes the informative context of SME statistics based on administrative data. The imputation process is described in Section 3, and some results about the evaluation of the estimation procedure are reported and discussed in Section 4.

2. Informative context

The administrative data sources are the Financial Statements, Studi di settore, and Tax Return data. The units of Financial statements (FS) are the companies, mainly corporate firms, liable to fill in the financial statement. The *"Studi di settore"* (SDS) is a Fiscal Authority survey that aims at evaluating the capacity of enterprises to produce income and at indirectly assessing whether they pay taxes correctly. The units compiling the SDS form, composed of detailed information on costs and income, are the enterprises with a turnover less than 7,500,000 Euros belonging to many activity sectors. Tax return data are mainly based on the fiscal form *"Unico"* and, for a residual part of units representing corporate firms, on *"Irap"* (the Italian regional tax on productive activities).

All the analyses described in the paper have, as a starting point, the quality assessment of the administrative data carried out by the subject matter experts (Curatolo *et al.*, 2015). Although in principle many variables observed in the administrative data sources could be used for the SBS estimates, only some of them can be considered enough reliable both in terms of consistency of definitions with the ones described by the SBS regulation, and in terms of reported values compared to the SME observations. The list of the variables used in the imputation process is reported in Table 1.

Table 1 - Variables used in the imputation process

Section	Label	Variable
Revenues	Y_1	Income from sales and services (Turnover)
	Y_2	Changes in stock of finished and semi-finished products
	Y_3	Changes in contract work in progress
	Y_4	Changes in internal work capitalized under fixed assets
	Y_5	Other income and earnings (neither financial, nor extraordinary)
Costs	Y_6	Purchases of goods
	Y_7	Purchases of services
	Y_8	Use of third party assets
	Y_9	Changes in stocks of raw materials and for resale
	Y_{10}	Other operating charges
	PC	Personnel Costs

It is worthwhile to remark that the variable *personnel costs* is always observed and it is used as auxiliary variable in the imputation procedure. In addition to PC, some derived variables are used as auxiliary variables in the imputation process. In fact in some cases they are considered more reliable than the variables used for the derivation, this is due to a kind of compensation process that is not easy to model. The derived variables, related to the Cost section, are listed in Table 2.

Table 2 - Derived variables

Derived Variable	Transformation	Variable description
CS	$Y_2 - Y_9$	Total Change in Stock
GS	$Y_6 + Y_7$	Purchases of Goods and Services
IC	$GS + Y_8 + Y_{10}$	Total Intermediate Costs

We remark that some variables are observed in more than one data sources, this means that for each of them (generally) different values are available. In this application a hierarchical approach is chosen. It consists in assigning a hierarchy to the administrative sources and consequently values of the variables are chosen according

to this raking. The hierarchy has been established by subject matter experts according to some quality criteria such as coverage of administrative sources with respect to the business register, steadiness of the supply in terms of timing and variable content (Curatolo *et al.*, 2015), and it assigns the first rank to FS, then to SDS, and finally to Tax Return data. Based on the assumed hierarchy, the coverage of the administrative sources for the 2011 is reported in Table 3.

Table 3 - Coverage of administrative sources for the 2011

Source	Frequency	Relative frequency (%)
FS	714885	16.1
SDS	2836100	64.0
Unico	714894	16.1
Irap	4201	0.1
NA	162848	3.7
Total	4432928	100.0

The subset of population not covered is small and it is composed of the smaller units in terms of size. In our procedure we also need to take into account that for all the units the number of employees is known from the business register ASIA and that for most of the units in ASIA there is an important information related to turnover (i.e.,'amount of business') and it is a good proxy for the turnover mainly coming from the VAT declarations.

It is however worthwhile to remark that the coverage of each single variable depends on its availability in the different data sources.

In Table 4 the pattern of missing data per variables and data sources is illustrated. The symbols 'X' and '?' stand for observed and missing data respectively. In this table, SDS-F, SDS-G and Unico1-Unico8 refer to the different kinds of SDS and Unico that enterprises have to fill in depending on their legal status. In particular, the units compiling SDS-G and Unico5 are represented by professionals and "minimum taxpayers", respectively.

Table 4 - Pattern of missing data per variables and data sources

Source	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	PC	CS	GS	IC	Coverage rate (%)
FS	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16.13
SDS-F	X	?	X	X	X	X	X	X	?	X	X	X	X	X	50.08
SDS-G	X	X	X	X	X	?	?	?	X	X	X	X	?	X	13.90
Unico1	X	X	X	X	X	?	?	?	X	?	X	X	X	?	0.78
Unico2	X	X	X	X	X	?	?	X	X	?	X	X	X	?	0.04
Unico3	X	?	X	X	X	?	?	?	?	?	X	X	X	?	2.73
Unico4	X	?	X	X	X	X	X	?	?	?	X	X	X	?	0.76
Unico5	X	X	X	X	X	?	?	?	X	?	X	X	?	X	10.86
Unico6	X	?	?	?	?	?	?	?	?	?	X	?	?	?	0.16
Unico7	X	?	?	?	?	?	?	?	?	?	X	?	?	?	0.31
Unico8	X	?	?	?	?	?	?	?	?	?	X	?	?	?	0.49
Irap	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0.09
NA	?	?	?	?	?	?	?	?	?	?	X	?	?	?	3.67

The rate of missing data per variable, taking also into account the information available in the Business Register (ASIA), is reported in Table 5. We notice that the minimun amount of missing data is related to the variable Y_1 (turnover) and it is lower than the minimum of Table 3. This is because in the Business Register there is a variable closely related to the turnover (named *amount of business*) that in some cases can be used to predict Y_1. On the other hand, a very high rate of missing data (approximately 58%) affects variables Y_2 and Y_9, associated with the two components of the change in stock. It is worthwile to mention that however, for a large set of units, the difference $CS = Y_2 - Y_9$ is known even though the separate components are not observed. This mitigates the impact of the missing data on the estimates of some crucial derived variables such as the *value added*, where only the total change in stock is relevant.

The economic data described in this paragraph are used both by the SBS sector and by National Accounts, requiring many domains of estimation. In order to avoid consistency problems, missing data are imputed to obtain a microdata file.

3. The imputation process

The imputation procedure is based on a combination of different techniques.
The entire imputation process is composed by 4 sequential steps:

1. deterministic imputation based on the guidelines of subject matter experts;
2. imputation of the variables Y_1, Y_6, Y_7 and CS, through Predictive Mean Matching (PMM);
3. imputation of the variables Y_3, Y_4, Y_8, Y_{10} and Y_5, through Nearest Neighbor Donor (NND);
4. imputation of the variables Y_9, Y_2 through a two-step procedure composed by a logistic and a linear regression model.

In this paper we focus on the description and evaluation of the imputation process related to the last three steps.

Table 5 - Rate of missing data per variable

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	CS	GS	IC
3.7	58.2	4.7	4.7	4.7	19.2	19.2	19.8	58.3	19.8	6.7	15.6	15.6

The pattern of missing data depicted in Table 4 is the one obtained after the deterministic imputation in step 1.

The steps from 2 to 4 have been carried out inside strata based on the economic divisions Nace2 cross-classified with the two subsets of observations characterised by having or not personnel costs. For the PMM, also the item GS has been used to define strata, the distinction is made between units either with or without purchases of goods and services.

The enterprises with information coming from SDS-G and Unico5 are represented by professionals and "minimum taxpayers", and they are considered to behave

quite differently from the rest. Since the imputation process tends to reproduce in the non-observed part of the population the behaviour of the observed units, for this subset of population the imputation is made by resorting to the SME survey, details will be given later in the paper.

The choice of each imputation method for different groups of variables is due to: the percentage of missing values, the variable distribution characteristics (only positive, zero-inflated, etc.), the presence of a (weak/strong) relationship between variables and the presence of balance edits. All these characteristics influence the choice of a statistical model in the imputation process.

3.1 Methods

The PMM can be considered as a NND imputation technique based on a distance function where matching variables are weighted through their predictive power with respect to the variables that have to be imputed. In a multivariate context, the PMM is typically applied to match each recipient to the donor having the closest predictive mean with respect to a regression model of the target variables on a set of covariates. Selection of donors is based on the Mahalanobis distance defined in terms of the residual covariance matrix from the regression model. Intuitively, Mahalanobis matrix gives largest weights to the variables with the smallest prediction error. More in detail, when the variables are continuous and in presence of arbitrary patterns of missing items, a typical application of the PMM is the following (Little, 1988).

1. The parameters of a multivariate Gaussian distribution are estimated through the EM algorithm (Dempster et al., 1977) using all the available data (complete and incomplete).
2. Based on the estimates from EM, for each incomplete unit (recipient), predictions of the missing items conditional on the observed ones are computed. The same predictive means (i.e., corresponding to the same missing pattern) are computed for all the complete observations (donors).
3. Each recipient is matched to the donor having the closest predictive mean with respect to the Mahalanobis distance defined through the residual covariance matrix from the regression of the missing items on the observed ones.
4. Missing items are imputed in each recipient by transferring the corresponding values from its closest donor.

The NND method is a common hot deck method, in which a donor is selected from the complete cases in order to minimize some similarity measure, such as the Euclidean distance. In this application, the matching variables used to compute the Euclidean distance are Y_1, Y_6, Y_7 and PC (if present), and the variables to be imputed are the ratios of Y_3, Y_4, Y_5, Y_8 and Y_{10} to Y_1. The final imputed value is obtained by multiplying the imputed ratios by the size variable Y_1 of the recipient unit, this technique is also known as *ratio hot-deck* (de Waal et al., 2011). This method is preferable to the classical one that imputes directly the value observed in the closest donor, because it ensures that the values of the variables to be imputed are coherent with respect to the value of the reference variable. The reason why Y_1 has been treated

as a size variable, instead of the commonly used *Number of employees*, is that it has both the lowest rate of missing data and the highest quality from a content point of view. When Y_1 is zero the ratio cannot be computed, in this case the standard NND is used.

In this context, both the PMM and NND approaches have the advantage to recover live values from donors. Since the PMM technique relies on a multivariate normal model, it has been used to treat variables having a genuine continuous distribution. On the contrary, the NND method has been used to treat variables with distribution characterized by 0 inflation and a non-linear relation.

Finally, the imputation of Y_2 and Y_9, representing the two components of CS has been carried out through a two-step process, composed by a logistic and a linear regression model. In the first step, we applied a stepwise logistic regression using as covariates Y_1, Y_3, Y_6, Y_7, CS, a modified version of the economic divisions Nace2 and PC (if present) in order to assign each enterprise to one of the 3 subpopulations characterized by the presence or absence of the two components (yes/yes, yes/no, no/yes). The assignment is based on random drawing from a multinomial distribution with parameters corresponding to the probabilities estimated through the logistic model. In the second step, for the enterprises which have been assigned to the subpopulation with only one component, the total value of CS has been imputed to such component. For the other enterprises, we estimated the value of the two components through a linear regression model with the same covariates used in the logistic model. This approach has been compared with a NND approach through a simulation study, resulting in a better efficiency (both in terms of time consuming and accuracy of the estimates) of the two-step approach. The difficulty in the imputation of these variables is both in their nature and in the nature of the total CS that is semi-continuous and not positive. This means that the value CS could be generated by any linear combination of the two components. As an hypothesis, when the variable CS is equal to 0, the two components have been imputed to be equal to 0.

For enterprises with information coming from SDS-G and Unico5, all the information on revenues is complete. Costs are imputed through random ratio hot-deck within suitable defined imputation cells. The donors are chosen from the SME survey, this is the same as drawing a vector of ratios from the estimated distribution of the ratios in SME. In particular, we imputed the composition of the costs using as size variable the total costs and transferring the compositional information from the survey data.

4. Evaluation

The complexity of the imputation procedure and the particular nature of administrative data make the evaluation of the accuracy of the estimates a difficult task. A first overall evaluation has been obtained by comparing the estimates based on administrative data with the ones resulting from the classical procedure obtained by means of the SME sample survey data. The comparison has been made by using two different sets of estimation domains: the first is Nace2 (aggregation of economic sectors), and the second corresponds to the different administrative sources where information is

taken from. While in the former case the domains are aggregations of planned survey domains - thus they are composed of sampling strata - in the latter case the domains - that have not been planned in the survey design phase - are used to analyze possible different levels of discrepancies between administrative data and survey data across the available sources.

For each typology of domain and each analysed variable, relative differences between total estimates based on administrative and sample data are considered. In detail, for a given variable Y with corresponding population total T_y, we have computed the indicator:

$$d_y^t = \frac{\hat{T}_y^s - \hat{T}_y^{ad}}{\hat{T}_y^{ad}} \times 100,$$

where $\hat{T}_y^s$ is the estimate of T_y obtained with the sample data through the calibration estimator currently used for SME survey, and $\hat{T}_y^{ad}$ is the estimate computed on the entire archive by summing up all the values. In order to distinguish the source of discrepancies due to the sampling and the measurement error, we have also considered, for each domain, the additional estimate $\hat{T}_y^{ad,s}$, that results from using the SME survey estimator on the sampled units, with the replacement of the survey data with the administrative data. As approximate measures of the measurement effect and sample error respectively, we introduce the following two indicators:

$$d_y^m = \frac{\hat{T}_y^s - \hat{T}_y^{ad,s}}{\hat{T}_y^{ad}} \times 100, \qquad d_y^s = \frac{\hat{T}_y^{ad,s} - \hat{T}_y^{ad}}{\hat{T}_y^{ad}} \times 100.$$

Thus, the total difference is decomposed into the sum of two differences associated with the two mechanisms:

$$d_y^t = d_y^m + d_y^s. \tag{1}$$

Note that the indicator d_y^m that evaluates the "measurement effect", being based on the comparison of different measures only on the sample units, is also affected by sampling error. In particular, a few gross errors may have an high impact on the indicator.

Table 6 reports the indicators d^t and d^m for the three variables Y_1, IC, and the Value Added computed as $VA = \sum_{i=1}^{5} Y_i - IC - Y_9$ by source. In Table 7 results are shown for the following economic divisions Nace2: *Manufacture of textiles* (Nace2 =13), *Construction of buildings* (Nace2=41), *Wholesale and retail trade and repair of motor vehicles and motorcycles* (Nace2=45) and *Architectural and engineering activities; technical testing and analysis* (Nace2=71). In the tables, the size of domains in the population N and in the sample n are also reported.

Results in Tables 6 and 7 show that the largest component in the decomposition (1) is the one associated with the sampling error. This result is encouraging because it implies that the transition from designed based inference to an estimation approach based on administrative sources would result in a significant improvement of the estimation accuracy.

Table 6 - Discrepancies between sample estimates and estimates based on administrative data for different administrative data sources: total differences (d^t) and measurement component (d^m)

			d^t			d^m		
Source	N	n	Y_1	TC	VA	Y_1	TC	VA
Tot	4432928	74112	-6.5	-8.8	0.2	-0.9	-0.5	-0.9
FS	714885	34284	-2.6	-4.6	3.1	-0.9	-0.7	-2.3
SDS-F	2220050	31732	11.4	11.3	12.7	-0.5	-1.2	1.6
SDS-G	616050	3844	23.1	26.5	26.7	-0.2	13.5	-0.3
Unico1	34570	296	-38.9	-69	-26.7	-1	-3.3	0.3
Unico2	1746	32	-41	-40.6	-40	-1.7	-2.5	0.9
Unico3	120876	756	-70.7	-78.5	-55.7	-0.7	-5.3	8.4
Unico4	33676	356	-69	-81.1	-34.6	-0.5	-5	14.6
Unico5	481517	1434	-58.2	-51.5	-59.1	-1.4	3.3	-1.3
Unico6	7371	151	-70.1	-76.9	-59.5	0.1	-2.5	-6.5
Unico7	13553	361	-68.5	-68.5	-59.4	-0.4	-2.5	13.8
Unico8	21585	381	-63.7	-55.3	-58.2	-2.4	8.7	-1.4
Irap	4201	89	-55.8	-63.1	-61	-0.1	-0.9	4.1
NA	162848	396	-91	-90.9	-89.8	-2.5	-1.3	-4.6

Table 7 - Discrepancies between sample estimates and estimates based on administrative data for some economic divisions (Nace2): total differences (d^t) and measurement component (d^m)

			d^t			d^m		
Nace2	N	n	Y_1	TC	VA	Y_1	TC	VA
13	15669	1275	8.4	10.1	3.8	0	1.1	-3.5
41	150417	2625	-18	-21.7	-9.2	-0.8	2	1.5
45	118985	2649	0.4	0.3	1.7	1	0.9	-0.7
71	212880	1009	-9.5	-15.3	-4.6	-2.6	1.5	-0.5

An important issue in the evaluation of an estimation procedure is the assessment of the estimate accuracy. According to the estimation approach so far used in Istat, the SBS estimates for SME are based on a sample survey, hence the assessment of their accuracy relies on designed-based inference. As already mentioned, massive use of administrative information requires a change of paradigm. In fact, differently from the context of sample survey, the availability of administrative information is not under control of the researcher, so that some model assumptions are necessary. In particular, one has to think of data as *iid* realizations from a statistical (possibly not explicitly specified) model. This is generally referred to as super-population model. In this framework, the inferential approach is predictive, i.e., the missing values are imputed (predicted) on the basis of the available information. Thus, the uncertainty of the resulting estimates are essentially due to the prediction error associated with the imputation procedure.

Some limitations of the present evaluation methods should be mentioned. First, comparison is performed at one point in time, so that results should be assessed in future occasions. Second, evaluation of measurement component of the total error is based on survey data and thus it is affected by sampling error. It is interesting to note that, because of compensations, substantial differences in the estimates of Turnover

(Y_1) and Total Costs (TC) do not result in significant discrepancy for the variable Value Added (VA).

If predictions were based on some parametric regression model and the missing patterns were enough simple, standard analytic techniques could be used to evaluate the estimate of the estimator variance (Valliant, 2000). In cases where missing patterns are arbitrary, but imputations are obtained from a unique multivariate normal, the Rubin multiple imputation approach can be (relatively) simply applied to assess the precision of the estimate of any finite population quantity (Rubin, 1987). In the present case, however, the imputation procedure is complex and is composed of many different techniques. This complexity makes it difficult to use standard procedures for the assessment of the uncertainty in the final output. In particular, the assumed super-population model is not explicitly specified and it is only implicitly defined through the imputation procedures that have been used. Because of this characteristic, a replication approach seems to be more appropriate than an analytic approach. However, common univariate techniques for the variance estimation such as jacknife and bootstrap (Wolter, 2007) are difficult to extend to our context and further research is needed.

References

Curatolo S., V. De Giorgi, F. Oropallo, A. Puggioni, G. Siesto. 2016. "Quality analysis and harmonization issues in the context of the SBS frame", *Rivista Statistica Ufficiale*, N.1/2016.

De Waal T., J. Pannekoek, S. Scholtus. 2011 *Handbook of Statistical Data Editing and Imputation* Wiley

Dempster A.P., N.M. Laird, D.B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, B 39; 1-38.

Istat. 2014. I nuovi conti nazionali in SEC 2010. "Innovazioni e ricostruzione delle serie storiche (1995-2013)". *Nota informativa Istat*, Ottobre 2014.

Little R.J.A. 1988. "Missing-Data Adjustments in Large Surveys". *Journal of Business & Economic Statistics*, 6, 3; 287-296.

Rubin D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*, Wiley

Valliant R., A.H. Dorfman, R.M. Royal. 2000. *Finite Population Sampling and Inference: A Prediction Approach*, Wiley

Wolter K.M. 2007. *Introduction to Variance Estimation*, Springer

Estimation procedure and inference for component totals of the economic aggregates in the "Frame SBS"[1]

Paolo Righi[2]

Abstract

Recently the Italian National Institute of Statistics - Istat - implemented a Business frame where several variables are collected from administrative registers. Nevertheless, these variables do not cover all statistical interests and some variables are collected only by the Small and Medium Enterprise survey – SME survey. The paper deals with the estimation of totals of variables strictly observed in Istat SME survey and proposes an estimation procedure, based on the projection estimator, exploiting the variables of the Business frame and coherent with respect to the totals of the variables in the frame. The result is an integrated output in the Business frame and a flexible tool useful for other statistical purposes. Inferential properties are shown theoretically and empirically and conditions to obtain unbiased estimates are pointed out.

Keywords: Administrative data sources, projection estimator, design based inference.

1. Introduction

Most of the new Istat Business *frame* variables (Luzi e Monducci, 2016) come from the several Italian administrative data sources. They cover only partially the business economic information demand. Nevertheless, other variables and the respective parameters such as totals or means are required by EU Structural Business Statistics (SBS) Regulation. In particular, they are fundamental for implementing econometric models analyzing trend and the performance of the economic system. Usually such variables represent the *components* of economic aggregates which are known from the archives or are imputed in the frame by previous steps (see Di Zio *et al*., 2015).

The only direct informative source of these components is essentially the Small and Medium Enterprises (SME) sampling survey conducted by Istat. In the SME survey the calibration estimator (Deville and Särndal, 1992) is used. However, the large amount of auxiliary information, now available, is inefficiently exploited by this estimator. Furthermore, the output of the estimator is not suitable for the frame purposes. In the paper we propose an estimation process, exploiting the auxiliary variables in an enhanced way, for the totals of these components. The new estimator takes into account some appealing requirements: the sum of the estimated totals for the elementary components belonging to a given economic aggregate must be coherent to the (estimated) total of the economic

[1] The views expressed in this paper are solely those of the author and do not involve the responsibility of Istat.

[2] Istat, e-mail: parighi@istat.it

aggregate at domain level according to the current SBS Regulations; the output should be a flexible statistical tool and it can be used for other aims. In particular, the Istat National Account (NA) sector bases its procedures on the frame, so the coherence of the estimates of the components must be fulfilled for the NA domains that are generally highly detailed.

To obtain these objectives, the *projection estimator* (Kim and Rao, 2012) has been used. The method imputes or *projects* the component values of the not sampled enterprises in the SME survey by using an estimated regression models. The final estimates are achieved as the sum of the projected values by the models.

There are some advantages in using the projection estimator. At first glance the micro level estimates (projected values) seem the most appealing feature of the estimator. Nevertheless, such feature has to be used carefully because it hides a dangerous drawback of producing biased estimates at certain level of detail (section 2.1). The most relevant properties involve the inferential process. The projection estimator takes into account the randomization process of the SME sampling design and the inference is performed by a model assisted approach. That greatly simplifies the computation of the precision of the estimates, especially when a large scale population (about 4.4 millions of enterprises) has to be investigated. Model assisted approach is, commonly, used in the national statistical office and the variance estimation of the projection estimator can be found in classical textbooks. Moreover, the approach guarantees unbiased and robust estimates at least at certain domain level (see section 2.1) without an overwhelming model diagnostic required when a model based approach is taken into account.

Finally, the projection estimator is a more flexible tool compared to the generalized regression estimator (Särndal *et al.*, 1992), approximating the calibration estimators. The regression (and calibration) estimator considers a unique set of covariates in the regression model; the projection estimator varies the set of covariates in the regression model when the variables of interest change. That means each component is projected by a specific statistical model and that allows the improving precision of the estimates.

These conditions justify the choice to identify the projection estimator as a tool to complete the Business frame.

The outline of the paper is as follows: section 2 is devoted to the description of the projection estimator, highlighting the theoretical aspects and the bias issue. Section 3 describes the practical implementation of the estimator. Since the SME sample is affected by unit non response, the weight adjustment process for unit nonresponse is shown. The projection estimator has been implemented using the adjusted sampling weight. Section 4 gives an approximate estimate of the sampling errors. Section 5 presents brief conclusions.

2. Projection estimators

The projection estimator was introduced long ago in the sampling literature, but recently has had considerable attention (Hidiroglou, 2001; Merkourios, 2004; 2010) and the paper by Kim and Rao (2012) well formalizes the fundamental properties. Schenker and Raghunathan (2007) reported several applications using a model-based inference. Unlike Kim and Rao proposed a model assisted framework that is robust against failure of the working model used to generate the synthetic or projected values.

The estimator arises to deal with a nonnested two-phase sampling design. This design

involves two independent surveys from the same target population U consisting of N elements. A large sample s_1 from survey 1 collects information only on a vector of variable $\mathbf{x} = (x_1,...,x_q,...x_Q)'$ and a much smaller sample s_2 from survey 2 provides information on both y and $\mathbf{x}$, being y the variable of interest. It is assumed that the observed variables $\mathbf{x}$ are comparable. The concept of comparability refers to the classical test theory in which two kinds of measurement errors are distinguished (Bakker, 2012): validity and reliability. According to McCall (2001), reliability refers to whether the measurement procedures assign the same value to a characteristic each time it is measured under essentially the same circumstances. Unreliable measurement leads to random error. Validity refers to how accurately the values assigned in the measurement procedures reflect the actual conceptual variable measured. Invalid measurement leads to systematic error or bias in estimates (McCall, 2001). In the following we make the approximation of the absence of two kinds of measurement errors.

The main aim of the estimator is in creating a single synthetic dataset of proxy values $\tilde{y}_k$ (k=1, …, N) for the unobserved y_k values in survey 1 and then using the proxy data together with the associated survey weights, w_{1k} of survey 1 to produce projection estimates of the population and domain (or subpopulation) totals of y. Since the estimator creates an imputed dataset associated with the sample s_1, the method is classified as mass imputation technique too.

We focus on the estimator of the totals $Y_d = \sum_{k \in U} y_k \delta_k(d)$, where $\delta_k(d)$ is the domain membership indicator variable. The total for the overall population is a specific case obtained setting $\delta_k(d)$=1 always. The sub-population total is obtained setting $\delta_k(d)$=1 if unit $k \in U_d$ and $\delta_k(d)$=0 otherwise being U_d the dth domain (d=1, …, D).

Projection estimator is assisted by a superpopulation working model. Let a general formulation of the working model be $E(y_k \mid \mathbf{x}_k) = f(\mathbf{x}_k, \boldsymbol{\beta}) = \mu_k$, with $Var(y_k \mid \mathbf{x}_k) = \sigma^2 a(\mu_k)$ for some known function $a(\mu_k)$ and that $Cov(y_k, y_j \mid \mathbf{x}_k, \mathbf{x}_j) = 0$ for $k \neq j$. For a continuous variable y as the case of the components to be projected in the frame the linear model is a suitable choice. The working model is fitted by relating y to $\mathbf{x}$ using the data $\{(y_k, \mathbf{x}_k) : k \in s_2\}$ and $\tilde{y}_k = f(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is obtained as a solution to

$$\sum_{k \in s_2} w_{2k} [(\partial \mu_k / \partial \boldsymbol{\beta}) / a(\mu_k)](y_k - \mu_k) = 0 ,$$

according to the estimation function theory (Godambe and Thompson, 1986). We point out that the ordinary and weighted least square methods for linear model belong to this class of parameter estimators.

Finally, the projection estimator at domain level is given by

$$\hat{Y}_{d,p} = \sum_{k \in s_1} \tilde{y}_k \delta_k(d) w_{1k} . \qquad (2.1)$$

In our estimation context we assume the SME survey as the second survey while the

first large survey is the business register covering the entire population, being w_{1k}=1. In the SME survey and in the business frame the $\mathbf{x}$ variables are the economic aggregates and some other auxiliary variables such as number of employed persons and economic activity. The $\mathbf{x}$ variables of both the data sources are comparable being the systematic errors and the unreliability reduced (Luzi e Monducci, 2016). The estimator (2.1) becomes

$$\hat{Y}_{d,p} = \sum_U \tilde{y}_k \delta_k(d) \qquad (2.2)$$

As far the overall population total is concerned the projection estimator is given by $\hat{Y}_p = \sum_U \tilde{y}_k$.

2.1 Bias and variance

The working model introduced in section 2 is domain independent. However, there are some advantages to consider the domain if we want produce unbiased estimates.

Usually the estimator (2.2) produces biased estimates being an approximate expression of the bias given by

$$B(\hat{Y}_{d,p}) \cong \sum_{k \in U} \delta_k(d)[y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0)], \qquad (2.3)$$

in which $\boldsymbol{\beta}_0$ is the probability limit of $\hat{\boldsymbol{\beta}}$ with respect to the second sampling design. An estimate of the domain bias is given by $\hat{Y}_{d,bc} = \sum_{k \in s_2} \delta_k(d) w_{2k}(y_k - \tilde{y}_k)$, so a bias-corrected version domain estimator is

$$\hat{Y}_{d,p,bc} = \hat{Y}_{d,p} + \hat{Y}_{d,bc} . \qquad (2.4)$$

Unlike the projection estimator (2.2) the bias corrected estimator requires the use of the data and of the survey weights from the second survey and the issue could be unattractive.

Nevertheless, there are some conditions in which (a) $\hat{Y}_{d,bc}$=0 or (b) the bias of the estimator (2.3) is asymptotically negligible with respect to the domain total Y_d .

The condition $\hat{Y}_{d,bc}$=0 is achieved when the $\mathbf{x}_k$ vector include the $\delta_k(d)$ value. That means the domain intercept has to be included in the linear model underling the projection estimator. When a heteroscedastic linear model is used, with $Var(y_k \mid \mathbf{x}_k) = \sigma^2 x_{qk}$ then the variable $\delta_k(d) x_{qk}$ must be included in the regression line in order to obtain $\hat{Y}_{d,bc}$=0.

It is worthwhile to note that the bias-corrected estimator has internal consistency property: if condition (a) is fulfilled for a domain U_d the estimates when summed over the sub-domains defining a partition of U_d agrees with $\hat{Y}_{d,p}$.

As far condition (b) is concerned the asymptotic bias of the projection domain estimator relative to the domain total is given by

$$RB(\hat{Y}_{d,p}) = -\frac{N\,Cov(\delta_k(d), r_k)}{Y_d}, \qquad (2.5)$$

where $Cov(\delta_k(d), r_k)$ is the population covariance of $\delta_k(d)$ and $r_k = y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0)$. It follows from (2.5) that $RB(\hat{Y}_{d,p})$ is negligible if $\delta_k(d)$ is approximately unrelated to r_k. Roughly speaking, such condition is verified when in the scatter plot of the *d*th domain points is fairly distributed over and under the regression line. The expression (2.5) is equivalent to the formula (13) proposed by Kim and Rao (see Appendix 1). The (2.5) highlights that for large Y_d the relative bias becomes relatively small.

As far the variance is concerned, in the standard nonnested two-phase sampling design estimator (2.1) an approximate expression, when the condition (a) or (b) holds, is

$$Var(\hat{Y}_{d,p}) \cong Var_1[\sum_{k \in s_1} \delta_k(d) w_{1k} f(\mathbf{x}_k, \boldsymbol{\beta}_0)] + Var_2[\sum_{k \in s_2} \delta_k(d) w_{2k} (y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0))], \qquad (2.6)$$

where $Var_1[.]$ and $Var_2[.]$ are the design variances respectively of the first and the second sampling design. In our survey context is quite dissimilar from the usual one if we treats the first survey as census. In this case $Var_1[.]$ disappears and the variance of the projection estimator becomes

$$Var(\hat{Y}_{d,p}) \cong Var_2[\sum_{k \in s_2} \delta_k(d) w_{2k} (y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0))], \qquad (2.7)$$

which is the standard formula used for the generalized regression estimator. By the consequence we may use the standard variance estimator of the generalized regression estimator (Särndal *et al.*, 1992). The assumption is that the economic aggregate **x** values are really observed. Ignoring the imputation process implemented for some units (Di Zio *et al.*, 2015) the expression (2.7) is a downward variance approximation. The goodness of the subsequent inference will depend on the performances of the imputation step and the rate of the imputed values.

Introducing the imputation uncertainty the variance expression becomes more complex (Appendix 2) and it is not dealt with in the application.

3. Estimates of the economic component totals: application of the projection estimator

The procedure is based on the sample of respondents of the SME survey. Bias conditions and variance of the projection estimator have been taken into account for setting the regression models. There is a trade-off between bias and precision; models defined including the domain intercept at highly detailed domain level allows to compute unbiased detailed estimates, but variance estimation could increase. So we cannot use too specific regression models. The estimation procedure considered the coverage of about 33,600 respondents of the whole population (SME survey, year 2011). The analysis led to consider

models for domains defined according to the Nace Rev. 2 three digit economic activity by the size class of employed persons (0-5, 6-19, 20-99).

We started with about 600 domains (and regression models) with generally a minimum number of sampled units equal to 25 and an average number of about 45 of sampled units. In some cases we obtained smaller sample size domains but with a high sample rate (Table 3.1). Finally, it was necessary to collapse some economic activities / or classes of employed persons to gather an enough number of respondents for estimating the models, obtaining 583 domains.

Table 3.1 - Rules for not collapsing the domain

(If) Number of respondents	(than) Sample rate (respondents/population size) must be
1-2	1.00
4-5	0.50
6-8	0.10
9-14	0.02

Under this level of detail the estimates could be significantly biased according to the condition (2.5) and a simple solution it should be use the estimator (2.4), guaranteeing the internal consistency although the variance problem still remains. Otherwise, the small area estimation approach (Rao, 2003) could be used for more reliable estimates, but the procedure could be complex if internal consistency must be satisfied.

SME survey is affected by unit nonresponse. So we used the Response Homogeneity Group technique (Oh and Scheuren, 1983) to adjust the sampling weights for unit nonresponse. The sampling weights are inflated by the inverse of non-response rate measured at RHG level, being the sample size of 2011 SME survey of about 97,000 enterprises. After studying the best way to deal with the nonresponse the RHGs coincided with the domains of regression models. In particular, the logistics model and different nonresponse classes for nonresponse adjustment has been compared. The results have been not significantly different from the ones using the domains of the projection estimator. On the other hand implementing the logistic model for estimating the response probability can be cumbersome if the process has to be performed in each survey occasion.

The regression models assisting the *projection estimator* have been defined taking into account the space of the possible projection values. For all the components the constraint is to obtain non negative projected values but the components of the change economic aggregates. So, for the former type of variables the heteroscedastic ratio models have been used, where each component has as covariate the economic aggregate to which it belongs to. We point out that with this model the sum of the components of a given aggregate is equal to economic aggregate at enterprises level. The regression model for the components of the change economic aggregate uses standard heteroscedastic model where the heteroscedastic term is the square root of the number of employees.

4. Sampling errors of the projection estimator: some evidences

The efficiency of the projection estimator has been compared with the one of the

calibration estimator currently used in the SME survey. The analysis of the results has to envision two issues: (i) the variance of the projection estimator (computed on 33,600 respondents) does not take into account the previous imputation step on the economic aggregates; (ii) the variances of the calibration estimator are computed on the respondents and the integrated non respondents of the SME survey (73,200 enterprises) according to the procedure described by Casciano *et al.*, (2012). We point out that the component values of the integrated non respondents are imputed but the estimator treats them as if they were observed and the accuracy of the estimates will be generally overstated (Kalton and Kasprzyk, 1986; Righi *et al*., 2014). We remark a fundamental different role of the imputed values in the two sampling contexts. In the projection estimator, the imputed variables are the auxiliary variables, while in the current estimation strategy they are the interest variables. That means: the true projection estimator variance will be larger than the variance measured in the analysis; bias is introduced in the calibration estimator and the true mean square error will be larger than the one observed in the analysis.

Section 2.1 introduces the complexity for tackling the point (i). To deal with the point (ii) it should be necessary to known the imputation procedure, making the comparison too burdensome. Therefore, we consider the results as general evidences of the two estimator performances, underling when the imputation step affects the final evaluation.

Figures 4.1, 4.2 and 4.3 depict the Coefficient of Variations (CVs) of the projection and the calibration estimators for the totals for the entire target population. For sake of brevity, only some of the most important component variables are shown: *income from sales and services* (*turnover*), *purchases of services*, *purchases of goods*, *use of third party assets* and *other operating charges*.

Generally, the projection estimator outperforms the current estimation procedure. The results on the purchases of services components are more controversial. Especially for the components with small amount (and large CV), sometime the current procedure shows lower CV than the projection estimator. The calibration estimator outclasses the projection estimator also the component *C12905* of the other operating charges.

Figure 4.1 – CV (%) of the components (label from SME questionnaire) belong to *income from sales and services* realized by the generalized regression and projection estimator.

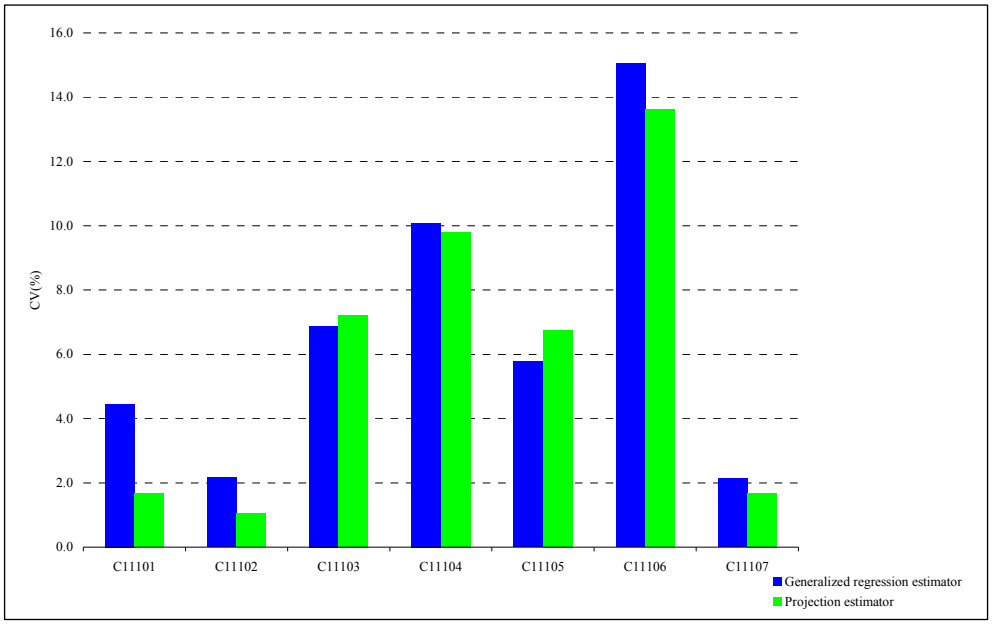


Figure 4.2 – CV (%) of the components (label from SME questionnaire) belong to *purchases of services* realized by the generalized regression and projection estimator.

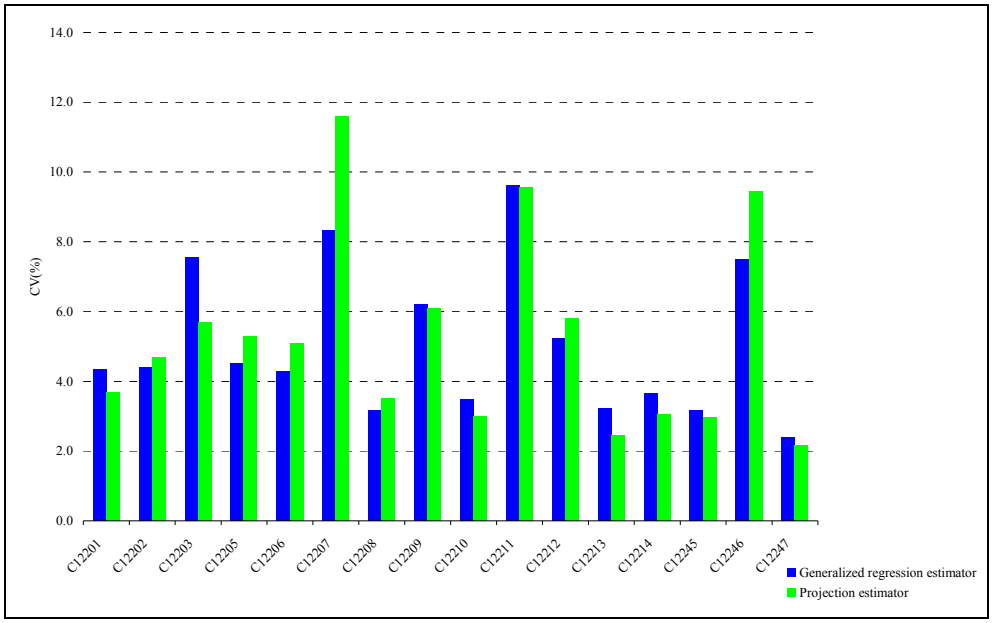
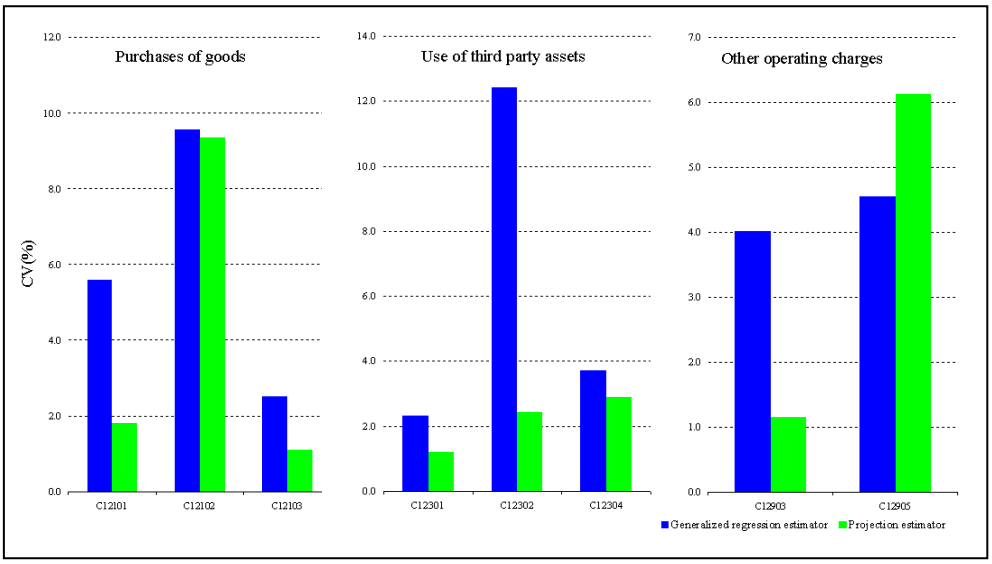


Figure 4.3 – CV (%) of the components (label from SME questionnaire) belong to *purchases of goods*, *use of third part assets* and *other operating charges* realized by the generalized regression and projection estimator.



The CV computed for the specific estimators at overall population level represents the average performance of such estimators a domain level.

Nonetheless, to get a really insight into the performances of the two estimators, we studied the CV distributions at domain level as well. The analysis reverses the relationships between the two estimators and the calibration estimator looks like better than the projection estimators. In particular, the former one produces lower CVs for totals relatively small (figure 4.4, 4.5 and 4.6). Figures 4.7 shows the median of the projection estimates of each component observed in the distribution of the domain estimates. When the median is quite small the CV distribution of the projection estimator is worse than the calibration estimator distribution. This evidence probably depends on the different sample sizes used since, for rare phenomena (or small amounts), the number of units have a greater impact on the precision on the estimates so ignoring the imputation step in the current estimation strategy the bias could be prevalent. On the other hand, the projection estimator shows his weakness for the small area estimation as usual for a direct estimator.

Figure 4.4 – Distribution of the CV (%) of the 583 domains for the components (label from SME questionnaire) belong to *income from sales and services* realized by the projection estimator (Projection) and the generalized regression estimator (GREG).

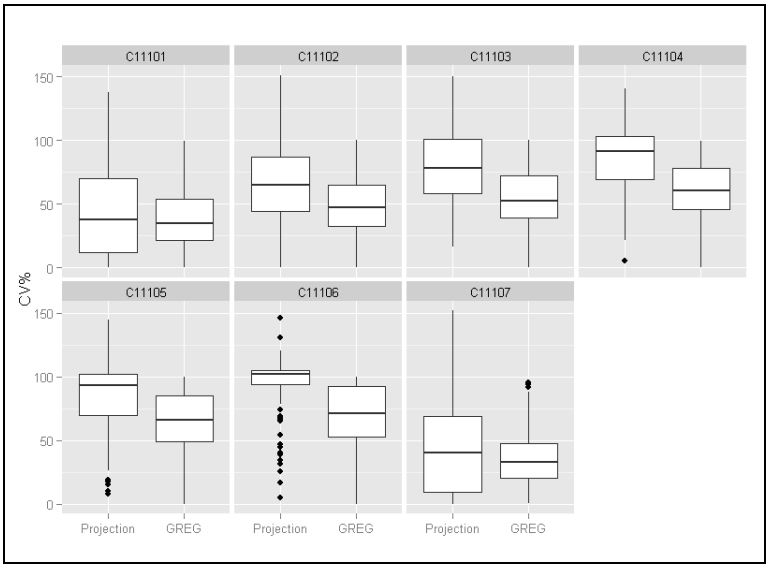


Figure 4.5 – Distribution of the CV (%) of the 583 domains for the components (label from SME questionnaire) belong to *purchases of services* realized by the projection estimator (Projection) and the generalized regression estimator (GREG).

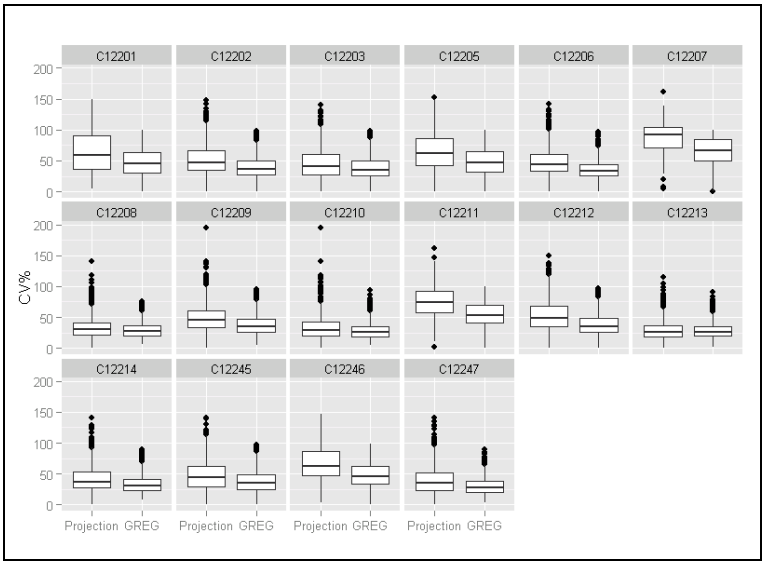


Figure 4.6 – Distribution of the CV (%) of the 583 domains for the components (label from SME questionnaire) belong to purchases of goods (upper left), use of third party assets (upper right)and other operating charges (lower left) realized by the projection estimator (Projection) and the generalized regression estimator (GREG)

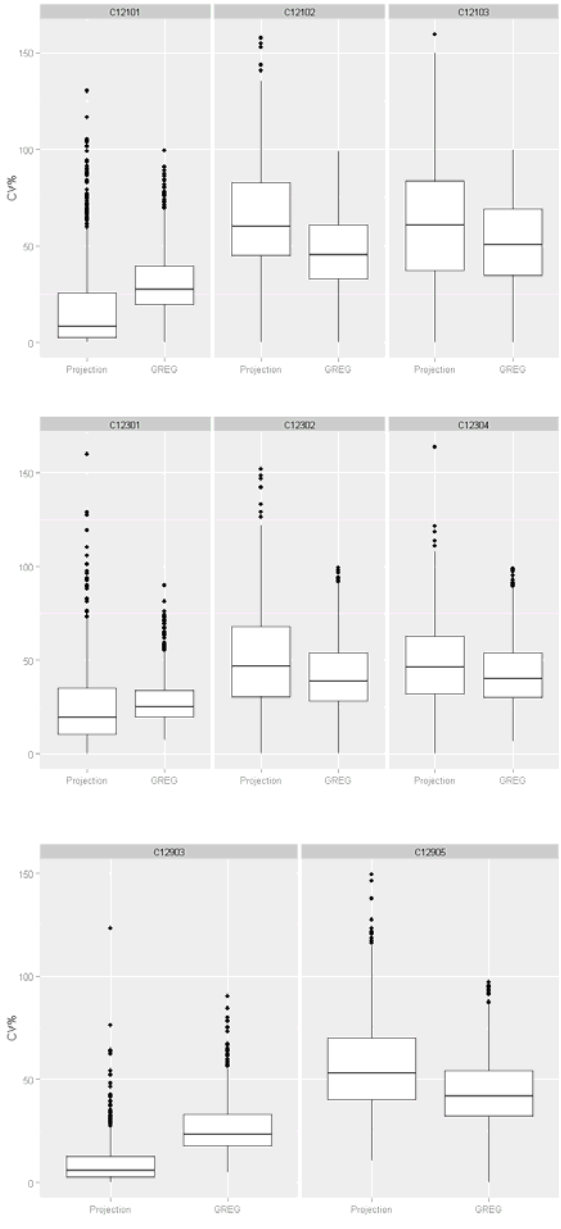
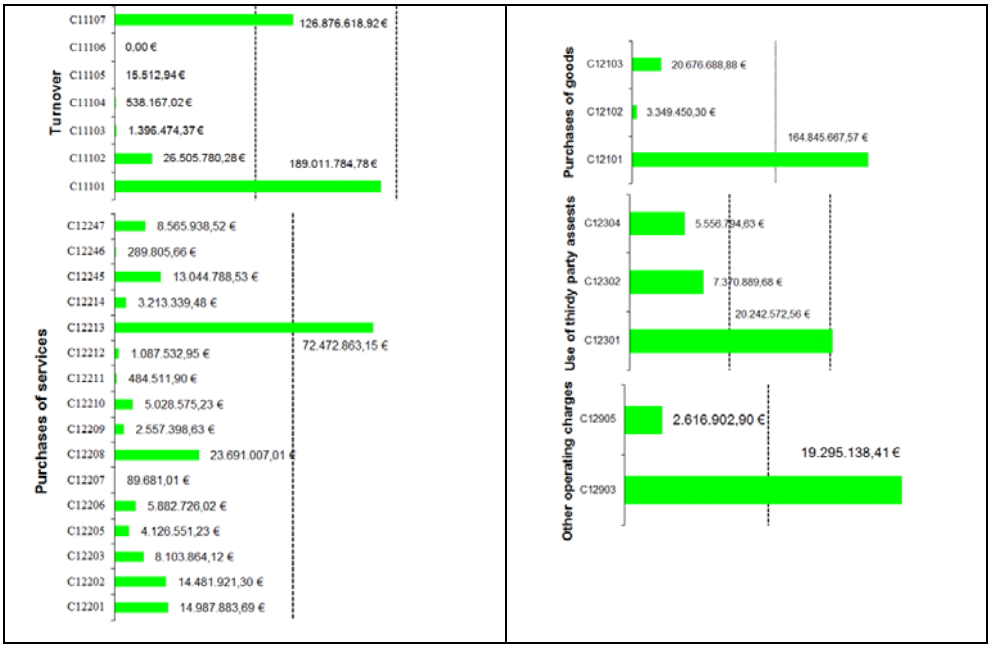


Figure 4.7 - Median values of the domain estimate distribution obtained by the projection estimator



5. Conclusions

The administrative data sources such as Balance sheets, Sector Studies, Tax returns, etc., although offer a large amount of economic variables are not exhaustive of the business information demand. The paper shows an estimation procedure, based on the projection estimator, to complete the set of estimates of the new Italian Business frame (Luzi e Monducci, 2016). The choice of using the projection estimator comes from a mix of operative conditions, theoretical properties and applicative opportunities. The estimation process is involved in a general context in which large data set and highly detailed domains are deemed. So an automatized and easy to implement method is quite appealing. The proposed process meets requirements and offers a well-founded inferential framework in which the sampling errors and bias are simple to compute and internal consistency is always satisfied. The outcome of the process is the input of other statistical processes. In particular, the Istat National Account (NA) sector bases its procedures on the frame and the imputation carried out by the projection estimator give interesting applicative opportunities. Anyway the projected values are not the true values and the inference must be carried out carefully at certain level of detail. In this case some tricks can be used. Otherwise, other estimation approaches, such as small area estimators, must be applied with the risk to complicate the sampling strategy. The projection estimator has been implemented from 2010 data onwards. The paper focuses on the precision of the estimates of the totals and a comparison with the current Structural Business Statistics estimates based on the SME survey is performed (year 2011). The SME survey uses the calibration estimator based on

the sample of respondent and the integrated non respondent. So the estimator uses about the double number of units with respect to the projection estimator that considers the sample respondents only. The findings have to be assessed taking into account that a part of the variance of the two estimators is ignored: in fact, the imputed values of the integrated respondents (for the current procedure) and the imputed economic aggregate variables of the frame (for the projection estimator) are treated as if they were observed.

As main results, the proposed technique for non small areas outperforms the current estimation strategy, because the auxiliary information of the new business frame are powerful predictors of the interest variables, underlying that the new estimation strategy will enhance the quality of the business statistics.

When in a given domain either or both the phenomenon is rare or the sample size is small (small area estimation problem) the comparison seems to be highly affected by the number of units used in the two estimators. In this case the performances of the current procedure is favored because a double sample size is used. As general indication, the result highlights a possible mean square error underestimation of the current procedure. As far the projection estimator is concerned, there are large CVs in many domains (Nace Rev. 2 3 digit by size class) even though the worst values should be only for the residual domains because for the overall population totals the CVs (considered as an average of the CV domain estimates) are quite low. The evidence recommends of using very carefully the estimates at high level of detail.

For these residual domains it should be better to use suitable estimators as small area estimators. But in this case other issues should be opened: integrate model assisted and model based estimates; know the domain types involved in the procedure (they are the domains of SBS Regulation, the domains of NA sector or types of domains that are not possible to foresee before processing the data); define a time spending process relate to the dimension of the data set, number of estimates and the estimation procedure itself.

Eventually, the proposed estimation procedure does not take properly into account the model uncertainty due to the imputation step of the economic aggregates implemented for some enterprises of the business register. The matter should be dealt with in the future for achieving a correct inferential analysis using data from the frame.

Appendix 1.

Proof of formula (2.5). Let us consider the expression (13) proposed by Kim and Rao (2012)

$$RB(\hat{Y}_{d,p}) = -\frac{Cov(\delta_k(d), r_k)}{\bar{\delta}(d)\bar{Y}_d},$$

where $\bar{\delta}(d) = (1/N)\sum_U \delta_k(d)$ and $\bar{Y}_d = (1/N_d)\sum_U y_k \delta_k(d)$.

Then

$$RB(\hat{Y}_{d,p}) = -\frac{N\,Cov(\delta_k(d), r_k)}{N\bar{\delta}(d)\bar{Y}_d} = -\frac{N\,Cov(\delta_k(d), r_k)}{[\sum_U \delta_k(d)][(1/N_d)\sum_U y_k\delta_k(d)]}$$
$$= -\frac{N\,Cov(\delta_k(d), r_k)}{\sum_U y_k\delta_k(d)} = -\frac{N\,Cov(\delta_k(d), r_k)}{Y_d}.$$

Appendix 2.

For obtaining a reduced downward approximate expression of the variance let us consider the working model used for imputing the missing values of the economic aggregate in the frame (Di Zio *et al*., 2015). We reformulate the first addendum of formula (2.6) as

$$Var_1[\sum_{U_O} f(\mathbf{x}_k, \boldsymbol{\beta}_0) + \sum_{U_M} f(\tilde{\mathbf{x}}_k, \boldsymbol{\beta}_0)], \qquad (A.1)$$

where the sample s_1 is replaced by $U = U_O \cup U_M$ with U_O and U_M respectively the population with observed and missing values and $\tilde{\mathbf{x}}_k$ the vector of imputed covariates in the frame. For sake of brevity we suppose a common pattern of missingness among the variables x_{qk} (q=1, …, Q). The Var_1 operator reflects the design variance so we need to introduce the model uncertainty of the previous imputation step.

Assume that the imputation of the x_q is ruled by the model $E_M(x_{qk} \mid \mathbf{z}_k) = g(\mathbf{z}_k, \boldsymbol{\gamma}) = \tilde{x}_{qk}$, with $Var_M(u_{qk} \mid \mathbf{z}_k) = \psi^2 b(\mathbf{z}_k)$, being u_{qk} the residual term, for some known function $b(.)$ and that $Cov_M(u_{qk}, u_{qj} \mid \mathbf{z}_k, \mathbf{z}_j) = 0$ for $k \neq j$, where the operators $E_M(.)$, $Var_M(.)$ and $Cov_M(.)$ are referred to the M imputation model. Since the expected values are equal to the true values, the model expectation is unbiased. Instead of the design variance we jointly consider the model and design variance. The model variance influences only the first addendum of the expression (2.6). Then we replace the expression (A.1) with

$$E_p E_M[\sum_{U_O} f(\mathbf{x}_k, \boldsymbol{\beta}_0) + \sum_{U_M} f(\tilde{\mathbf{x}}_k, \boldsymbol{\beta}_0) - \sum_U f(\mathbf{x}_k, \boldsymbol{\beta}_0)]^2, \qquad (A.2)$$

where $E_p(.)$ is the design expectation. Nevertheless the operator $E_p(.)$ disappears because we have a census and it can be shown that the (A.2) is equal to $\sum_{U_M} E_M\left\{f[g(u_{qk}, \boldsymbol{\gamma})]\right\}^2$. In case of $f(.)$ and $g(.)$ are linear function the model variance becomes $\sum_{U_M} \psi^2 b(\mathbf{z}_k)\boldsymbol{\beta}_0$.

References

Bakker B. F. M. 2012. "Estimating the validity of administrative variables". *Statistica Neerlandica*, 66: 8-17.

Casciano M. C., V. De Giorgi, F. Oropallo, G. Siesto. 2012. "Estimation of Structural Business Statistics for Small Firms by Using Administrative Data". *Rivista di Statistica Ufficiale,* n. 2-3/2012.

Deville, J.-C., C.-E. Särndal. 1992. "Calibration estimators in survey sampling". *Journal of the American Statistical Association*. 87: 376-382.

Di Zio M., U. Guarnera, R. Varriale. 2015. "The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". *Rivista di Statistica Ufficiale*, n. 1/2016.

Godambe V., Thompson M. 1986. "Parameters of superpopulation and survey population: their relationship and estimation". *International Statistical. Review*, 54: 127-38.

Hidiroglou M. 2001. "Double sampling". *Survey Methodology*, 27: 143-54.

Luzi O., R. Monducci. 2016. "The new statistical register "Frame SBS": overview and perspectives". *Rivista di Statistica Ufficiale*, pp. 5-14.

Kalton G., D. Kasprzyk. 1986. "The Treatment of Missing Survey Data". *Survey Methodology*, 12: 1-16.

Kim J. K., J.N.K. Rao. 2012. "Combining data from two independent surveys: a model-assisted approach". *Biometrika*, 99: 85-100.

McCall R. B. 2001. *Fundamental statistics for behavioural sciences*. Wadsworth, Belmont.

Merkouris T. 2004. "Combining independent regression estimators from multiple surveys". *Journal American Statistical Association*, 99: 1131-9.

Merkouris T. 2010. "Combining information from multiple surveys by using regression for efficient small domain estimation". *Journal of Royal Statistical Society B*, 72: 27-48.

Särndal, C. E., B. Swensson and J.H. Wretman. 1992. Model-assisted Survey Sampling. New York: Springer.

Oh H.L., F.J. Scheuren. 1983. Weighting adjustment for unit nonresponse, in: W.G. Madow, I. Olkin and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press: 143-184.

Rao J.N.K. 2003. *Small Area Estimation*. Wiley, New York.

Righi P., Falorsi S., Fasulo A. 2014. "A modified Delete a Group Jackknife variance estimator under random hot deck imputation in business surveys". In F. Mecatti F., P. L.Conti and M. G. Ranalli (eds), *Contributions to Sampling Statistics*. Springer: 219-233.

Schenker N., T. Raghunathan. 2007. "Combining information from multiple surveys to enhance estimation of measures of health". *Statist. Med.*, 26: 1802-11.

New experiences in the production of business statistics: the construction of the "Frame SBS" and SBS - data warehouse[1]

Francesco Altarocca [2], Diego Bellisai[3], Antonio Laureti Palma [4], Roberto Sanzo[5]

Abstract

This paper describes the first experience of data integration using administrative sources to estimate Structural Business Statistics (SBS) aggregates. It briefly shows the process that led to the construction of the "Frame", the "integrated" dataset that is the base to derive SBS estimates from the year 2012. This approach represents an innovative way to produce business statistics: the use of direct surveys is limited in favour of the use of administrative data. Moreover, this paper presents an innovative IT proposal on how to combine the integrated production model and the warehouse approach for the production of Structural Business Statistics. This consists in a metadata-driven data warehouse of integrated SBS information which is well-suited for supporting the management of modules in generic workflows. Such a modular approach can improve the efficiency of collaboration among different statistical experts on common data.

Keywords: Frame, Business statistics, SBS, Administrative sources, Economic aggregates, Scientific workflow, Data Warehouse, Data integration, Process integration

Introduction

Structural Business Statistics (SBS) are a powerful instrument to obtain a lot of detailed information about most aspects of Italian enterprises. The Eurostat Regulation n.58/97 and SBS EU Council Regulation n. 295/2008 define the main economical aggregates that have to be sent to Eurostat; these aggregates have to be provided for a defined estimation set and these estimates have a statistical significance only if referred to a certain period and to a certain universe.

The Italian Statistical Institute (ISTAT) since 1998 has derived these estimates carrying

[1] This work is the results of the common effort of the authors. However the paragraphs can be attributed as follows: Introduction and par.1 Roberto Sanzo; Conclusions and par. 2 Antonio Laureti Palma; par. 2.1 Diego Bellisai; par. 2.2 Francesco Altarocca. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

[2] Istat, e-mail: fraltaro@istat.it.

[3] Istat, e-mail: bellisai@istat.it.

[4] Istat, e-mail: lauretip@istat.it.

[5] Istat, e-mail: sanzo@istat.it.

out two different surveys depending on enterprise size[6]: the PMI survey (*Small and medium enterprise survey, including professional and artistic activities,* "*Rilevazione sulle Piccole e Medie Imprese e sull'esercizio di arti e professioni*"), a sample survey referred to enterprises with less than one hundred employees and involving about 120,000 enterprises, and SCI survey (*Survey on enterprise accounting system*, "*Sistema dei Conti delle Imprese*"), a total survey referred to enterprises with one hundred or more employees and involving about 10,000 enterprises.

In recent years, however, the increasing availability of information deriving from administrative sources has made possible to use them for statistical purposes both to make easier the analysis of respondents' data and to impute the statistical information when it was not available. In fact since the end of the '90s, the data about the Balance became an essential element of the production process of SCI, both during the editing step and during the integration of total non-response phase. In the 2000s, also the PMI survey began to use some administrative sources to try to estimate the information for non-responding firms.

The agreements between the ISTAT and the government (in particular the "*Agenzia delle entrate*" and "*Unioncamere*"), have ensured ISTAT to obtain an unprecedented amount of data that suggested the massive use of it for making statistics; the aim was to try to reduce the "statistical burden" for the respondents too.

The new approach for SBS was to obtain business information for each unit in Italian Business Register (ASIA, "*Archivio Statistico delle Imprese attive"*) using administrative data: in this way, the aggregates could be simply derived using "sum of values" among different units instead of using sampling and calibration techniques. This approach was tested in 2010-11 on PMI Universe; for enterprises with 100 employees or more the SCI survey was preferred because it ensured a better quality of estimation. In fact, not all the SBS aggregates can be derived from administrative data: in an administrative source, the business information can be or not be present; moreover the data were not available for all units: some techniques of imputation or estimation are however needed to ensure a "complete sum of values" for all the units to obtain the main SBS aggregates.

The SBS Frame is the final result of this integration activity; it represents the first product completely derived from administrative data and used to obtain statistical aggregates: it provides a dataset including the values of main economical aggregates for each unit in SBS universe by Asia Register.

Finally, given the big amount of information to be managed and to increase its usability, the Frame has to be produced in an innovative application framework based on a data warehouse supported by a workflow engine. .

This paper is structured in two parts: in the first chapter, the steps to make the Frame will be synthetically described and some simple results will be showed; the second chapter presents an innovative IT proposal on how to combine the integrated production model and the warehouse approach for the production of Structural Business Statistics, consisting in a metadata-driven data warehouse of integrated SBS information.

[6] For more details, refer to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2015. "Quality analysis and harmonization issues in the context of the SBS frame". *Rivista di Statistica Ufficiale*. n.1/2016.

1. The "Frame": the making of

In order to build an information structure able to estimate the SBS variables and the aggregates of national accounts (NA) using administrative sources, the following sources were considered :

- *Financial Statements* (hereafter BIL);
- *Sector Studies survey* ("*Studi di settore*", hereafter SDS);
- *Tax returns* ("*Modello Unico*", hereafter UNI);
- *Regional Tax on Productive Activities* ("*Modello IRAP*", hereafter IRAP).

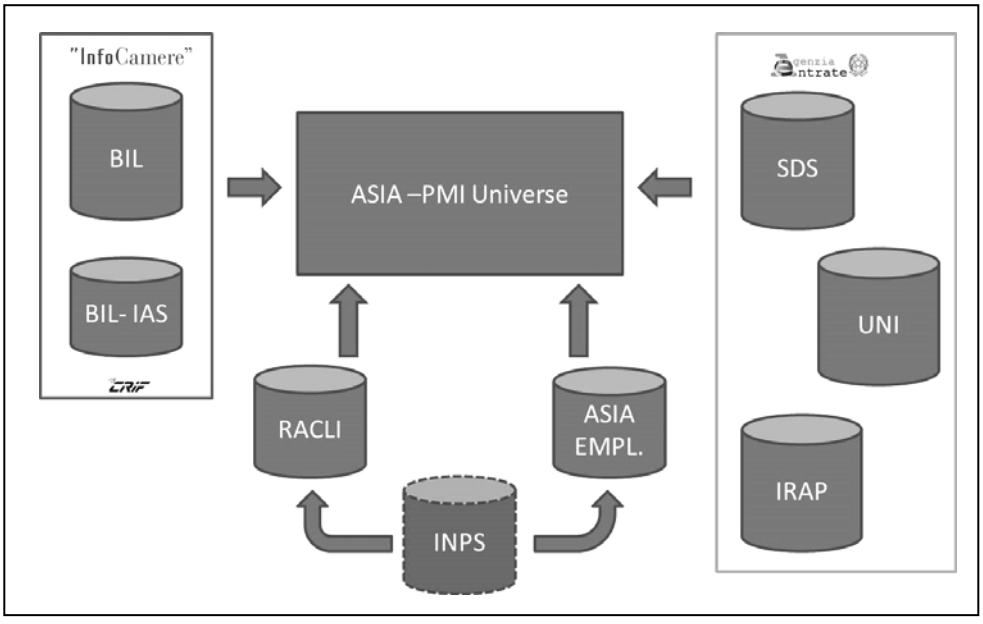
The Financial Statements (BIL) represents the main information document about enterprises dynamics: it includes the costs and revenues of the entire accounting period and includes the assets and the liabilities that a company has incurred during the year.

The SDS source is a tool that the Italian tax authorities used to detect the economical parameters for professionals, self-employed and small firms. The main aim is to identify the activity and the economic environment in which the enterprise operates, in order to assess its ability to produce real income.

The UNI source is the model of ordinary income tax return while IRAP is a local tax that applies to productive activities in each region: it must be paid only by those who carry out business activity and not by individuals.

As a reference universe is considered what is defined by the SBS Regulation and its identification is made through the Business Register (ASIA): in particular, we considered all the firms in industry and services (excluding financial firms and insurance) with less than 100 employees and active for more than six months in the concerned year (PMI Universe).

Figure 1: Logical scheme for the administrative data integration model.



Moreover, as support archives used in the different phases of construction of this information structure , from now on called Frame, are also used information from :

• Archive RACLI ("*Registro anagrafico del costo del lavoro per impresa*"), with regard to the cost of labour and its items[7]; it derives the Social Security Data from the Italian National Security Institute (INPS).

• ASIA-Employment, regarding the information about atypical workers and their remuneration.

The Figure 1 shows the logical scheme used for deriving statistical information from administrative source: the center of the scheme is the PMI universe as defined by ASIA according to SBS Regulation. The last two archives (RACLI and ASIA-Employment) entered the process to ensure a verification of information about personnel costs (see § 1.1).

1.1 Administrative items and statistical variables: general aspects and the personnel costs correction

The use for statistical purposes of information from administrative files, which, by definition, are kept for reasons other than those for which we will try to use them in this context, involves a whole series of preliminary activities to

• identify what items can be used to build the economic variables needed;

• evaluate the consistency with the statistical definitions;

• verify the information content of the items in the various administrative sources.

Regarding these three points, the different administrative forms were analyzed by experienced ISTAT personnel and the items useful for statistical purposes were identified. The next analysis was about the fitting of administrative and statistical values (in this case, comparing the values obtained by respondents to PMI survey) for identifying, for each administrative source, the different degree of fitting of the main economic aggregates used in SBS[8].

It was not easy to identify a systematic behaviour in compiling of the administrative forms by accountants or tax operators able to explain differences with surveys questionnaires. However, when the administrative item and the statistical variable should not be different according the definition, these differences could be considered "noise" and those could be treat using statistical methods.

An exception is represented by the variable "personnel costs" that has an additional source of comparison (RACLI): for enterprises using atypical workers the discrepancies between statistical and administrative information could be explained by the remuneration of the atypical workers; also these differences could be explained by independent workers even if the enterprise does not use atypical workers. In fact, the empirical analysis showed that sometimes the remunerations of the atypical or independent workers were wrongly

[7] For more details, refer to: Arnaldi S., Baldi C., Filippello R., Mastrantonio L., Pacini S., Sassaroli P., Tartamella F. 2016. "The labour cost variables in the building of the frame". *Rivista di Statistica Ufficiale*. N.1/2016.

[8] For more details, refer to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2016. "Quality analysis and harmonization issues in the context of the SBS frame". *Rivista di Statistica Ufficiale*. n.1/2016.

included in the personnel costs item instead of in “costs of services” or in profits [9].

For this reason, it was necessary to apply a correction procedure: however to ensure consistency with other economic variables, for constructing variable “personnel costs” in the Frame is used the information coming from administrative sources (except RACLI); RACLI is used in a procedure of "correction" (called COMBO) for separating the "noise" caused by atypical and/or independent workers from the administrative data.

Only in cases where the administrative data about labour cost were higher than the value in RACLI, the COMBO procedure tries to determine, according to the information derived by ASIA Employment, the best possible combination of remuneration values of atypical workers and independent to be deducted from the personnel costs, to get a value as close as possible to the RACLI value. Then the amounts deducted from the personnel costs were added to the costs of services (if related to atypical workers) or to the profit (if related to independents) to ensure consistency with other information.

The COMBO procedure was applied separately in the sources and has produced changes in a limited number of units (about 3.5% in BIL, less than 1% in the SDS and UNI) with variations respect the original data that, overall, ranging from -1.5% to -0.7% in terms of personnel costs and 0.8% to -0.2 % in terms of added value. In fact a conservative perspective was used: the correction is made only when there are a sufficient number of clues to ensure the successful application of the procedure. The problem did not concern for the source IRAP where the personnel costs is not used.

A similar correction, but in the opposite direction, was made (if COMBO did not act) for all units for which ASIA-Employment provided an indication of the presence of independent reclassified by ASIA Register as para-subordinates: their remuneration was deducted from the cost of services and moved in the profit.

1.2 Pre-treatment of administrative data: the data editing procedures

When an administrative item useful to statistical purpose was identified, at least other two steps were necessary:

- checking that the information relating to the same unit within each source is unique;
- checking and correcting all the inconsistencies that occur among items within each administrative source.

Regarding these points, it was necessary to implement some statistical procedures of data-editing to check, correct and/or eventually erase anomalies in the sources .

To ensure the “internal consistency”, the administrative data were treated directly, as those have been acquired, also verifying relationships with variables that are not useful for the construction of statistical aggregates. But because of the high amount of data (millions of records for hundreds of items), not all items were involved in the check activity: in fact, for each of these sources only items directly or indirectly useful to the construction of economic aggregates were tested.

Then, the main activities carried out were:

1. identification and elimination of "duplicate records",
2. identification and elimination of internal inconsistencies in each source,

[9] For more details, refer to: Arnaldi S., Baldi C., Filippello R., Mastrantonio L., Pacini S., Sassaroli P., Tartamella F. 2016. The labour cost variables in the building of the frame. *Rivista di Statistica Ufficiale*. n.1/2016.

3. exclusion of useless records

All these activities were performed independently on each of the administrative source.

In all considered sources were sometimes present more record referring to the same unit: these duplications were due to the fact that it is possible for the enterprises to re-present an administrative form already presented if they have to correct or to improve their declaration. But it is not simple for the statistician to understand what to do when there are more records for a single unit: the flags deriving from administrative form often are not satisfactory in all the cases because these flags seldom are nonzero. For this reason an interactive analysis of duplications was necessary: referring to the 2010 and 2011 occurrences, some rules were identified to help the choice. These rules made it possible to remove a lot of duplications in the 2012 and 2013 too; but since these rules are empirical, they are not exhaustive and then every year an interactive step is necessary to treat a residual number of cases.

The next step was to check and eventually correct inconsistent data. For this purpose for each source a consistency plan was built; a certain number of rules were considered: to accept a single record all of these rules had to be come true. In particular, domain edits, prorate edits or "less than" edits for "specification items" were the edits considered in the consistency plan: to correct inconsistencies, a deterministic method was used and the correction rules were based on the relationships among the variable.

At the moment, the check procedure does not provide comparisons with information deriving from the other sources: only for BIL is made a comparison with SDS data. In perspective we expect to use similar procedures for UNI, SDS and IRAP too.

When a record is formally exact, we had to decide if it was useful for the Frame: in fact if no item of the administrative form was used for derive a statistical variables, the related record was deleted from the dataset.

1.3 Harmonization of variables

The activity of "harmonization"[10] is crucial for binding the administrative items to statistical variables. It is carried out only after the elimination of the causes of "noise" previously described: to make the information recognizable and comparable as much as possible, the same names were used in the different sources, diversifying them only by a suffix indicating the origin of the source. These names come from names used in SBS.

All the steps described above must be driven by a set of metadata as comprehensive and clear as possible; unfortunately the greatest difficulty encountered in the first part of the work was just the availability of accurate metadata in electronic format. It is hoped, however, that for the next realizations of the Frame the metadata problem will be reduced through agreements with the government, the "owner" of the data.

The application of the procedures in the various steps of pre-treatment of data decreases the number of units available: this is the result of the "cleaning" of the data that includes not only the elimination of duplicate records but also of record with data incorrect or useless (e.g. all values equal to "1"); other deleted records are those records that do not have "significant" data for the purposes of this context (e.g. missing values for all item used for

[10] For more details, refer to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2016. "Quality analysis and harmonization issues in the context of the SBS frame". *Rivista di Statistica Ufficiale*. n.1/2016.

the construction of the economic variables here considered).

The Figure 2 shows a summary of the preliminary activities previously described for the construction of the Frame, starting from "Gross data" to arrive to "harmonized data".

Figure 2: Steps of pre-treatment of administrative data for the construction of the Frame.

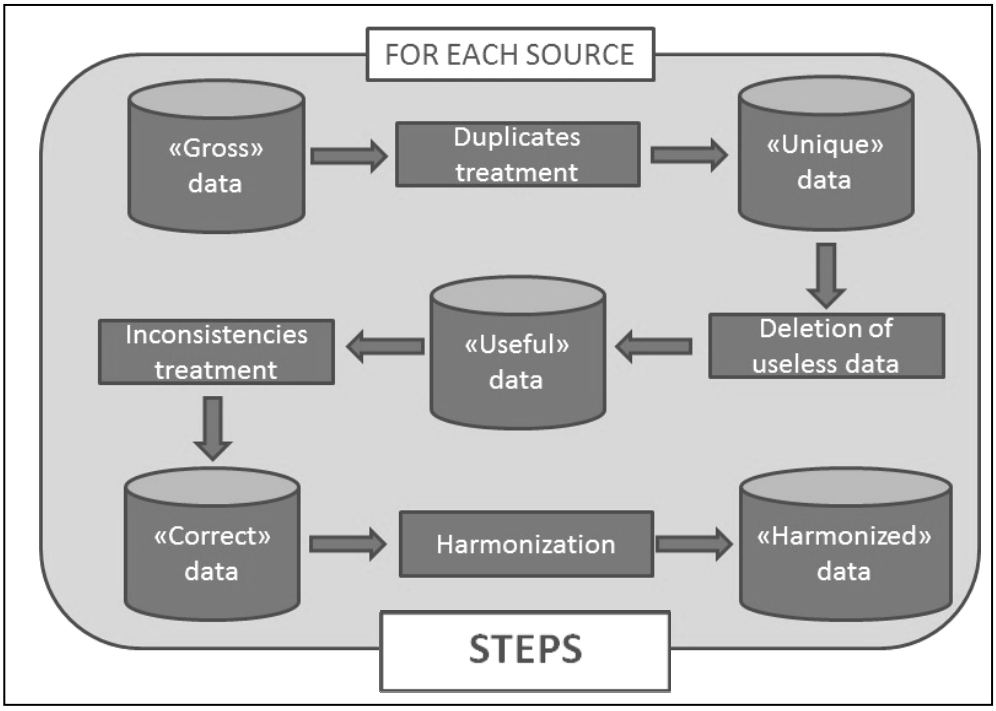


Table 1: Number of treated units in each source for steps of pre-treatment. Year 2012.

Steps	BIL	SDS	UNI	IRAP
Gross data	747,492	3,973,230	4,498,531	3,658,585
Duplicates deletion	746,953		4,329,326	3,532,179
Data Editing	743,737	3,973,230	4,319,252	3,495,617
Harmonized Data	743,737	3,580,745	4,313,632	3,494,493

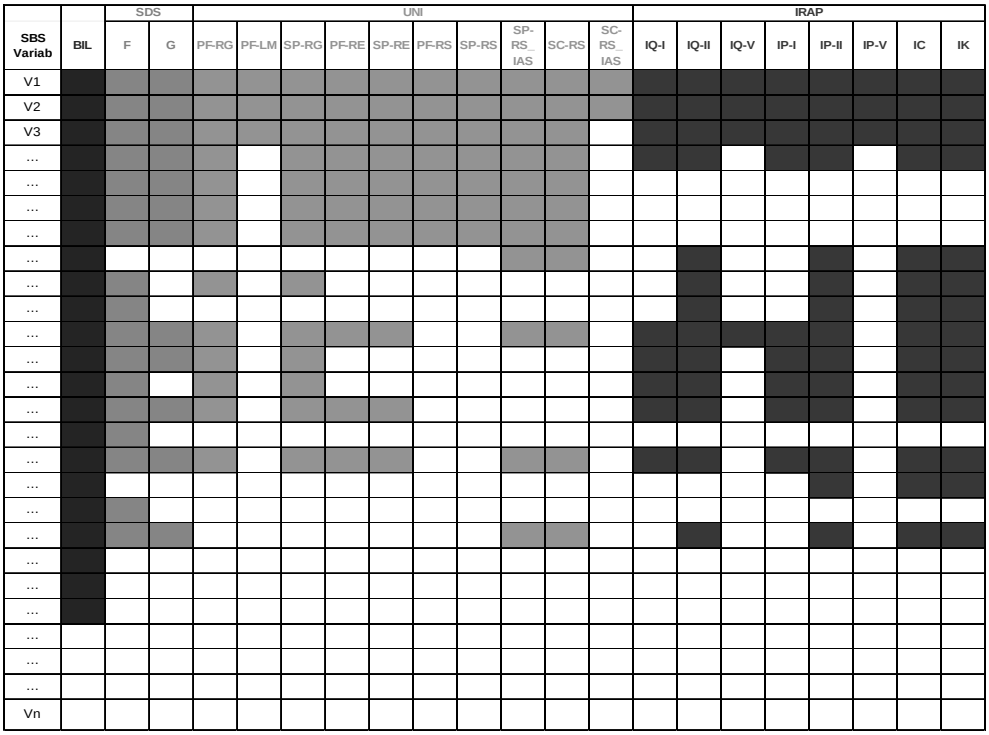
The Table 1 shows the number of units treated in the different pre-treatment steps.

The harmonization step produced the table shown in Figure 3: this table is just a fictional representation of what really happened: the colored cells indicate the presence of SBS variable within a source. The Figure shows clearly that not all SBS variables are covered by administrative source and that the different administrative sources do not provide all the SBS variables: some sources provide more SBS variables then others. Moreover, also within the source, the same variable could be or not present, according to

the type of enterprise and to compiled form.

In conclusion, paradoxically in some cases the information is too broad, in others it is virtually absent: for this reason the next step was to arrange this huge amount of data - with different degrees of quality - for identifying the best business information to assign to each unit in PMI universe.

Figure 3: Fictional representation of availability of SBS variables in administrative data.



1.4 Data integration: hierarchical order and source priorities

To assign a set of statistical variable to each unit of PMI universe; three situations were shown:

1. More sources for the same unit
2. Units with only a source
3. Units without information.

The choice is certain for point 2: if information is available only from a source, the SBS aggregates are derived from that source.

Regarding to point 3, it can be considered as a non-respondent case in a survey, then a statistical method of imputation of total response can be apply. In this context, the

"predictive mean matching" and the "minimum distance donor" methods have been used[11].

The choice of source for units with more sources available was more complex: it was necessary a set of analysis for determining the accuracy between administrative and survey data in each source. Furthermore, these analysis showed different results depending on the variables used. For this reason not all economic variables derived in the different sources have been considered at the same level of accuracy[12]. The main reasons were:

- Definitions are very similar but not identical;
- "Noise" introduced by the administrative purposes of data collection from different sources (errors in the sources);
- "Noise" introduced by the statistical definitions that are not easily attributed to administrative definitions (errors in the survey).

Then, the choice was made taking into account two conditions:

a. The "importance" of variables with an high degree of accuracy;
b. The number of accurate variables.

Based on these two conditions, a list of source priorities was identified for deciding which source to use if more sources had data for the same unit: BIL showed a better accuracy to the survey data and this result was easily predictable because PMI variables definitions are based on Financial Statements legislation. In addition, BIL represents the richest source in terms of the number of variables.

Despite a smaller number of usable variables than BIL, SDS was a rich source considering the number of statistical units: more than 3.5 million units. The advantage of this source, compared to UNI, was that data are derivable from the same section (Section F) for all units, in spite of differences of firms activities; the only exception is the professionals who fill out the section G. Then, SDS was identified as an alternative source to BIL because of its uniformity and its relative richness of information.

The UNI source was used as third choice in case of lack of the first two sources: the reasons are to be found in the high heterogeneity of the variables, in a number quite limited of information and in the difficulty of standardizing the data because of the type of firm (individuals, partnerships and corporations), which filled different forms also because of the used accounting system.

For the year 2011 the IRAP source was acquired only when the making of the Frame was already nearing completion. Because the accuracy of it with PMI data has not been discussed in depth, it was used only as a last resort in case of absence of any other source and only for the corporations, because the analysis performed for the integration of non-response in the SBS survey about enterprises with 100 employees and over (SCI) showed a good accuracy with survey data.

In a second time, more analyses were conducted on other type of enterprises too: the results were the accuracy depends more on Section than on Form. In fact a very good accuracy for data coming from Section II of both forms IP and IQ, a good accuracy from Section I, a bad accuracy from Section V. These analysis permitted to change the order of

[11] For more details, refer to Di Zio M., Guarnera U., Varriale R. 2016. "The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". Rivista di Statistica Ufficiale. N.1/2016.

[12] For more details, refer again to: Curatolo S., De Giorgi V., Oropallo F., Puggioni A., Siesto G. 2016. "Quality analysis and harmonization issues in the context of the SBS frame". Rivista di Statistica Ufficiale. n.1/2016.

priority of sources for the next years, although a great problem of using IRAP source is that the stability of information is not ensured because it is more prone to policy decisions than other sources. However, in 2012 and 2013, IRAP source entered in the construction procedure as the others.

In conclusion, the order of priority used was the following:

1. BIL, if there is not then
2. SDS, if there is not then
3. UNI or IRAP, depending on sections and forms.

As previously said, for the enterprises as defined in the Business Register (ASIA) that are not present in any of the considered sources, it was needed a step of statistical integration of the missing information, relying on information from the RACLI source (for the personnel costs) and ASIA with regard to the turnover, derived from "Value Added Tax", VAT, in particular.

1.5 The SBS Frame

After the correction for the remuneration of the atypical and independent workers (see § 1.1) and the harmonization of variables, the next step was the record linkage by the ASIA identification code to derive data from administrative sources (BIL, SDS, UNI, IRAP) referring to the SBS universe (in particular, PMI universe), discarding all information relating to non-active or not present in ASIA units.

The choice of data for every enterprise was made according to the order of priority as defined above. At the end, information from RACLI were linked too.

Some "exceptions to the priority" have been considered: the first exception regards the non-solvable inconsistencies: if the causes of inconsistencies are impossible to find, the involved source for that record was considered missing. Another exception was identified in BIL: records with an accounting period other than 12 months were discarded: in these cases the sources SDS, UNI or IRAP are considered more reliable, because those are referred to the calendar year.

Another exception came from a subsequent analysis of per-capita, in particular the personnel costs per employee. The cases of too high and/or too different from RACLI per-capita advised to waive the priority and to use the source following in hierarchical order: if no data was available or consistent, the unit was considered as a non-respondent and the information were imputed.

The Figure 4 shows the data integration step, as previously described: this figure is important because it shows the moment when the four different sources were linked to identify the statistical information to be used in all the following steps. Before step, the administrative datasets were treated independently and only in this time the different information are processed together.

The result of this step is a data set containing for each unit in PMI universe, all possible information related to a set of economic variables; the resulting dataset is exemplified in Figure 5, where the colored cells indicate the presence of information: for each unit one, more than one or none administrative source were available and the choice of data was made according to the rules previously described.

Figure 4: Administrative data integration: the Frame_base dataset.

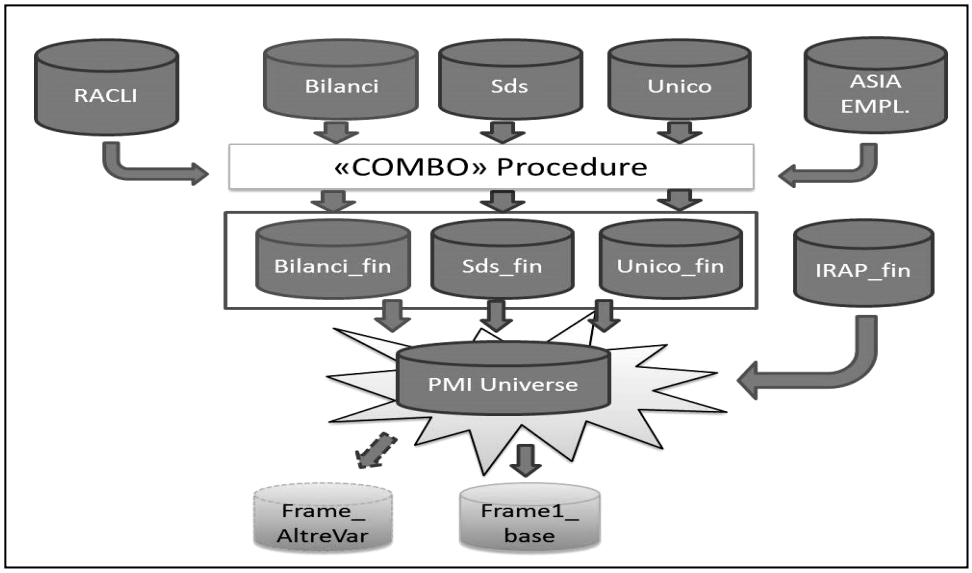


Figure 5: Exemplified representation of integrated dataset after linkage step.

	SOURCE			
Unit	BIL	SDS	UNI	IRAP
U1				
U2				
U3				
U4				
U5				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
...				
Un				

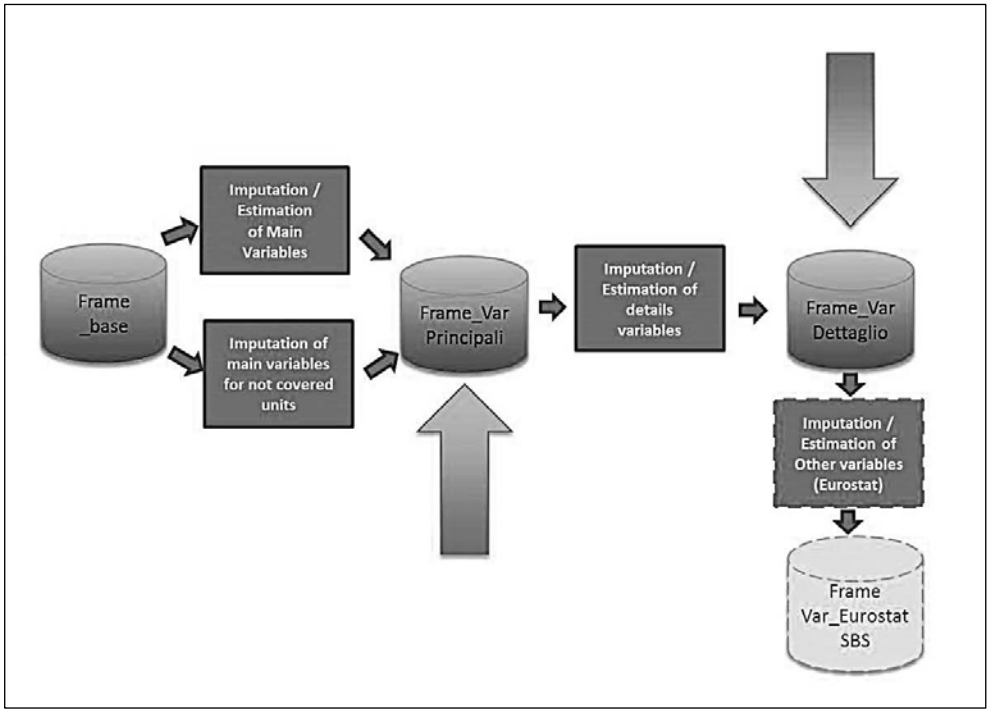
At the end of this process for each unit - with at least one administrative source - was present a set of information provided by one and only one source; for the remaining units (without information available from administrative data) only the structural information

deriving from ASIA and economic variable deriving from RACLI were considered: the resulting dataset is called *Frame_base*. This dataset represents the starting point for the construction of the SBS Frame, a complete set of information for the SBS estimates and for the National Accounts (CN) aggregates too.

A byproduct of the linking above is a dataset containing a set of variables useful for estimating the non-regular economy: on the contrary to the previous ones, the variables present in it, were not checked. The estimates of some other variables of detail for other revenue, made by the CN using the information from the "Notes" of BIL, are added too.

Starting from *Frame_base*, the next step is the imputation of missing values using both deterministic procedures and statistical ones, such as the predictive mean matching or the donor of minimum distance method[13]. The result was a new dataset called *Frame_VarPrincipali* where the values for all the main variables for all units of the Frame were available; it contained the main economic variables like turnover, costs, value added, personnel costs and so on. Also the details of personnel costs were present in it: those were estimated, if necessary, considering the structure provided from RACLI.

Figure 6: Final steps for the making of the Frame.



[13] For more details, refer to: Di Zio M., Guarnera U., Varriale R. 2016. "The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". *Rivista di Statistica Ufficiale*. N.1/2016.

At the end, the last step was the estimation of the details of revenue or costs not previously calculated from *Frame_VarPrincipali*; the result was another dataset called *Frame_VarDettaglio*. The variables was estimated by the "projection estimator"[14] using survey data from PMI.

The Figure 6 shows the final steps of making of the Frame: the arrows indicate the datasets currently available to determine SBS statistics; they are the starting point for NA aggregates estimation. The Frame is available for years 2011, 2012 and 2013; for 2010 is available a prototype version.

The Table 2 shows the number of units in the Frame for the year 2012 according to the source of origin of the data: it is also shown the contribution that each source provides to the total of the number of employees and of some economic variables (turnover, value added and labour cost). The table has been derived from the elaboration of the Frame after the step of imputation and estimation of the main variables; in fact, also it shows the estimated values for not covered units. For 2013 the results are similar.

1.6 Future perspectives

Since 2012, the SBS Frame, with SCI survey, has been the core of SBS estimates sent to Eurostat: the main economical aggregates (e.g. turnover, added value, labour cost, etc.) are built simply using the "sum function"; instead to obtain other aggregates (e.g. detail of costs, detail of turnover, etc.) is necessary to apply a statistical method for estimation: also in these cases the SBS Frame provides data for each unit but they are significant just within the estimation domains of projection model.

Table 2 - Number of units and main economic aggregates of the Frame for source of data. Year 2012. (to be continued).

SOURCE	Number of enterprises	%	Number of employees	%	Turnover (mln €)	%	Value-Added (mln €)	%	Personnel costs (mln €)	%
BIL	707,167	16.0	4,572,661	37.8	1,138,480	64.4	221,203	53.2	139,348	66.4
SDS	2,985,929	67.5	6,064,848	50.2	433,273	24.5	153,962	37.0	51,612	24.6
- Section F	*2,280,386*	*76.4*	*5,093,983*	*84.0*	*378,115*	*87.3*	*113,779*	*73.9*	*46,952*	*91.0*
- Section G	*705,543*	*23.6*	*970,865*	*16.0*	*55,159*	*12.7*	*40,183*	*26.1*	*4,660*	*9.0*
UNI	542,721	12.3	722,050	6.0	34,369	1.9	13,648	3.3	3,190	1.5
- Professionals and self-employed										
- Form RE	*66,270*	*12.2*	*77,421*	*10.7*	*3,628*	*10.6*	*2,545*	*18.6*	*243*	*7.6*
- Form RF	*357*	*0.1*	*813*	*0.1*	*76*	*0.2*	*21*	*0.2*	*10*	*0.3*
- Form RG	*200,539*	*37.0*	*295,174*	*40.9*	*13,273*	*38.6*	*3,552*	*26.0*	*1,428*	*44.8*
- Form CM	*230,026*	*42.4*	*223,725*	*31.0*	*4,416*	*12.8*	*3,283*	*24.1*	*6*	*0.2*

[14] For more details, refer to: Righi P. 2016. "Estimation procedure and inference for component totals of the economic aggregates in the new Italian Business frame". *Rivista di Statistica Ufficiale*. N.1/2016.

Table 2 (continues): Number of units and main economic aggregates of the Frame for source of data. Year 2012.

SOURCE	Number of enterprises	%	Number of employees	%	Turnover (mln €)	%	Value-Added (mln €)	%	Labour costs (mln €)	%
UNI										
- Partnership										
- Form RE	*1,963*	*0.4*	*9,832*	*1.4*	*1,335*	*3.9*	*725*	*5.3*	*99*	*3.1*
- Form RF	*284*	*0.1*	*2,209*	*0.3*	*392*	*1.1*	*84*	*0.6*	*49*	*1.5*
- Form RG	*38,335*	*7.1*	*89,226*	*12.4*	*3,732*	*10.9*	*901*	*6.6*	*548*	*17.2*
- Company										
- IAS	*698*	*0.1*	*4,454*	*0.6*	*3,645*	*10.6*	*1,393*	*10.2*	*254*	*8.0*
- Others	*4,249*	*0.8*	*19,196*	*2.7*	*3,873*	*11.3*	*1,144*	*8.4*	*552*	*17.3*
IRAP	78,667	1.8	397,482	3.3	124,693	7.1	18,285	4.4	9,956	4.7
- Professionals and self-employed										
- Form IQ, Sect.I	*17,998*	*22.9*	*62,485*	*15.7*	*7,945*	*6.4*	*1,344*	*7.4*	*936*	*9.4*
- Form IQ, Sect.II	*2,467*	*3.1*	*10,893*	*2.7*	*1,658*	*1.3*	*341*	*1.9*	*199*	*2.0*
- Form IQ, Sect.V	*484*	*0.6*	*934*	*0.2*	*44*	*0.0*	*15*	*0.1*	*7*	*0.1*
- Partnership										
- Form IP, Sect.I	*15,984*	*20.3*	*87,989*	*22.1*	*15,638*	*12.5*	*2,205*	*12.1*	*1,599*	*16.1*
- Form IP, Sect.II	*3,437*	*4.4*	*27,432*	*6.9*	*7,667*	*6.1*	*1,184*	*6.5*	*658*	*6.6*
- Form IP, Sect.V	*60*	*0.1*	*242*	*0.1*	*12*	*0.0*	*4*	*0.0*	*2*	*0.0*
- Company (Form IC)	38,195	48.6	206,760	52.0	91,698	73.5	13,182	72.1	6,526	65.5
- Economic Public Corporation (Form IK)	42	0.1	746	0.2	29	0.0	10	0.1	29	0.3
NOT COVERED	109,489	3.9	328,388	5.1	37,038	6.2	8,559	5.3	5,725	6.3
- Professionals and self-employed	*47,219*	*43.1*	*89,488*	*27.3*	*6,393*	*17.3*	*1,594*	*18.6*	*807*	*14.1*
- Partnership	*13,145*	*12.0*	*40,311*	*12.3*	*3,357*	*9.1*	*846*	*9.9*	*443*	*7.7*
- Company	*40,787*	*37.3*	*148,503*	*45.2*	*20,709*	*55.9*	*4,645*	*54.3*	*3,273*	*57.2*
- Cooperative society	*5,667*	*5.2*	*40,701*	*12.4*	*3,528*	*9.5*	*921*	*10.8*	*735*	*12.8*
- Consortium	*1,196*	*1.1*	*3,019*	*0.9*	*667*	*1.8*	*127*	*1.5*	*106*	*1.8*
- Economic Public Corporation	*183*	*0.2*	*2,928*	*0.9*	*584*	*1.6*	*107.0*	*1.2*	*94*	*1.6*
- Foreign firm	*1,292*	*1.2*	*3,439*	*1.0*	*1,801*	*4.9*	*319*	*3.7*	*267*	*4.7*
TOTAL	**4,423,973**	**100**	**12,085,428**	**100.0**	**1,767,852**	**100.0**	**415,658**	**100.0**	**209,830**	**100.0**

As a new experience of massive use of administrative data for statistical purpose, the SBS Frame needs some years for adjusting the methods and the techniques; however, at this time the results are encouraging because the analysis of the differences between the Frame and the PMI estimates shows a similar behaviour between one year and the previous.

In recent years, ISTAT has invested a lot of skills (statisticians, IT resources etc.) to guarantee a long-term solution for the Frame construction process: some workgroups and task-forces have arranged to study different solutions for improving all the steps of the process, both about methodological and IT issues. This activity needs a continuous development and support to ensure an increasing quality of SBS estimates.

Also, the Frame will have to ensure the possibility of provide estimates for variables not available at this time: the use of other administrative sources (i.e. VAT or Financial Statements notes) could help to derive, for example, estimates about investments. The short-term statistics could be used too, for example to estimates the number of worked hours; this solution probably will be implemented in Frame 2014 while for the estimate of investments it is probably necessary at least one more year for properly studying the new sources.

Other activities are focused on improving the editing method introducing the use of probabilistic methods too, or on modifying the imputation techniques for units without administrative data (e.g. estimation of turnover by VAT using regression models in homogeneous domains); furthermore, a systematic selective editing procedure for the most influent units will soon be introduced in the process, as new analysis on increasing the quality of the choice of the administrative source when more sources are available will be further detailed.

Moreover, for the next Frame occurrences, it will be strategical to have a solid IT architecture to ensure the safety and the replicability of the process: this is a long-term activity but it will be one of the most important task in the future (in fact, it is currently in progress).

2. The warehouse system supporting Frame production

In statistical production the statistical burden is one of the main problems to be faced. To cope with the problem the two main directions are: the intensive use of administrative data and the reduction of the stovepipe production models. In fact, in a stovepipe model different surveys are totally independent from each other in almost every phase of statistical production processes.

Using administrative data often means to elaborate several large archives of data. In this case, an IT Data Warehousing (DWH) model[15] could be a suitable approach. In particular, we can consider a Statistical-Data Warehouse (S-DWH), i.e. a DWH specialized and optimized for producing statistical information and for data reuse (by storing data once but using it for multiple purposes). This means that we could manage several production processes for different topics in the same statistical domain, i.e. a data base for a single environment in which we could have data integration and sustain process integration. In fact, the data model underlying a S-DWH is not oriented to produce specific reports as well as for on line analytical processing. Instead of focusing on a process-oriented design, the underlying repository design is based on data inter-relationships that are fundamental for different processes of a common statistical domain.

[15] Inmon, Bill (1992).Building the Data Warehouse. Wiley. ISBN 0-471-56960-7

The S-DWH data model is based on the ability of realizing data integration at micro or macro data granularity levels: micro data integration is based on the combination of different data sources with a common unit of analysis, while macro data integration is based on integration of different aggregate information in a common estimation domain. In this paper, we will consider only micro data integration in order to support the complex procedure of the Frame production in which the integration of different sources of micro data is one of the crucial steps in the process.

In fact, the production process is articulated in a number of different phases or sub-processes. Schematically, each phase collects some input variables and produces some output variables. One way to find a common ground between different statistical actions is to focus on a generalized data input interface in which it is possible to identify and select the variables needed for data processing in each production phase. Adaptation of the data input and output interfaces of each phase of a process gives us the opportunity of managing the elaboration phases by using generic software components, i.e. using almost any statistical editing tool in a common application framework. The adaptability of the data input/output interfaces to the procedures are particularly helpful in statistical production based on administrative data when the input data layout and variable meanings are not under the direct control of the statistical producer, so that they can change for each supply due to national regulation changes.

In this way, the production of the Frame can be seen as a workflow of separated activities, which must be realized in a common environment where all the statistical experts involved in the different production phases can work. In such an environment the role of knowledge sharing is central and this is sustained by the S-DWH in which all information from the collaborative workflow is stored.

From an IT point of view this corresponds to a workflow management system able to sustain a data centric workflow of activities (scientific workflow[16]), i.e. a common software environment in which all the statistical experts (or data scientists), involved in the different production phases of the same process, work by testing hypotheses.

The Workflow management system then allows a controlled process through the standardization of working methods in a flexible environment. On the other hand, a scientific workflow based on a S-DWH can increase efficiency, reducing the risk of data loss and integration errors by eliminating any manual steps in data retrieval.

2.1 Data integration

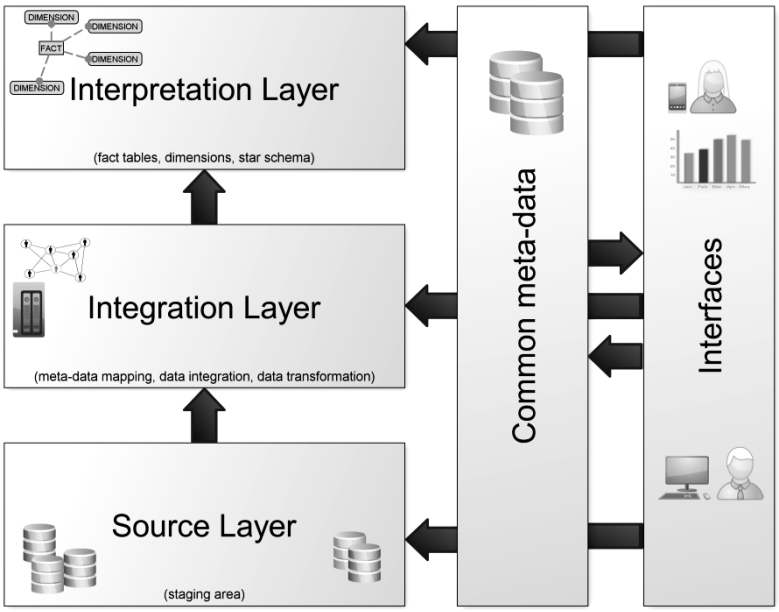
One of the most important activities in the process of building the SBS Frame is represented by *data integration*. This term refers to a set of activities aiming at creating a unifying view of data information coming from different and heterogeneous sources.

Often data integration problems are overcome through ad hoc approaches: each instance of the problem is treated case by case. In the Frame context, the key element to achieve this goal is given by a S-DWH.

[16] G. Scherp. A Framework for Model-Driven Scientific Workflow Engineering 05/2010 Procedia Computer Science. [17] N. Russell, A. Ter Hofstede, D. Edmond, W. van der Aalst. 2005. Workflow data patterns. In Proc. of 24th Int. Conf. on Conceptual Modeling Springer. Verlag: october

Statistical data warehousing systems extract, transform, and load data from different heterogeneous statistical data sources into a single schema so that data become compatible with each other and can be processed, compared, queried or analyzed regardless of the original data structure and semantics.

In classical DWH concept, it is first necessary to set up a database and a system that replicates data from production databases to the DWH's one. In addition to this, some tools to produce reports and view data are needed. All of the tools have to use and display metadata information.



The S-DWH is organized into three contiguous but essentially disjoint areas (the three boxes on the left in the picture above):

- the *source layer* contains the staging area. Here the supply of raw data from external sources takes place in the form of the *data provision* (which includes the set of elementary data, metadata, and classifications of a source for a given reference period). In this area also some preliminary consistency check operations are performed. This processing, called *provision acceptance*, allows to determine the completeness and the consistency of a given data supply;
  the *integration layer* deals with several Extraction Transformation and Loading (ETL) activities. Among them one can enumerate: finding and correcting inconsistent data, transforming data to standard formats in the DWH, classifying and coding, deriving new values, cleaning data and mapping the variables in the sources record layouts to the variables in the DWH dictionary. This layer, thanks to a suitable organization of metadata and thanks to the support of common metadata, allows decoupling the internal structure of the data sources from the interfaces. Thus, the analysts are able to

perform queries on the data sources based on the dictionary of the concepts of the DWH regardless of the structure of the single sources;

- the *interpretation layer* contains, in the form of dimensions and fact tables, the elementary data coming from the ETL operations performed in the previous layer. Only data related to variables of data provisions, that have been associated to the S-DWH dictionary items in the integration layer, can be accessed by analysts for data mining purposes.

The three layers are accessible by suitable interfaces through a layer of common metadata. As far as this layer is concerned, the S-DWH metadata consist of many logical elements. The first one is the data provision component which represents the part that provides information on the data source available for the subsequent elaborations. It is defined by the collection of data, metadata such as classifications, record layout and other objects. The second element is the Statistical dictionary, also called dictionary of meta-attributes, which contains the definitions of the statistical objects to be used in order to perform data analysis or data mining. Another important component of the common metadata architecture is the meta-source abstraction. It represents associations (mappings) between meta-attributes (i.e. the interpretation layer metadata) and elementary variables collected by the sources. It thus helps giving a support for interpreting the information of the data supply. Two of the main advantages of the mapping component are given by a direct possible association between information and statistical data dictionary and by different interpretations of the same variable as the content of the variable itself changes in time. In particular, the interpretation layer can be accessed by an interface through the metadata represented by the S-DWH statistical dictionary.

Differently from a classical DWH, in which data flow only from one layer to the subsequent one, the S-DWH system can play a key role in circular processes of microdata re-use in the following way:

- it is employed (through the interpretation layer) to extract linked raw or cleaned micro data for specific variables from the internal and external sources to be used as input for a new editing and imputation statistical process. This constitutes a data transformation process which takes place by an asynchronous elaboration and uses the S-DWH as input/output data repository;
- part of the output of the new statistical process, i.e. the one corresponding to validated micro-data (with associated metadata and classifications), goes to feed the S-DWH through the source layer. The dictionary of S-DWH (in the metadata layer) is then enriched with meta-attributes created during the new statistical process.

This circular process, in the framework of a S-DWH, sustains the production of the SBS Frame. This type of process can be supported using a blackboard design pattern's paradigm; i.e. a shared area (the blackboard) which can be accessed by autonomous processes or actors in some coordinated and cooperative way. Generally, this working scheme allows to achieve complex goals that require multidisciplinary skills.

The first step consists in setting up a S-DWH on the domain of Structural Business Statistics: it contains data coming from different kinds of sources, administrative data, registers and surveys, both internal and external to the National Statistical Institute (NSI).

Subsequently, all available information are analyzed in the S-DWH and the variables, needed for the process of construction of the SBS frame, are chosen from the S-DWH and

extracted for any required elaboration. In fact, the resulting data enter the statistical process of construction of the SBS frame and end up in an output, the Frame itself, which can be loaded in the S-DWH, becoming itself an internal source (besides being a byproduct of the integration of other sources).

More in detail, at the source layer level, one can find:

1. NSI surveys data;
2. NSI Statistical Business Register data;
3. Admin data.

In particular, the source layer of the S-DWH under implementation contains: the raw admin data coming from the Italian Revenue Agency (Income Tax Returns, "Studi di Settore") and from the Union of Chambers of Commerce (Financial Statements), the NSI Statistical Business Register (ASIA) data and the validated Small-Medium Enterprises survey data.

In the source layer some preliminary consistency checks between micro-data and record layout metadata are performed which make the data provisions ready to enter the subsequent layer for data integration. In the integration layer, after some ETL operations, data from the different sources are reconciled and mapped against common statistical concepts and subsequently loaded in fact and dimension tables in the interpretation layer. Here, data are thus ready for querying and analysis.

During the researchers' activities it is often necessary to inspect and view data in order to get detailed information on the survey units.

2.2 Process integration

In the previous paragraph some of the issues arisen in the construction of the SBS Frame have been discussed: in particular we focused on data and their integration. On the other hand, this paragraph deals with the integration of processes, tasks and all the elements needed for the Frame construction. Naturally, data integration and process integration are strongly intertwined.

The building process of SBS Frame can be defined as "scientific workflow" (SWF). This type of workflow is characterized by a strong dependence on data (data-centric WF), while a classic WF focuses more on processes than on the objects to manipulate (flow-centric WF). In fact, industrial (classic) WF is used when the process is not subject to substantial changes along its lifecycle and its number of instances (iterations) is large. Conversely, the scientific WF is characterized by frequently modified processes and by few instances.

One example of trial-and-error approach supported by SWF in the Frame construction, is the search of duplicated records in the UNICO source. Very often the sources are little documented or they are not documented at all (regarding to the statistics domain); therefore the only strategy researchers can use is to try some hypothesis on data. Subsequently it is necessary to test the hypothesis but effects could be observed only at the end of the sub-process or even afterwards. If the test succeeds, it is possible to move on; otherwise it is necessary to go back and try another strategy or modify some element of the existing one in order to remove all the duplicates.

Another great difficulty met in the production of the SBS Frame consists in collecting and organizing all the elements which concur to the results. As an example, every year

some administrative sources metadata change and therefore several methodological adaptations to processes and procedures could be necessary.
Another example could be represented by a variable semantic change: this kind of event affects not only the mapping of the variables, but also check and correction rules.
These two simple examples are enough to show that the SBS Frame production involves different heterogeneous skills and tools (each sub-process, particularly statistical ones, uses its own instruments). Besides the issues outlined above, the huge amounts of data and strict time constraints involved, make the building process of the SBS Frame particularly complex and critical to manage.

In order to efficiently organize the WF[17] with the aim to support the production processes and improve quality, it is necessary to connect several entities such as the source variables and the related documentation. It is also important to gather and record the versions of any entity in order to fully document the process and guarantee its quality, reliability, replicability and reusability. A systematic collection of all the tested attempts could also contribute to the production efficiency because the researcher's team would be able to examine all past discarded hypotheses.

In the next paragraphs, it is described the design of an integrated environment for setting up, executing and documenting the Frame production process. This is articulated in following item functionalities:

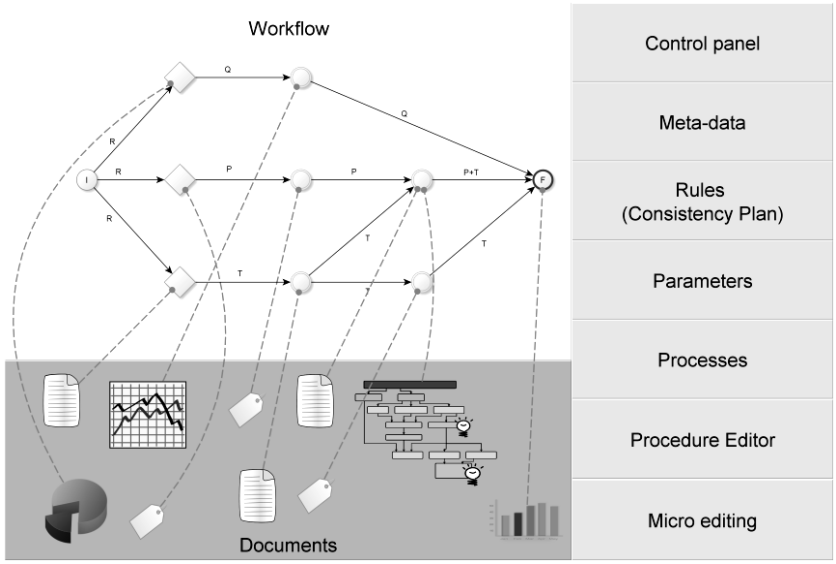
- *design and management of a statistical workflow*. It allows designing, modifying and executing the main phases, sub-processes and elementary activities which constitute the statistical production process;
- *activities and processes schedule.* The activities, the remote processes and the procedures can be run by a scheduler in an automatic way. This is particularly useful when one deals with huge amounts of data. The scheduler's purpose is to translate the workflow design into a sequence of activities to be submitted to the distributed processing nodes. This sequence has to satisfy priority constraints planned during the design phase;
- *local and remote services call.* Each elementary activity can be either a native procedure (e.g. a SAS procedure, a *PL/SQL* program or an R procedure) or an external service, such as a web service encapsulating a high-level domain service (i.e. *BANFF*) that can be invoked from the platform. It is necessary to provide some mechanism of sharing information between systems;
- *integration of statistical abstractions*. A statistical production process has its own rules, constraints, methodologies and paradigms. The aim of the statistical abstraction layer is to supply a set of abstractions that make the researcher's work flexible, independent of technical details and more focused on research objectives. Among the possible abstractions there could be:
    - *meta-parameters*: the use of global parameters reduces the need to modify the scripts and variables necessary for other systems to operate correctly;
    - *partitioning or filtering units*: each type of record (unit) has its own processing path in the WF. The value of some variable could be used to filter units to the

[17] N. Russell, A. Ter Hofstede, D. Edmond, W. van der Aalst. 2005. Workflow data patterns. In Proc. of 24th Int. Conf. on Conceptual Modeling Springer. Verlag: october

next processing step;
- *sampling test*: when the amount of data is very large, it is useful to test some hypothesis or programs on a subset of data in order to avoid loss of time and to early discover weak hypotheses;
- *rule checker*: a tool for finding inconsistencies in a formally defined set of rules and to manage efficiently semantic and definitional changes in sources;

- *documentation management and versioning. It is possible to associate one or more* documents and metadata to each WF element and, at any time, recall previous versions of the WF and all the elements connected.

The following picture presents a WF architecture that summarizes concepts and paradigms presented above.



The upper left box represents the key concept of the architecture. The WF is graphically represented by a custom graph and consists of sub-processes and domain modules. Each element can be linked to one or more documents (the bottom left box), for example: charts, reports, regulations, metadata and parameters, record layouts, text documents, etc. as well as information related to the execution of the sub-processes (actual and mean execution time, error and warning reporting, results, previous and subsequent sub-processes, etc.).

The right area contains "services" that the researchers could use to support most of their activities; in the following they are briefly described from top to bottom:

- *metadata* module implements a decoupling approach in data mapping. This type of abstraction introduces a new layer between data sources and statistical variables so that a semantic change in one administrative source does not affect statistical sub-processes that depend on the related statistical variables;
- *rules* module allow the researcher to write the consistency plan, check possible contradictions in the edits set, run the plan, log error and warnings and produce reports.

Moreover, this module assists the researcher in the activity of modifying an existing check plan in case some variables are introduced or deleted;

- *parameters* module is used to implement a basic form of parametric changes in all of the components of the WF. It can be thought as similar to a dashboard through which modifying thresholds, setting parameters, choosing elaborative units, switching on and off options, etc. For instance, suppose one parameter is shared by many sub-processes: a change in this value has an impact on all the sub-processes containing that parameter. The parameter is a placeholder that at runtime is set to the actual value (e.g. some sub-process can possibly change the parameters' value during processing);
- *processes* module provides information on actual state of active elaborations. It is possible to view the scheduled sequence of sub-processes and to recall the log of previous ones;
- *procedure editor* module is the development environment needed to create procedures or modify existing code. Such a module should support at least one statistical language (SAS, R) and one data manipulation language (PL/SQL). New languages can be added to this system in a modular and incremental way. The editor integrates a versioning system in order to restore a previous version of a procedure, document code changes and to monitor the improvements of the implemented functions;
- the *micro-editing* component is used in manual and interactive micro data editing activities. It can be a useful tool for statisticians to analyze some sample of micro-data.

Conclusions

In 2012 for the first time, ISTAT sent to Eurostat the SBS estimates using massively administrative data to derive the main SBS aggregates: this experience has shown that it is possible to have a lot of business information also without the use of direct surveys and if this information is appropriately supported by statistical methods it can be used to cover the most important aspects of enterprises' results. Of course the current process of Frame construction could get better: in particular, some steps could be improved using better and/or newer techniques, more efficient statistical methods or more generalized programs or software. But the confirmation of the quality of the results also for 2013 represents the most encouraging factor: accordingly, the SBS Frame is becoming the reference product for all the structural business statistics.

Furthermore, the analysis of the SBS Frame production system has shown the need to define a new informative infrastructure, which allows users to interact with a number of administrative data sources, and to yearly adapt and modify the Frame production procedures. The proposed infrastructure, that relies on concepts and paradigms of scientific workflows' management systems, involves the construction of an environment for process modelling, flexible and integrable with the standard statistical processing tools, and a data warehouse of microdata. The latter has been implemented for the specific domain of structural business statistics and allows to: manage the variable mapping, reconcile and follow definitional changes of the different sources in time, as well as build many customizable environments supporting the workflow process or data analysis.

References

Corsini V., T. Di Francescantonio, S. Filiberti, R. Sanzo. 2000. *Utilizzo integrato di fonti amministrative e fonti statistiche per la produzione di stime preliminari di alcuni principali aggregati economici previsti dal regolamento della Ue n.58/97 sulle statistiche strutturali*. ISTAT (Documento interno UDAS.10.00.3).

Dabbicco G. 2002. *Utilizzo dei bilanci aziendali civilistici ai fini del soddisfacimento del regolamento UE 58/97 sulle statistiche strutturali sulle imprese*. ISTAT (Documento interno UDAS.03.01.1).

De Carli R. 2003. *Integrazione tra dati statistici e dati amministrativi sui risultati economici delle imprese: prime evidenze dal confronto tra i dati individuali delle principali rilevazioni statistiche strutturali, di quelli desumibili dai bilanci civilistici e di quelli derivanti dalle dichiarazioni fiscali*. ISTAT (Documento interno UDAS).

Monducci R., G. Dabbicco, C.M. De Gregorio, T. Di Francescantonio, S. Filiberti, U. Sansone, R. Sanzo, A. Volpe Rinonapoli. 2003. "Prime esperienze sull'utilizzo integrato di fonti statistiche ed amministrative per la produzione di statistiche strutturali sui risultati economici delle imprese". In *Temi di ricerca ed esperienze sull'utilizzo a fini statistici di dati di fonte amministrativa*, a cura di P.D. Falorsi, A. Pallara, A. Russo. Milano: Franco Angeli.

Siesto G., F. Branchi, C. Casciano, T. Di Francescantonio, P.D. Falorsi, S. Filiberti, G. Marsigliesi, U. Sansone, E. Santi, R. Sanzo, A Zeli. 2006. *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI.* ISTAT (Documenti, 17).

Grazzi M., R. Sanzo, A. Secchi, A Zeli. 2009. *MICRO 3-ISTAT: A new integrated system of business micro-data 1989-2004*. ISTAT (Documenti, 11).

Grazzi M., R. Sanzo, A. Secchi, A Zeli. 2013. "*The building process of a new integrated system of business micro-data 1989-2004"*. *Journal of Economic and Social Measurement,* 38: 291-324.

Sanzo R., D. Bellisai, T. Di Francescantonio. 2014. "Il processo di produzione del FRAME". Relazione presentata al Workshop scientifico: "Il nuovo «frame» delle statistiche sulle imprese: innovazioni metodologiche, uso delle fonti, risultati e potenziale informativo", Roma, 25 marzo 2014.

Buschmann F., R. Meunier, H. Rohnert, P. Sommerlad, M. Stal. *Pattern-Oriented Software Architecture, A System of Patterns*.

Inmon Bill. 1992. *Building the Data Warehouse*. Wiley.

Scherp G. *A Framework for Model-Driven Scientific Workflow Engineering.* 05/2010 Procedia Computer Science.

Russell N., A. Ter Hofstede, D. Edmond, W. van der Aalst. 2005. *Workflow data patterns.* In Proc. of 24th Int. Conf. on Conceptual Modeling Springer Verlag: October 2005.

Norme redazionali

La Rivista di statistica ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche corredati da una nota informativa dell'autore contenente attività, qualifica, indirizzo, recapiti e autorizzazione alla pubblicazione. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di due referenti scelti tra gli esperti dei diversi temi affrontati.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file RSU stili o nella classe LaTex, entrambi disponibili on line. La lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 35 pagine. Una volta che il lavoro abbia superato il vaglio per la pubblicazione, gli autori sono tenuti ad allegare in formato originale tavole e grafici presenti nel contributo, al fine di facilitare l'iter di impaginazione e stampa. Per gli standard da adottare nella stesura della bibliografia si rimanda alle indicazioni presenti nel file on line.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 120 parole); quelli in italiano dovranno prevedere anche un abstract in inglese.

Nel testo dovrà essere di norma utilizzato il corsivo per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare la redazione o per inviare lavori: rivista@istat.it. Oppure scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
all'attenzione di Gilda Sonetti
Istat
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012016001000000

ISSN 1828-1982

€ 15,00