European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Quality evaluation for statistical register: the Italian FRAME-SBS

Orietta Luzi[1], Fabiana Rocci[2], Roberto Sanzo[3], Roberta Varriale[4], Giovanna Brancato[5]

[1] *Istat, Rome, Italy; luzi@istat.it*
[2] *Istat, Rome, Italy; rocci@istat.it*
[3] *Istat, Rome, Italy; sanzo@istat.it*
[4] *Istat, Rome, Italy; varriale@istat.it*
[5] *Istat, Rome, Italy; brancato@istat.it*

Abstract

At the Italian National Statistical Institute (Istat), the direct use of administrative data for estimating business statistics has progressively increased, stimulated by the augmented availability and quality of secondary data on both private and public businesses. In this context, in 2013 Istat has implemented a new statistical register (called Frame-SBS) for the annual production of economic accounts statistics based on the integrated use of administrative and survey data. Given the peculiarities of the target population and the characteristics of the available sources, the development of the system has implied the management of a number of challenging issues, like the harmonization of concepts in the original sources to the statistical purposes (target populations/units, target variables), the evaluation of their quality and usability, the analysis and treatment of measurement, coverage and response errors.

The transition from the traditional survey-based estimation procedure to the new production strategy has required a deep change in terms of both the methodological and organizational strategies adopted. Furthermore, the use of integrated administrative data has determined the need of developing new methods and tools for the evaluation of the quality of all the components of the statistical process: input data sources, data processing, and outputs. Starting from the results of European projects like the BLUE-ETS and the AdminData, and taking into account some recent theoretical frameworks, in this paper we delineate the basic steps of the data production flow and propose a first scheme of indicators for measuring and documenting the quality of the Frame-SBS. The final goal is to implement a comprehensive control system to regularly monitoring the quality of the Frame-SBS, taking into account all the quality dimensions (accuracy, timeliness, coherence, etc.), allowing the identification of the possible weaknesses of the process, their impact on quality dimensions and supporting the evaluation of quality improvements.

Keywords: Administrative data, Quality.

1. Introduction

In the last years, Istat has strongly increased the number of administrative (hereafter *admin*) datasets that are centrally acquired for several statistical purposes. Such an increase calls for a tailoring of the current quality measurement and assessment approaches, by introducing a wider framework based on: the measurement of the quality of input sources when they are centrally acquired by Istat (Ambroselli and Di Bella, 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.

This paper deals with the quality assessment of the statistical register called Frame-SBS, which is currently used at Istat for the annual estimation of Structural Business Statistics (hereafter SBS). Actually, given the availability of stable, timely and reliable admin and fiscal sources providing high quality and detailed information on enterprises' profit and loss accounts, since 2013 Istat has been using a new estimation strategy based on the intensive and integrated use of such secondary data, complemented by direct survey data and driven by the Italian Business Register (Asia). The Frame-SBS we refer to contains microdata for the *main* economic variables for all 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.

In order to build the statistical register, the following admin sources are used (Luzi et al., 2014): Financial Statements (hereafter *FS*), Sector Studies survey (hereafter *SS*), Tax returns (hereafter *Unico*), Regional Tax on Productive Activities (hereafter *IRAP*). Moreover, information from Social Security Data (hereafter *SSD*) from the Italian National Security Institute are used as auxiliary data to support some of the data analysis and data processing activities at different stages of the Frame-SBS production process. Based on the new register, estimates for the *main* SBS can be nowadays computed at an extremely refined level of detail, overcoming some limitations of the previous estimation strategy. Improvements have been achieved in terms of both accuracy of cross-sectional estimates (as for example the sampling

error components have been removed), and consistency of estimates over time and among related statistical domains, with particular reference to National Accounts.

In this paper, we focus on the production process of the *main* Frame-SBS variables. The paper is structured as follows. Section 2 contains a description of the proposed quality framework, together with the list of the corresponding quality indicators. In section 3 some concluding remarks are provided and the directions for further developments are delineated.

2. Quality indicators: the proposed quality evaluation framework

2.1. Zhang's two-phase framework for quality assessment

In order to assess the quality of the Frame-SBS data and, indirectly, of the statistics based on its use, we adapted the framework proposed by Zhang (2012), which main aim is to provide a well-defined list of errors that can occur when the production of statistic is based on the combinations of various admin and statistical datasets. The framework consists of two phases and is represented in the lifecycle diagram shown in Figure 1 and Figure 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 give a quality measure of the source itself. The second phase focuses on errors arising when data from several sources are combined to produce a statistical output, in order to give a quality measure of the transformation process to adapt the data from their original purpose to the statistical one. Indeed, in this phase the reference point corresponds to the statistical population and to the statistical concepts to be measured. For more details on the Zhang's framework for quality assessment see Zhang (2012) and Zabala (2013).

2.2. The Frame-SBS production process

In this paragraph the production process of the statistical register Frame-SBS is described. We start from the Zhang's framework, which is useful in order to examine the design of any mixed-source statistical process in a clear way and, consequently, to understand which are the

sources of error potentially affecting the output data as a result of both the original design of each individual admin archive and the design choices of the statistical process.

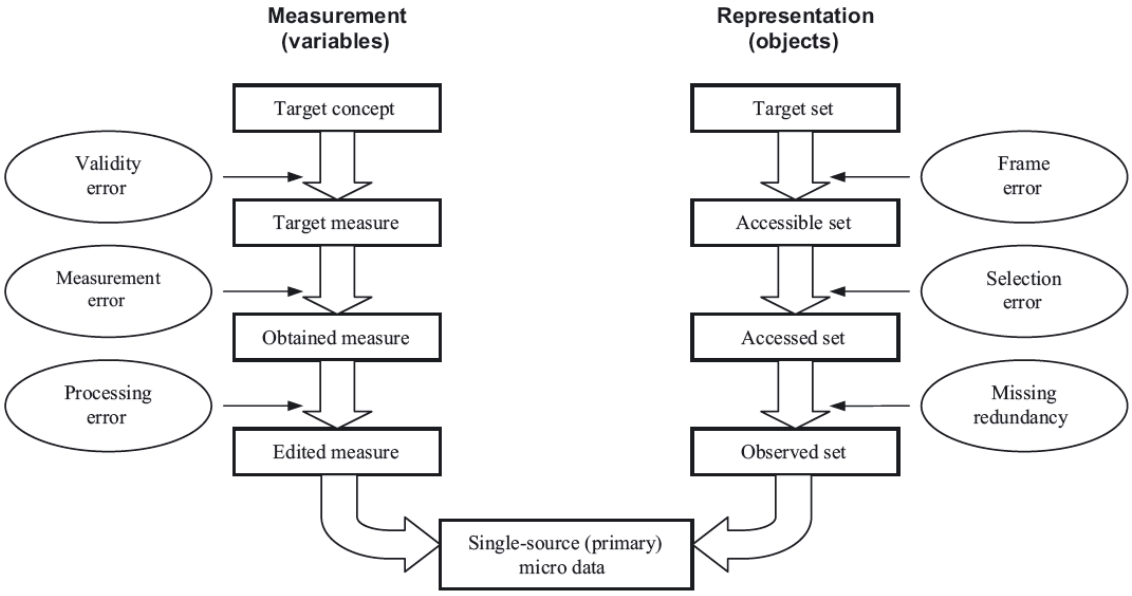

Figure 1: Sources of error in phase one of Zhang's framework.

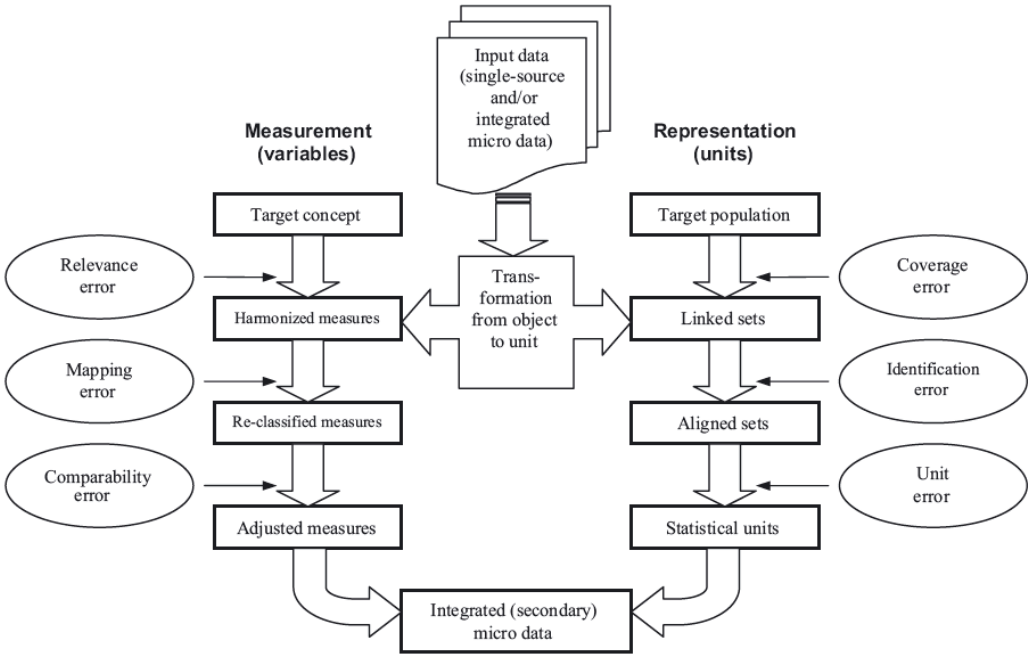

Figure 2: Sources of error in phase two of Zhang's framework.

Nevertheless, we propose to represent the process in three-phases: the first phase can be assimilated to the Zhang's phase one, while the second one has been divided into two sub-phases to better distinguish the specific steps of the transformation process the original data have to go through. Indeed, in phase two the admin data are evaluated according to the SBS target (both units and variables), but still there is a first sub-phase where each admin source is evaluated separately (phase 2a) in order to both determine the criteria according which to select and combine the data. Subsequently, the integrated dataset is created and is further elaborated (phase 2b) to attain the final SBS data.

Phase 1. Pre-treatment of the admin sources. The first phase of the Frame-SBS production process consists in the pre-treatment of each admin source's data. This phase is carried out separately for every source, covering each a different population and characterized by a peculiar structure and contents. Firstly, only the subset of items related to the Frame-SBS is selected, specifically the ones which are useful for deriving the SBS target variables. Hence, for each admin source, the aim is to verify whether there are changes over the time in the population coverage and in the supply timing and to identify and to eliminate the duplicate or unacceptable records (on the objects side). On the measurements side, an initial assessment of formal data inconsistencies is carried out, based on the use of accounting rules (edits).

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 harmonized w.r.t. the target SBS variables. The harmonization 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. As it is not always possible to directly "reconcile" the admin and the statistical definitions, the admin information is used to obtain the harmonized 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 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 compared with auxiliary information available from the SSD, 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 (Luzi et al., 2014). A macroediting 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.

2.3. A new framework for quality assessment of Frame-SBS

In the following schemes we propose a set of quality indicators for the assessment and the documentation of the quality of the Frame-SBS, consisting of both new measures and some adaptation of the indicators proposed by Zabala (2013). The indicators are presented divided by process phase, subject (variables, objects and units), process step and error type (see Figure 1 and Figure 2).

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 developed conceptual schemes provide 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.

3. Final remarks

In this paper a new quality framework for the quality assessment of the statistical register Frame-SBS on enterprises' accounts is presented. 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's production process, in order to identify the actual sources of errors and the corresponding quality measures on both the variables and the objects/units sides. It has to be remarked that the identification of the error sources represents the basis for the systematic and continuous improvement of the production process through the removal (or at least the reduction) of the error sources themselves in the subsequent replications of the Frame-SBS production process. Furthermore, the availability of such indicators for different reference years will allow to analyze the quality of Frame-SBS data and production process 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 is just an initial step, as additional developments in terms of quality indicators will follow, as a consequence of the possible extension of the admin sources used and the detection of further error sources. Furthermore, more specific indicators to assess the estimates' accuracy based on the use of admin microdata are also needed: to this aim, model-based approaches could be used, especially for variables characterized by lower coverage rates in the admin sources.

Phase 1 indicators	**Objects. Accessible Set -> Accessed Set; Selection error**	
	Proportion of missing units w.r.t. FS theoretical population	[1-*No. units in the source/ Total No. units in the theoretical population in BR*] x 100
	Proportion of units of BR population in the source, by source	[1-*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	*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 S in Source S with multiple identification 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 ref. 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 n.of units checked* x 100
	Proportion of units with all missing values, by source	*No. units with all values equal (missing or 0 or 1) / Total n.of units checked* x 100
	Proportion of units with all implausible 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	$Modification\ rate: \frac{Tot.no.values\ changed\ from\ a\ code\ to\ another\ code\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \text{x } 100$ $Net\ imput.rate: \frac{Tot.no.values\ changed\ from\ missing\ or\ zero\ to\ a\ code\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \text{x } 100$ $Cancellation\ rate: \frac{Tot.no.values\ changed\ from\ a\ code\ to\ zero\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \text{x } 100$

Phase 2a indicators	**Units. Target Population -> Linked Sets; Coverage error**	
	Proportion of SBS population units in source FS	*No. corporate enterprises of SBS population in source FS/ No.of corporate enterprises of SBS population* x 100
	Proportion of SBS population units in sources SS, Unico, Irap	*No. units of SBS population in source S / No.of units of SBS population* x 100
	Variables. Target Concept -> Harmonized 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. Harmonized 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

Phase 2b indicators	**Units. Target Population -> Linked Sets; Coverage error**	
	Proportion of units of SBS population in the integrated dataset (undercoverage). Also in longitudinal perspective.	*No. units of SBS population in the integrated dataset/ No. units in the SBS population* x 100
	Proportion of units of SBS population in the integrated dataset. Also in longitudinal perspective, by source	*No. units of SBS population in the integrated dataset from source S/ No. units in the SBS population* x 100
	Proportion of units of SBS population in the integrated dataset with information present in only one source, by source	*No. units of SBS population in only one source S/ No. units of SBS population in at least one source S* x 100
	Proportion of units of SBS population in the integrated dataset with information present in more than one source	*No. units of SBS pop. in more than one source S/No. units of SBS population in at least in one source S* x 100

Phase 2b indicators (*continues*)	**Variables. Re-classified Measures -> Adjusted Measure; Comparability error**	
	Proportion of units with influential values, by variable	***No.of units with influential error)/ Total no.of units*** **x 100**
	Proportion of outliers, by variable	***No.of units outliers/ Total no.of units*** **x 100**
	Proportion of units with imputed values	***No.of units with imputed values/ Total number 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's values imputed, by variable	***N.of units with imputed values for variable Y/Total number of unit*** **x 100**
	Composition of the proportion of variable's values imputed, by variable	$\textbf{Modification rate}: \frac{\textbf{N. of values of the variable Y changed from a code to a different code}}{\textbf{Total n. of imputed values of variable Y}}$ **x** $\textbf{Net imputation rate}: \frac{\textbf{N. of values of variable Y changed from missing or zero to a code}}{\textbf{Total n. of imputed values of variable Y}}$ **x 1** $\textbf{Cancellation rate}: \frac{\textbf{N. of values of the variable Y changed from a code to zero}}{\textbf{Total n. of imputed values of variable Y}}$ **x 100**
	Impact of data editing and imputation on microdata, by variable	***Simple and quadratic distance between the pre-edited (Y) and post-edited (Y*) microdata of variable Y*** $DL_1(Y_i,Y_i^*) = \Sigma_i N \lvert Y_i - Y_i^* \rvert$ ***/Total N. of units N*** $DL_2(Y_i,Y_i^*) = \sqrt{\Sigma_i N (Y_i - Y_i^*)^2}$ ***/ Total N. 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 (histograms, univariate statistics, etc.) pre- and post- editing and imputation*
	Impact of data editing and imputation on statistical relations	*Pearson correlation index, Covariance matrix*
	Impact of data editing and imputation on statistical aggregates, by variable	***Tot. of the variable before editing and imputation /Overall total of the variable after editing and imputation*** **x 100**

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 measuring the sampling and non-sampling components of the difference between the Frame-SBS estimates and the corresponding ones resulting from the direct surveys on enterprises' accounts. In addition, iterative procedures (based e.g. on bootstrapping or on multiple imputation strategies) could be tested to measure the additional uncertainty due to the imputation process, under appropriate assumptions on the missing data mechanisms.

4. References

Ambroselli, S. and Di Bella, G. (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), Vienna. 3-5 June.

Brancato, G., Boggia, A., Barbalace and F., Busetti C. (2014), Quality Guidelines for statistical processes using administrative data, European Conference on Quality in Official Statistics (Q2014), Vienna. 3-5 June.

Luzi, O., Guarnera, U. and Righi, P. (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. 3-5 June.

Zabala, F., Reid, G., Gudgeon, J. and Feyen, M. (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, 66, n.1, pp. 41-63.