A quality evaluation framework for the statistical register *Frame-SBS*

Orietta Luzi, Fabiana Rocci, Roberto Sanzo, Roberta Varriale [1]

Abstract

In 2013, Istat implemented the new statistical register "Frame-SBS" for the annual production of economic accounts statistics based on the integrated use of administrative and survey data, overcoming most of the limits of the traditional survey-based estimation strategy. The transition to a production strategy essentially based on the use of administrative data required the development of innovative methodological approaches, and determined the need of new tools for quality evaluation of both the data and the statistical process. In this paper we propose a first scheme of indicators for measuring and documenting the quality of the Frame-SBS. The final goal is to implement a quality control system to regularly monitor the register, by the identification of possible process and data weaknesses, and supporting quality improvements.

Keywords: Statistical Register, Administrative data, Quality.

1 Orietta Luzi (luzi@istat.it); Fabiana Rocci (rocci@istat.it); Roberto Sanzo (sanzo@istat.it); Roberta Varriale (varriale@istat.it), Italian National Institute of Statistics - Istat.
The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction

In the last years, Istat has strongly increased the amount of administrative (hereafter *admin*) archives that are centrally acquired and used for statistical production purposes. Such an increase calls for a tailoring of the current approaches for quality measurement and assessment, in order to build a wider framework based on: the measurement of the quality of input sources, that are centrally acquired by Istat (Ambroselli *et al.*, 2014); designing proper tools to extend quality auditing to the statistical processes using admin data (Brancato *et al.*, 2014); measuring, monitoring and assessing the quality of any statistical process and product derived by using admin data, which is the main aim of this paper. One of the most common output based on an intensive use of admin data is the establishment of statistical registers, representing the transformation of the admin information for statistical purposes, according to the statistical definition of the target population and variable (Wallgren *et al.*, 2007).

This paper deals with the quality assessment of the statistical register *Frame-SBS*, (Luzi *et al.*, 2016; Luzi *et al.*, 2014) which is currently used at Istat for the annual estimation of Structural Business Statistics (hereafter *SBS*). Its implementation has been guaranteed by the use of a number of admin sources integrated in an appropriate strategy to survey data. Hence, the availability of stable, timely and reliable admin sources providing high quality and detailed information on enterprises' profit and loss accounts, has allowed since 2013 Istat to use a new estimation strategy. The *Frame-SBS* contains microdata for the main economic variables for all the enterprises in industry and services (excluding financial companies and insurance) with less than 100 persons employed which are active for more than six months in the reference year (about 4.4 million of units), for every SBS domain required by the European Regulation.

Therefore, based on the *Frame-SBS*, estimates for the main SBS can be computed at an extremely refined level of detail, overcoming some limitations of the previous estimation strategy. As a consequence, improvements have been achieved in terms of both accuracy of cross-sectional estimates and consistency of estimates over time and among related statistical domains, with particular reference to National Accounts. Concerning accuracy, however, it has been underlined that even if sampling error components have been essentially removed, additional sources of non-sampling error need to

be assessed due to the admin data characteristics and coverage and to the features of the integration process.

The present work focusses on the definition and the implementation of the quality framework to assess the *Frame-SBS* production process, starting from the framework proposed by Zhang (2012). First considerations concerning the suitability of the Zhang proposal with respect to the Istat experience are also reported. In particular, in the paper a first application of the proposed quality framework is reported, with the introduction of an additional step to better deal with the admin sources integration phase. The focus is on the main register variables, that are those which can be directly derived by the admin sources with a high level of quality and coverage (see Curatolo *et al.*, 2016).

The paper is structured as follows. Section 2 contains a description of the main characteristics of the *Frame-SBS* register. In Section 3, the proposed quality framework associated to the *Frame-SBS* production process is illustrated, and the corresponding list of quality indicators is proposed. Some results from the established framework system are also provided, to show how the monitoring is currently assessed year by year. In section 4 some concluding remarks are provided and the directions for further developments are delineated.

2. The *Frame-SBS*

In this section we describe the main features and the production process of the *Frame-SBS*.

The target population of the register consists of all the Italian small and medium enterprise (enterprises with less than 100 persons employed) in the industrial, construction, trade and non-financial services sectors (about 4.3 million of units) which are active for more than six months in the reference year, for every SBS domain required by the European Regulation. This population is completely identified by the Italian Business Register (BR) ASIA[2] (Istat, 2016) which contains structural and classification information on the Italian active enterprises. Actually, ASIA allows to identify all the potential theoretical sub-populations (e.g. by legal form) which could be investigated for statistical purposes.

The *main* target variables of the register are the profit and loss account variables as identified by the E.U. regulation:

- Revenues
  - Income from sales and services (Turnover)
  - Changes in stock of finished and semi-finished products
  - Changes in contract work in progress
  - Changes in internal work capitalised under fixed assets
  - Other income and earnings (neither financial, nor extraordinary)
- Costs
  - Purchases of goods
  - Purchases of services
  - Use of third party assets
  - Changes in stocks of raw materials and for resale

2 The Italian Business Register represents the official source on the structure of the business population and demography that identifies the Italian enterprises, and their statistical variables. Asia has the role of the frame list for all Istat business survey. It is also a reference to update structural information on enterprises (economic activity, persons employed, employees, etc.) and allows linking all the available administrative sources through the fiscal code.

- Other operating charges
- Personnel Costs.

In order to estimate the target variables[3], in *Frame-SBS* micro-data from different admin data sources available in the Italian information system are properly integrated. Such sources are currently acquired by Istat through a unique entry point ensuring a standardised and consistent management of the relationships with data owners. The sources are (Curatolo *et al.*, 2016):

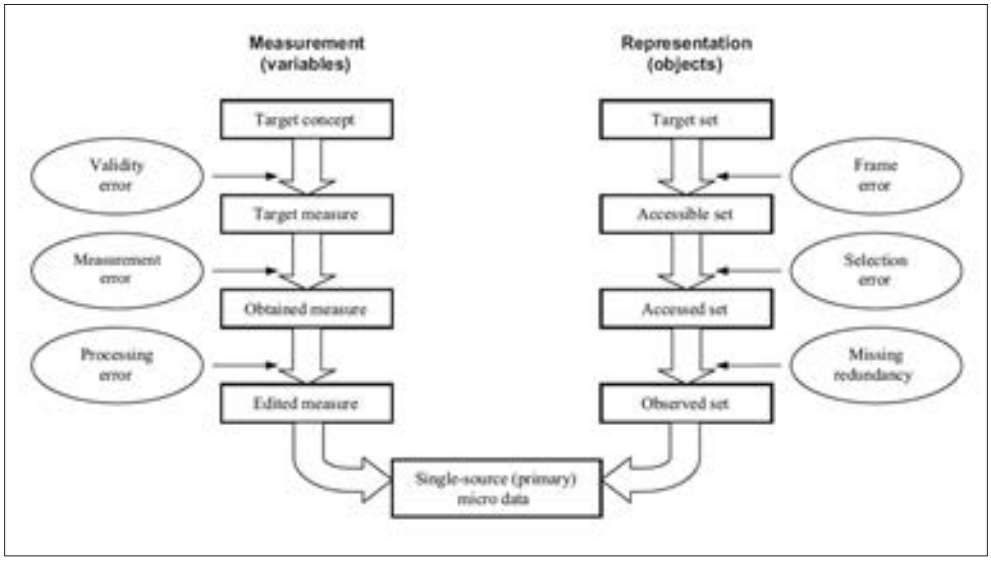
- Financial Statements (hereafter *FS*). FS are registered by the Italian Chambers of Commerce. Profit and loss account items of the financial statements are annually provided for limited liability companies (about 750,000 units). Variable definitions in FS, which are designed to check the balance sheet of corporate companies, are the closest to those required by SBS regulations. For this reason, this source plays a central role in the integration process described in the following;
- Sector Studies survey (hereafter *SS*). SS is a Fiscal Authority survey, including each year about 3.5 million of units, that aims at evaluating the capacity of enterprises to produce income and at indirectly assessing whether they pay taxes correctly. The units compiling the SS 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 returns (hereafter *Modello Unico*). The Modello Unico data is provided by the Ministry of Economy and Finance, is based on a unified model of tax declarations by legal form and containins economic information for different legal forms for about 4.5 million of units each year;
- Regional Tax on Productive Activities (hereafter *Irap*). 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 in accordance with the different type of the enterprises.

3 The variable personnel costs is always observed and it is used as auxiliary variable in the estimation process.

3. The proposed quality evaluation framework, first considerations based on *Frame-SBS* study

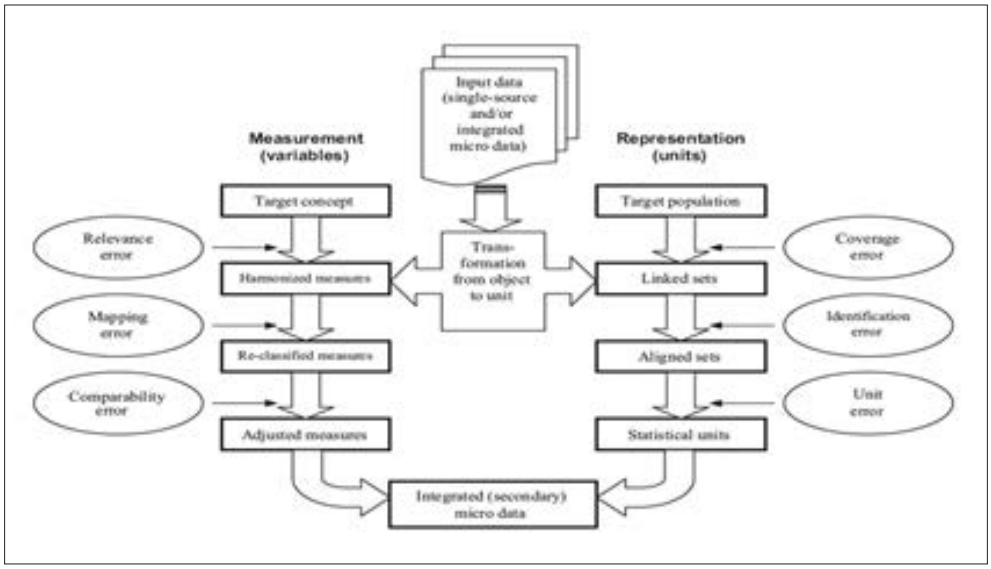
In order to assess the quality of the *Frame-SBS* data and, indirectly, of the statistics produced based on its data, we adapted the framework proposed by Zhang (2012), where a well-defined data processing scheme with the associated list of errors for the production of statistics based on the combination of various admin and statistical datasets is presented. The framework consists of two main phases, represented in the lifecycle diagrams reported in Figure 3.1 and Figure 3.2. The first phase, dealing with each single source, categorizes errors arising with respect to the original source's target population and concepts, in order to support the assessment of the quality of the source itself.

Figure 3.1 - Sources of error in phase one of Zhang's framework



Source: Zhang, 2012

Figure 3.2 - Sources of error in phase two of Zhang's framework



Source: Zhang, 2012

The second phase focusses on errors arising when data from several sources are combined to produce a statistical output. In this case, the aim is to measure the quality of the transformation process which is needed to adapt the data from their original purpose to the statistical one. Indeed, in this phase the targets correspond to the statistical population and to the statistical concepts to be measured. For more details see Zhang (2012) and Zabala (2013).

3.1 The *Frame-SBS* case study

In this paragraph the *Frame-SBS* production process is described. We start from the Zhang's framework, which is actually useful in order to clearly analyze the design of any mixed-source statistical process. The final aim is to understand the error sources potentially affecting the register's output data, that may result from the characteristics of each admin archive and/or from the design choices underlying the statistical production process.

Nevertheless, we propose to represent the process in three-phases: the first phase can be assimilated to the Zhang's phase one, while the Zhang's second phase has been split into two sub-phases in order to better distinguish the

specific steps of the transformation process the original data have to go through: in phase two the admin data are evaluated according to the SBS targets (both units and variables), however we define a first sub-phase (phase 2a), where each admin source is evaluated separately in order to determine the criteria according which to select the data and to combine them, and a second phase (phase 2b), where the integrated dataset is created and is further elaborated to attain the final register data.

Phase 1. Pre-treatment of admin sources. The first phase of the *Frame-SBS* production process consists of the pre-treatment of each admin source data. This phase is carried out separately for every source, covering each a different population and characterised by a peculiar structure and specific contents. Firstly, only the subset of items which are useful for deriving the target SBS are selected. On the objects side, for each admin source, the following actions are performed: verify if there are substantial changes over time in the population coverage and in the time of the source supply, identify and eliminate duplicated units or unacceptable information. On the measurements side, an initial assessment of formal data inconsistencies is carried out, based on the use of accounting rules (edits). At this stage, a proportion of FS units containing errors that cannot be resolved are discarded. The remaining errors are resolved by adopting a deterministic data imputation approach.

Phase 2a. Treatment of the admin sources, taking into account the SBS purposes. During phase 2a, the units belonging to the SBS population are selected from each source. Note that the statistical units in each source are identified at the archive acquisition stage from the external supplier, therefore units identification errors are not expected in the *Frame-SBS* production process. The admin (original) items of each source are harmonised w.r.t. the target SBS variables. The harmonisation process is a result of accurate preliminary analyses of admin data and their associated metadata, with the aim of comparing the economic contents derived from the admin items with the corresponding SBS definitions, as described by the SBS European regulation (Curatolo *et al.*, 2016). As it is not always possible to directly "reconcile" the admin and the statistical definitions, the admin information is used to obtain the harmonised variables, however a certain amount of information is

discarded and this causes a given amount of "item non response". Finally, the source coverage w.r.t. the target population is evaluated and the information content of the entries in the various admin sources are assessed. As a direct consequence of this assessment, different degrees of reliability are associated to the different admin sources, and a pre-defined priority is associated to each archive so that the best source is used for each target (sub)population in case of overlaps.

Phase 2b. Integration of the sources. In this phase, the final list of the units belonging to the target population is identified (based on the BR identification code) and a specific admin source is associated to each of them, following the predefined priority in case of concurrent (overlapping) sources. For each statistical unit all information from a single source (when available) is derived, to preserve the internal data consistency at unit level. There are some "exceptions to the priority", according to which the most reliable source is discarded and the source with next priority is used. For example, in case of inconsistencies resulting from the pre-treatment of each source (phase 1) that cannot be resolved. Another exception is based on the analysis of the *per capita* (per employee) labor cost of the enterprises, that when not coherent with auxiliary information available from the Istat Employee Wage Register (RACLI), may determine the selection of the units from the source with next priority. Once the above process is completed, an integrated dataset of target units and variables is determined. However, a certain amount of both under-coverage w.r.t. the SBS target population, and incompleteness w.r.t. SBS target variables remain, to be properly recovered. Therefore, after an editing activity aiming at identifying and treating possible outliers and influential errors, an imputation process to predict unit and item non-responses on the integrated data is performed (Di Zio *et al.*, 2016). A macro-editing strategy is used for the final cross-sectional and longitudinal validation of the final SBS estimates at the level of detail required by the Eurostat regulation.

For each phase of the *Frame-SBS* production process, in the Tables 3.1, 3.2 and 3.3 we propose a set of quality indicators consisting of both new measures and some adaptation of the indicators proposed by Zabala (2013). For each process phase, the indicators are presented by subject (variables, objects and units), process step and error type (as reported in Figure 3.1 and Figure 3.2).

Table 3.1 - Phase 1 quality indicators

Objects. Accessible Set -> Accessed Set; Selection error	
Proportion of *units* in FS w.r.t. the FS theoretical population in the BR	**[*No. units in the source/ Total no. units in the FS theoretical population in BR*] x 100**
Proportion of *units* in the source w.r.t. the BR population, by source (SS, Unico, Irap)	**[*No. units in the source/ Total No. units in BR*] x 100**
Adherence to reporting period, for FS	***[No. units that do not adhere to the reporting period/Total No. units]* x 100**
Qualitative indicators, by source (SS, Unico, Irap)	*Changes in population coverage (Does coverage change over time?)* *Updating of reporting units (How are changes recorded and actioned? Is it proactive or reactive?)*
Objects. Accessed Set -> Observed Set; Missing/Redundancy error	
Percentage of multiple records, by source	***[No. units in Source S with multiple id code / No. of unique identification codes] x 100***
Qualitative indicators	*Detecting duplicate records (Describe how duplicate reporting units are identified)* *Methods of treating duplicate records (Describe how duplicate reporting units are handled)*
Variables. Process step: Target Measure -> Obtained Measure; Type of error: Measurement error	
Punctuality, by source	***[Date of receipt - date agreed]***
Lagged time between reference period and receipt of data	***[Date of receipt by Istat-Date of the end of the reference period over which the data provider reports]***
Qualitative indicators, by source	*Changes in administrative forms*
Variables. Obtained Measure -> Edited Measure; Processing error	
Proportion of *units* failing edit checks, by source	***[No. units failing edit checks/ Total no. of units checked]* x 100**
Proportion of *units* with all implausible values, by source	***[No. units whose values are all missing, or all values are equal to 0, or all values are equal to 1 / Total no. of units checked]* x 100**
Proportion of *units* with all missing values, by source	***[No. units with all values missing/ Total n. of units checked]* x 100**
Proportion of edit rules failed at least once, by source	***[No. of failed edit rules for source S/ Total no. of edit rules for source S]* x 100**
Proportion of imputed *values*, by source	***[Total no. of imputed values in source S / Total no. of values in source S]* x 100**
Composition of the proportion of imputed *values*, by source	***Modification rate: [Total no. of values changed from a code to another code in source S / Total no. of imputed values in source S] x 100*** ***Net imputation rate: [Total no. of values changed from missing or 0 to a code in source S / Total no. of imputed values in source S] x 100*** ***Cancellation rate: [Total no. of values changed from a code to 0 in source S / Total no. of imputed values in source S] x 100***

The proposed indicators include both quantitative and qualitative measures. Actually, for some types of errors (e.g. *Measurement errors* in phase 1, Relevance errors and *Mapping errors* in phase 2a), the description of the conceptual schemes developed provides key information for the assessment of the quality of the production process. The indicators proposed for phases 1 and 2a are typical of all statistical processes based on the integrated use of admin data. The most part of indicators proposed for variables in phase 2b, on the other hand, are similar to measures which are typically used to assess the quality of data collected by direct surveys.

Table 3.2 - Phase 2a quality indicators

Units. Target Population -> Linked Sets; Coverage error	
Proportion of *units* in the FS source w.r.t. the SBS sub-population of corporate companies	***[No. corporate companies of SBS pop. in source FS/ No. of corporate companies of the SBS pop.]* x 100**
Proportion of *units* in the source w.r.t. the SBS population, by source (SS, Unico, Irap)	***[No. units of SBS population in source S / No. of units of SBS population]* x 100**
Variables. Target Concept -> Harmonised Measures; Relevance error	
Qualitative indicators, by source	*Changes in definitions of all variables in each source and changes in definitions of SBS variables (Does definitions change over time?)* *Conceptual scheme representing the re-classification of administrative concepts needed to produce the SBS variable definitions*
Variables. Harmonised Measures -> Re-classified Measures; Mapping error	
Quantitative indicators, by source	*Comparison of each harmonised variable with SBS benchmark variable (histograms, univariate statistics, statistical tests, etc.), to be repeated when variable definitions change*
Proportion of target variables which not require reclassification or mapping, by source	***[No. variables captured directly from source S / Tot. no. variables]* x 100**
Proportion of target variables which can be derived through reclassification or mapping, by source	***[No. variables derived from source S after reclassification/ Tot. no. variables]* x 100**

Table 3.3 - Phase 2b quality indicators

Units. Target Population -> Linked Sets; Coverage error	
Proportion of *units* of the SBS population in the integrated dataset (coverage). Also in longitudinal perspective.	***[No. of units of SBS pop. in the integrated dataset / No. of units of SBS pop.]* x 100**
Proportion of *units* of the SBS population in the integrated dataset, by source S.	***[No. of units of SBS pop. in the integrated dataset from source S/ No. of units of SBS pop.]* x 100**
Proportion of *units* of the SBS population in the integrated dataset with information present in only one source	***[No. of units of SBS pop. in only one source / No. of units of SBS pop. in at least one source]* x 100**
Proportion of *units* of the SBS population in the integrated dataset with information available in more than one source	***[No. units of SBS pop. in more than one source / No. of units of SBS pop. in at least one source]* x 100**
Variables. Re-classified Measures -> Adjusted Measure; Comparability error	
Proportion of *units* with influential values, by variable	***[No. of units with influential errors / Total no. of units]* x 100**
Proportion of outliers, by variable	***[No. of outliers / Total no. of units]* x 100**
Proportion of *units* with at least one imputed value	***[No. of units with at least one imputed value / Total no. of units]* x 100**
Proportion of *units* failing at least one edit rule	***[No. of units failing edit checks / Total no. of units checked]* x 100**
Proportion of *variable values* imputed, by variable	***[No. of units with imputed values for variable Y / Total no. of unit]* x 100**
Composition of the proportion of imputed *variable values*, by variable	***Modification rate: [Total no. of values of the variable Y changed from a code to another code in source S / Total no. of imputed values of variable Y]* x 100** ***Net imputation rate: [Total no. of values of the variable Y changed from missing or 0 to a code / Total no. imputed values of variable Y] x 100*** ***Cancellation rate: [Total no. values of the variable Y changed from a code to 0 /Total no. of imputed values of variable Y] x 100***
Impact of data editing and imputation on microdata, by variable	***Simple and quadratic distance between pre-edited (Y) and post-edited (Y*) values of variable Y*** $DL_1(Y_i, Y_i^*) = S^N_{i=1} \lvert Y_i - Y_i^* \rvert$ */Total no. of units N;* $DL_2(Y_i, Y_i^*) = \ddot{O}S^N_{i=1} (Y_i - Y_i^*)^2$ */Total no. of units* N_i
Impact of data editing and imputation on distributions, by variable	*Kolmogorov-Smirnov distance on pre-edited and post-edited distributions* *Comparison of variable distributions (univariate statistics, etc.) pre- and post- editing and imputation*
Impact of data editing and imputation on statistical relations	*Pearson correlation index, Covariance matrix between variables*
Impact of data editing and imputation on aggregates, by variable	***[Variable total before editing and imputation / Variable total after editing and imputation]* x 100**

3.2 Selected results

In this section, we provide an example of how the quality indicators included in the proposed evaluation framework can be used for the analysis of the *Frame-SBS* inputs, data processing and outputs. It is straightforward to mention that the availability of the indicators values for subsequent years allow longitudinal analyses in order to monitor the changes of the quality of both input and output data.

In Tables 3.4, 3.5 and 3.6 the values of a selected set of qualitative measures are reported for three reference years (2012, 2013 and 2014), for the designed phases of the *Frame-SBS* production process.

As it can be seen in Table 3.4, referring to *Objects: Selection error*, Unico is the archive with the lowest under-coverage rate w.r.t. its corresponding theoretical population. In particular, the under-coverage of FS w.r.t. its theoretical population (the Italian *corporate companies*) is essentially due to delays in the delivery of information to the Italian Chamber of Commerce by some of the enterprises, and to the fact that some deadlines for enterprises to supply their data are not compatible with the production of the register.

Concerning *Variables*, the proposed indicators relate to validation rules which identify within-records data inconsistencies with respect to the specific admin data coherence requirements. Note that a different number of rules has been defined to check the formal accuracy of data in the used sources[4]. From Table 3.4 it can be viewed that FS and SS have the highest quality in terms of proportions of units with all missing or implausible values. However, FS is the archive with the highest rate of edit rules failed at least once, while SS is the source with highest quality w.r.t. formal accounting rules. Very low proportions of imputed values result for all the sources involved in the production process, as imputation is performed on data after the elimination of the admin units containing unusable information (units with all missing values and units with all implausible values – missing, zero and 1).

4 FS: 29 edit rules defined (23 failed rules at least once in 2013); SS: 108 edit rules defined (40 rules failed at least once in 2013); Unico: 178 edit rules defined (52 rules failed at least once in 2013); Irap: 124 edit rules defined (26 rules failed at least once in 2013).

Table 3.4 - Phase 1, quality indicators by subject and error type. Years 2012, 2013 and 2014

INDICATOR	Year		
	2012	2013	2014
Objects. Selection error			
Proportion of units not in the source w.r.t. the theoretical population, by source			
FS	8.43	10.55	11.39
SS	12.80	12.55	10.60
Unico	4.48	5.52	6.39
Irap	22.62	22.17	26.00
Objects. Missing/Redundancy error			
Percentage of multiple records, by source			
FS	0.01	0.01	0.11
SS	0.00	0.00	0.00
Unico	2.24	2.13	2.03
Irap	2.23	1.21	0.95
Variables. Processing errors			
Proportion of units failing edit checks, by source			
FS	6.41	6.30	4.35
SS	0.01	0.00	0.00
Unico	18.46	0.68	0.59
Irap	0.01	10.88	10.56
Proportion of units with all missing values, by source			
FS	0.00	0.00	0.00
SS	0.00	0.00	0.00
Unico	1.42	16.57	0.01
Irap	1.19	12.55	0.03
Proportion of units with all implausible values, by source			
FS	0.01	0.01	0.01
SS	0.19	0.26	0.28
Unico	0.10	0.38	0.38
Irap	0.00	0.52	0.47
Proportion of edit rules failed at least once, by source			
FS	79.31	79.31	79.31
SS	-	37.04	40.74
Unico	-	29.21	28.73
Irap	0.01	20.97	15.45
Proportion of imputed values, by source			
FS	0.15	0.14	0.33
SS	0.25	0.00	0.00
Unico	0.41	0.01	0.01
Irap	0.00	0.40	0.39

Concerning phase 2a (Table 3.5), it results a high coverage rate of FS w.r.t. the SBS sub-population of corporate companies, as well as high coverage rates of the other sources SS, Unico and Irap w.r.t. the SBS target population.

Relating to variables, the *Variables: Mapping error* indicator is provided w.r.t. 20 key SBS, and taking into account that some of the sources are structured in multiple forms[5] corresponding to different classes of enterprises (e.g. different business legal forms) and providing different information accordingly (see Curatolo *et al.*, 2016, for more details). It is evident from the *mapping error* indicator that FS is the best harmonised source in terms of variables definitions w.r.t. the SBS estimation purposes. As expected, this indicator does not vary in the considered period: the variables definitions adopted for admin purposes did not change.

Table 3.5 - Phase 2a quality indicators by subject and error type. Years 2012, 2013 and 2014

INDICATOR	Year		
	2012	2013	2014
Units. Coverage error			
Proportion of units in the FS source w.r.t. the SBS sub-population of corporate companies	90.84	90.57	89.81
Proportion units in the source w.r.t. the SBS population, by source			
SS	79.99	80.84	80.11
Unico	77.57	78.30	74.54
Irap	95.42	94.76	93.91
Variables. Mapping errors			
Proportion of target variables which not require reclassification or mapping, by source			
FS	100.0	100.0	100.0
SS	86.0-90.0	86.0-90.0	86.0-90.0
Unico	6.0-73.0	6.0-73.0	6.0-73.0
Irap	25.0-80.0	25.0-80.0	25.0-80.0

Concerning phase 2b, in Table 3.6 the values of indicators on *coverage* in the integrated dataset are provided. Overall, the coverage of the target SBS population is about 97%, with the most part of information for each unit available from more than one source (about 93%). It has to be reminded that in the *Frame-SBS* the sources are used with a pre-defined priority, based on a preliminary assessment of the different quality levels of their information (see Curatolo *et al.*, 2016 for more details).

As it can be seen, SS is the source with the highest contribution in terms of proportion of units in the integrated dataset.

5 SS involves 2 different forms; Unico involves 8 different forms; Irap involves 8 different forms.

Table 3.6 - Phase 2b quality indicators by subject and error type. Years 2012, 2013 and 2014

INDICATOR	Year		
	2012	2013	2014
Units. Target Population -> Linked Sets; Coverage error			
Proportion of missing units of the SBS population in the integrated dataset (under-coverage)	2.50	2.63	3.76
Proportion of units of the SBS population in the integrated dataset, by source			
FS	16.17	16.87	16.88
SS	67.26	67.67	67.05
Unico	12.26	10.80	11.07
Irap	1.80	2.03	1.23
Variables. Re-classified Measures -> Adjusted Measure; Comparability error			
Proportion of units with at least one imputed value	19.95	19.05	24.59
Proportion of variable values imputed, by variable			
Revenues	2.78	2.74	7.84
Purchases goods&services	13.44	12.88	16.44
Value Added	10.96	10.56	9.68
Modification rate, by variable			
Revenues	0.00	0.00	3.95
Purchases goods&services	5.25	6.01	6.01
Value Added	8.20	7.72	5.72
Net imputation rate, by variable			
Revenues	2.78	2.74	3.89
Purchases goods&services	8.19	6.87	10.37
Value Added	2.75	2.84	3.97
Cancellation rate, by variable			
Revenues	0.00	0.00	0.00
Purchases goods&services	0.00	0.00	0.00
Value Added	0.00	0.00	0.00
DL_1 (Impact of data editing and imputation on microdata), by variable			
Revenues	10,377	8,781	16,339
Purchases goods&services	8,402	7,954	13,194
Value Added	4,236	4,063	5,432
DL_2 (Impact of data editing and imputation on microdata), by variable			
Revenues	592,973	482,945	2,497,389
Purchases goods&services	449,541	431,047	1,652,552
Value Added	294,411	299,485	550,086
Kolmogorov-Smirnov Index (Impact of data editing and imputation on distributions), by variable			
Revenues	0.03	0.03	0.04
Purchases goods&services	0.08	0.07	0.10
Value Added	0.03	0.03	0.04
Impact of data editing and imputation on aggregates, by variable			
Revenues	102.70	102.30	104.30
Purchases goods&services	102.60	102.50	104.50
Value Added	102.30	102.40	103.90

4. Conclusions and future work

In this paper a comprehensive framework for the quality assessment of the statistical register *Frame-SBS* on enterprises accounts is proposed. In the definition of the framework, an effort has been made to adapt the proposals from Zhang (2012) and Zabala (2013) to the peculiarities of the register production process, in order to identify the actual sources of errors by using appropriate quality measures on both the variables and the objects/units sides. In fact, the identification of the error sources represents the basis for the systematic and continuous improvement of the production process quality through their elimination (or at least the reduction) in the subsequent replications of the production process of the register. Furthermore, the availability of such indicators for different reference years allow the analysis of both data and process quality in a longitudinal perspective. In addition, based on the proposed framework, a complete quality report could be developed for documentation and dissemination purposes.

Concerning future work, it has to be remarked that this proposal has to be considered as an initial step of a complex project. An in depth analysis of the proposed set of indicators is necessary in order to fine tune it, in order to eliminate the possible redundancies and potentially add new indicators. Concerning the latter aspect, additional quality measures could be defined as a consequence of the possible extension of the admin sources used and the detection of further sources of error. Furthermore, as imputation models are used in phase 2b to compensate for not available information, an evaluation of their impact on final estimates should be provided, e.g. by adopting iterative procedures (based e.g. on bootstrapping or on multiple imputation) to measure the additional uncertainty due to the imputation process, under appropriate assumptions on the missing data mechanisms.

It has to be underlined that the proposed framework needs to be harmonised with the tools under development at Istat for the quality documentation of all the admin sources acquired from external suppliers: in particular, the indicators on "input data" included in the so called *source quality card* associated to each admin archive acquired by Istat, have to be properly incorporated in the quality evaluation framework proposed in this paper.

Finally, a relevant development relates to the need of identifying appropriate combinations of the proposed quality indicators (e.g. by using composite indicators) in order to have a complete representation of the overall quality of the register data and of its production process.

References

Ambroselli, S., and G. Di Bella. 2014. "Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat". *European Conference on Quality in Official Statistics (Q2014)*. Wien, 2-5 June 2014.

Brancato, G., A. Boggia, F. Barbalace, and C. Busetti. 2014. "Quality Guidelines for statistical processes using administrative data". *European Conference on Quality in Official Statistics (Q2014)*. Wien, 2-5 June 2014.

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

Di Zio, M., U. Guarnera, and R. Varriale. 2016. "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.

Istituto Nazionale di Statistica - Istat. 2016. "Il Registro statistico Asia-Imprese". *Nota metodologica*. Roma: Istat. https://www.istat.it/it/files//2016/06/Nota-metodologica-1.pdf

Luzi, O., and R. Monducci. 2016. "The new statistical register *Frame-SBS*: overview and perspectives". *Rivista di Statistica Ufficiale*, N. 1/2016.

Luzi, O., U. Guarnera, and 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)*. Wien, 2-5 June 2014.

Wallgren, A., and B. Wallgren. 2007. *Register-based Statistics: statistical methods for administrative data*. Chichester, U.K.: John Wiley & Sons Ltd.

Zabala, F., G. Reid, J. Gudgeon, and M. Feyen. 2013. "Quality Measures for Statistical Outputs using Administrative Data". *Statistical Methods*. Statistics New Zealand.

Zhang, L.-C. 2012. "Topics of statistical theory for register-based statistics and data integration". *Statistica Neerlandica*. Volume 66, Issue 1: 41-63. Hoboken, NJ, U.S.: John Wiley & Sons, Inc.