COMITATO CONSULTIVO PER LE METODOLOGIE STATISTICHE

ADVISORY COMMITTEE ON STATISTICAL METHODS

A new framework for quality assessment of processes based on survey, administrative and register combined data

Fabiana Rocci (rocci@istat.it)[1], Roberta Varriale (varriale@istat.it)[1], Orietta Luzi (luzi@istat.it)[1]

[1]Istat

Key words: Quality evaluation, statistical registers, Total Survey Error, multi-source data, combined data.

Introduction

The production of Official Statistics based on a combination of data from different sources - that most commonly comprehend administrative as far as Big Data - has spread out in recent years. As a consequence, across all the National Statistical Institutes (NSIs) many and intense developments have been worked out to achieve new strategies in producing the required outputs.

In this perspective, also the Italian National Statistical Institute (Istat) has launched the modernization of its overall production system, to deliver a whole new coherent production process based on the extensive use of data from external sources, for which the administrative data play a central role.

The new statistical production strategy is based on a coordinated system of statistical registers (named SIR), aiming each at representing specific phenomena concerning a given population. To this aim, for each statistical register all the administrative sources potentially containing information related to the target phenomena are considered. The final required statistical outputs are thought to be reachable through different designs: either directly from the registers or by integrating information from the registers with direct survey data.

In this view, the traditional processes, based on single data sources, obtained by direct surveys, are planned to be rethought. The challenge is to move towards processes where the combination of the available administrative data (AD in the following) should represent as far as possible the primary source, delivering strong and extensive information about the phenomena under study. This new approach implies the need to define new methodologies for efficient data treatment, integration and estimation in order to achieve the required estimates, determining statistical production processes different from the traditional ones as represented by the Generic Statistical Business Process Model (GSBPM, Unece, 2013).

To this extent, beyond the design and the operational implementation of such new processes, some theoretical analyses are outgoing in order to define proper guidelines for such new statistical processes. Among others, a theoretical and methodological key issue to

be considered relates to the development of a new quality framework to assess the quality of Official Statistics based on a multi-source process.

This paper focuses on this issue, with the final aim to propose an evaluation system framework for assessing and monitoring, through a system of indicators, the quality of new processes and the resulting outputs. The starting point is the adaptation of the two-phase life-cycle paradigm proposed by Zhang (2012) applied by Zabala *et al.* (2013), and the subsequent Total Survey Error (TSE) proposed by Reid et al. (2017) in the context of the use of AD supplemented by survey data, designed to help practical decisions about statistical design and monitoring of new processes. As far as different AD are available, different scenarios are possible, in terms of coverage, validity and feasible methods to treat data to reach a final combined dataset. From this point of view, the reflections about how to define the multi-source processes and to delineate their phases proposed by Sander (2017) have been taken as reference, with the final aim to delineate a comprehensive framework of evaluation of every stage of the production process .

As case study, we present the so called statistical register *Frame-SBS*, (Luzi *et al.*, 2016; Luzi *et al.*, 2014), which has been built to support the annual estimation of the Structural Business Statistics (hereafter *SBS*) on enterprises' profit and loss accounts. The application of the quality framework proposed by Zhang to this register has highlighted a number of new issues, mainly related to the need of adopting different methodological solutions to produce the register: specific decisions have driven to establish different methods for different sets of variables to achieve different (kinds of) outputs.

As a major result, a reflection about the need to formulate an additional phase in the Zhang's and Reid's TSE is gathered. In some situations, it seems to be necessary to introduce the definition of a specific phase, which we will refer to as *statistical-driven decision* phase, during which much decisions, strongly driven by the statistical purposes and by operational constraints, are taken about how to combine the external/administrative sources.

The whole system of indicators, defined along different phases representing the production process, should help as guidelines to identify potential source of errors, to measure their effect on the output and to prevent them, in order to progressively improve the new production system.

The paper is structured as follows. Section 2 describes the main reference literature on TSE and its adaptation to statistical processes based on using administrative sources (TSE*adm* hereafter). In section 3 the Frame-SBS register is presented as case study and the application is described of the TSE*adm*, focusing on the criticisms we encountered. Section 4 describes the proposed quality evaluation framework together with a first proposal of quality indicators referring to the Frame-SBS register. Section 5 concludes the work.

2. Existing literature

In many applied cases, the starting point to define a process based on the use of external data has been to reproduce what has been already defined for surveys. In such a way, some frameworks have already been proposed, but each of them leaves some open issues and calls for the need of re-formulate the undergoing *statistical thinking* (Reid *et al.*, 2017). Indeed, researchers agree that such a change calls for a tailoring of the current approaches for the new statistical processes in terms of: (i) design, (ii) implementation and (iii) quality measurement and assessment.

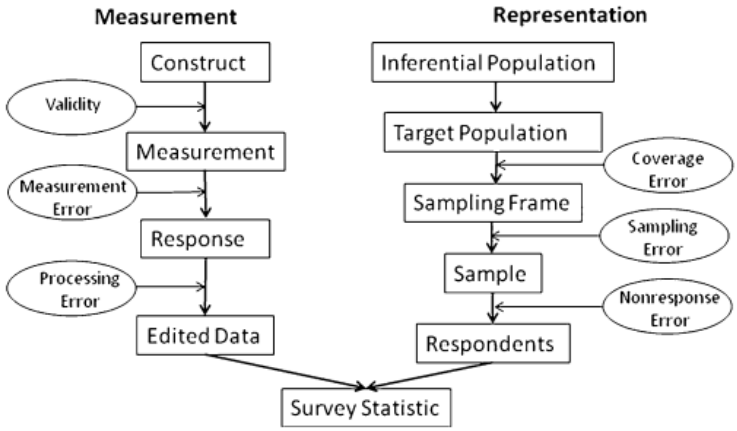
In the following, we describe the steps of the theoretical developments on quality measurement and assessment, in order to illustrate which are the issues that are still open.

2.1 Total survey error (TSE)

Every statistical process can gather outputs that actually differ from the true target parameter. Several frameworks have been developed to give an overall quality indicator of the given result, according to the process design and the several phases it is due to go through. Many components can be identified among the reasons why the distance between the true and the "measured" parameters can exists, generally divided between sampling and non-sampling components.

A starting point in the existing literature on quality assessment framework is the life-cycle approach described by Groves *et al.* (2004), that provided a systematic outlook on the potential error sources starting from the conception, collection and processing till the final production of estimates (Figure 1).

Figure 1 - Life-cycle of surveys, Groves (2004)



Afterwards, the most common scheme that refers to it is the TSE concept, for which Groves and Lyberg (2010) provide an interesting overview by illustrating the theoretical process leading to its development from the beginning of 40s.

TSE is a concept that tries to describe statistical properties of survey estimates, incorporating a variety of error sources. It is thought to help survey designers as a planning criterion among a set of alternative designs: the one that gives the smallest TSE (for a given fixed cost) should be chosen.

Nevertheless, the term TSE is not defined in a unique way and different researchers include different components of error within it, for which a number of typologies exist in the literature (Biemer *et al.*, 2017). Beyond considering the different problems arising in surveys, a comprehensive paradigm should also balance them with survey costs and other constraints (Weisberg, 2005). However, during the years, literature has developed along the

path of increasing the categorization of errors on dimensions of variance and bias on one hand, and errors of observation and non-observation on the other.

The main concept commonly recognized is to understand how to identify, in every phase of the survey, the nature of the error causing the gap between the theoretical concept, a priori defined, and the actual measure obtained through the survey. As described by Groves and Lyberg (2010) "researchers such as Kendall, Palmer, Deming, Stephan, Hansen, Hurwitz, Pritzker, Tepping, and Mahalanobis viewed a small error as an indication of survey usefulness, and Kish (1965) were the first to equate a small error with the broader concept of survey data quality. Later, other researchers and statistical organizations developed frameworks that include not only accuracy but also non-statistical indicators of data quality such as relevance, timeliness, accessibility, coherence, and comparability (e.g., Eurostat 2000; OECD 2003)".

To summarize, the roots of TSE are found in lists of problems facing surveys beyond those of sampling error and, during time, the term has evolved to a nested taxonomy of concepts of errors. In addition to the error typologies serving as checklists of possible errors and their estimation, it is important to find ways to eliminate their root causes.

2.2 TSE for administrative data based estimates (TSE*adm*)

Many reflections have started about how to enhance a framework for evaluating the quality of both the statistical production processes and the final results obtained by combining different sources, taking into consideration that AD nowadays play a different role from the traditional one (source of auxiliary information) in the production processes. As for TSE, also for such new processes the reasoning of identifying the causes of the distance between the theoretical value and the actual measurement has been identified. Starting again from the life-cycle represented in Figure 1, Bakker (2012) proposed its adaptation to register data, developing a theory for data treatment and integration based on "the general idea that it is likely that the errors that normally emerge in surveys will also occur in registers". Zhang (2012) has proposed a two-phase life-cycle model of integrated statistical microdata, reported in Figure 2 and Figure 3. The innovating idea is to apply a similar reasoning of identifying errors (i) first with respect to the original target of secondary data, and (ii) subsequently to assess how the integrated dataset, resulted from the use of external source, can satisfy the statistical purpose. In Zhang's framework, for the two phases different kinds of errors are theorized and the distinction between *object* and *units* is introduced, to specify clearly the purpose of each phase. Hence, a well-defined list of errors that can occur when the production of statistics is based on the combination of various administrative and statistical data is provided.

More in details, the first phase of the life-cycle model deals with each single source and categorizes errors arising with respect to the original administrative source's target population (*objects)*, in order to give a quality measure of every source itself. The second phase focuses on the assessment of potential errors in the dataset resulting from the integration of the original sources, aimed at producing the required statistical output. The aspect under analysis, in this case, is to assess the quality of the integrated dataset, that should give a measure of the cost to adapt the data from their original purpose (administrative) to the statistical one. Indeed, in this phase the reference point corresponds to the statistical target population (*units*) and to the statistical concepts to be measured.

Figure 2 - Sources of error in phase one of Zhang's framework (Zhang, 2012).

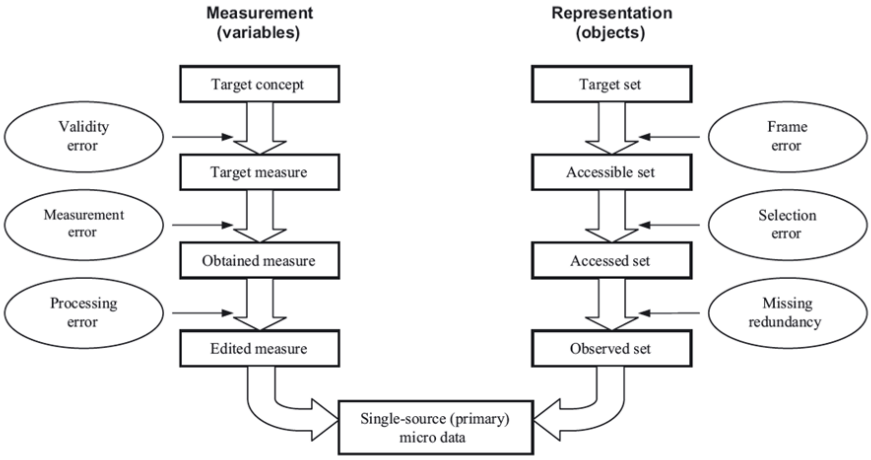
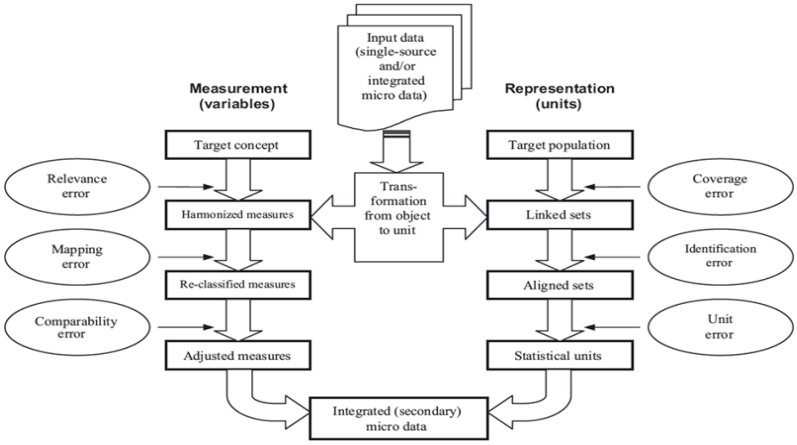


Figure 3 - Sources of error in phase two of Zhang's framework (Zhang, 2012).



Finally, Reid *et al.* (2017) proposes a three-phase framework applying the TSE paradigm to the new realm of statistical production. The Zhang's framework is interpreted as an extension of the TSE approach, since errors are linked to a life-cycle model, and a third phase consisting on the elaboration of the statistical output is added. This approach tries to comprehend an overall production process, where the integrated dataset in Zhang's framework can play different roles according to the type of statistical process. Indeed, it can be considered as the final output, complete in every observation and properly designed to achieve the statistical purpose, as commonly the statistical register are. Otherwise, it can be a transitional output: starting from the final estimates the final output can be achieved

through other estimation methodologies, that would add other variability components to the final estimates.

Furthermore, Reid *et al.* (2017) address two interesting issues: (i) the need of a more suited "statistical thinking" of the entire quality framework for the processes based on the combined use of AD, and (ii) the importance of a quality framework as a way to determine the strengths and limitations that different strategies (use of AD only, use of AD combined with direct survey data, use of survey data only) may have on the quality of a statistical output.

In the following, we will refer to Zhang's and Reid *et al.* frameworks as TSE*adm.*

3. Frame SBS case study, a critical application of TSE*adm*

In this Section the characteristics and the process of Italian register Frame-SBS is described and a critical application of TSE*adm* is carried out. The aim is to highlight the critical aspects emerging when the framework is applied to an actual process based on the (combined) use of administrative data.

3.1 Frame SBS: a short description

The statistical register Frame-SBS, designed to satisfy the European SBS regulation, is built for the annual release of statistics on loss and accounts of enterprises. It is designed with respect to the given international agreement on enterprises accountability and covers industry, construction, distributive trades and services, broken down to a very detailed sectoral level. Frame-SBS describes the structure and performance of businesses across the European Union.

In Italy, SBS variables are covered by a number of administrative sources, which can provide information on the enterprises' accountability variables at micro level**.** Such administrative sources are the Financial Statements (hereafter *FS*), the Sector Studies survey (hereafter *SS*), the Tax returns (hereafter *Unico*), the Regional Tax on Productive Activities (hereafter *Irap*).

Traditionally, SBS was estimated based on two direct annual surveys. The first one was the sample survey on Small and Medium Enterprises (SME), for which about 100,000 enterprises with less than 99 persons employed were sampled, representing a population of about 4.3 million of units. The second one was the total survey on Large Enterprises (LE), for a census of about 11,000 enterprises with 100 or more persons employed.

Istat has revised the estimation strategy for SBS in order to reduce the statistical burden on enterprises by extensively exploiting the available administrative sources, considered to be very reliable given the legal framework under the construction of the balance sheet, followed as well by the SBS regulation. Indeed, the AD sources often share similar aim, that all relates to fiscal control on specific types of enterprises and related issues.

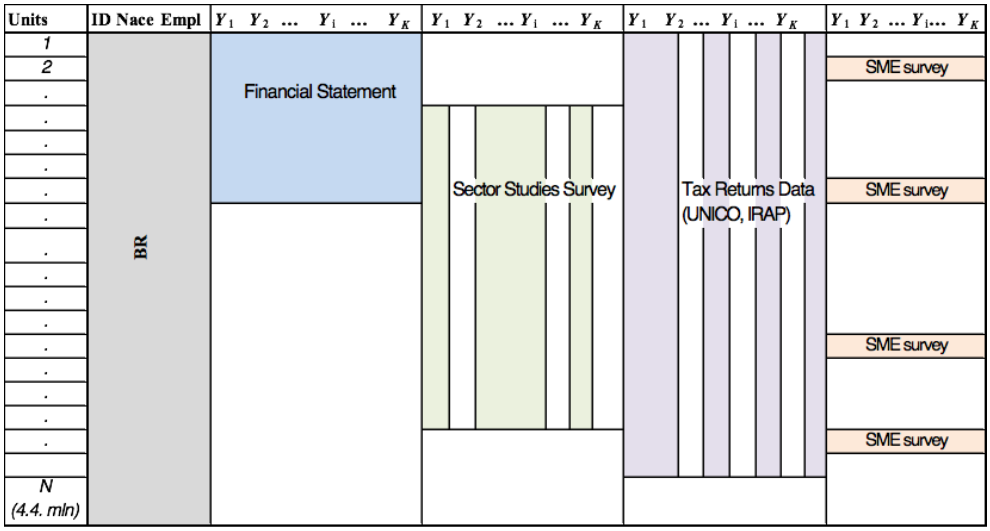
In the following the process is described, starting from the design issues that have been faced during the initial phase of the production process.

Step 1. At first, a quality assessment process on each candidate data source has been performed (Curatolo *et al.* 2016), in terms of quality with respect to the AD purpose, in order to evaluate to which extent they could assure coverage, both from the units and

variables side, and in terms of harmonization to the statistical definition. The quality of the each AD source was good, so a pre-treatment is need only to a certain extent; to eliminate possible "unacceptable" information (e.g. formal inconsistencies, duplicated objects, etc.).

Step 2. Afterwards, for each source, the quality assessment process has been based on a set of quality criteria such as *relevance and coverage* (in terms of target population), *completeness and validity* (in terms of target statistics), *accuracy*, *timeliness*. As a result, a final mapping of the overall coverage has been pictured for the whole system (Figure 4). The presence of the *K* variables required by the SBS was assessed on every source and quality indicators have been computed for each variable and available sources, with respect to the covered SBS target population, that has allowed to formulate a first idea to which degree the data source is capable of undergoing integration to achieve a complete dataset on the entire population.

Figure 4 Mapping of the coverage of AD for the SBS variables and population



At a first glance, the picture results as a chessboard, since some source are overlapping but no one of them could cover the same set of variables neither the whole population.

The main issues that were observed during the preliminary analyses are:

- different population coverage were guaranteed by different sources;
- different degrees of validity for each variable (validity error refers to the difference between the target administrative concept and the statistical one), that leads to different coverage for every variable, according to the source they are from;
- difference of measurement on some of the variables present in different sources is registered.

Step 3. Taking into account the issues addressed in Step 2 and in order to guarantee both the quantity of information gathered from administrative data and the internal coherence of the main variables, two different alternative strategies could be applied:

- **Strategy A**: for each statistical unit, integration of all available information coming from the administrative sources is performed and subsequently data are treated - edited or imputed - to achieve a "balance" check to ensure internal consistency.
  Strategy A maximizes the overall quantity of information.
- **Strategy B**: a "priority" is assigned to every source (FS, SS and Unico-Irap). So that for each statistical unit only one source is chosen, and the population coverage has different degrees. In this case, treatment is still necessary, but to impute data only if missing.
  Strategy B maximizes the internal coherence of the dataset.

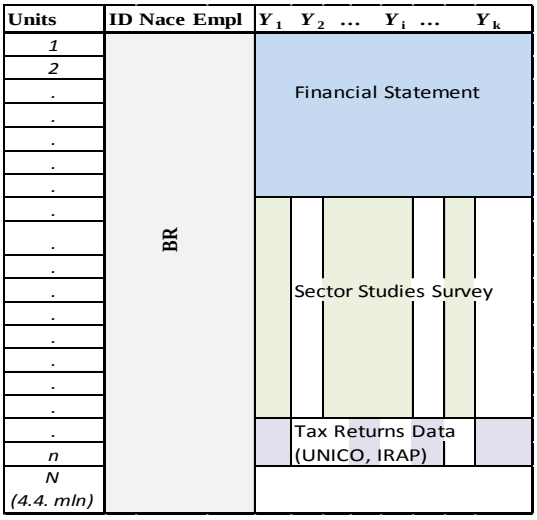
In Frame-SBS production process, Strategy B has been chosen, resulting in different coverage rates w.r.t. the various sub-populations of enterprises, as shown in Figure 5.
After the first integration, for each variable the coverage has been measured with respect the whole target population and different groups of variable have been classified:

Set of *BR* variables. A set of variables coinciding to those of the Businesses Register: economic activity (Nace) and Employment (Emp) of each enterprise

Set of *core* variables. The set of *core* variables Y_h ($h=1,...,H$; $H < K$) that are the variables "highly" covered by the administrative data, so that the integration of different administrative data cover up to 95% of the target population for each variable. None of those variables is completely gathered by any external source, so that some partial and total unit non response is observed (see Step 4 and 5).

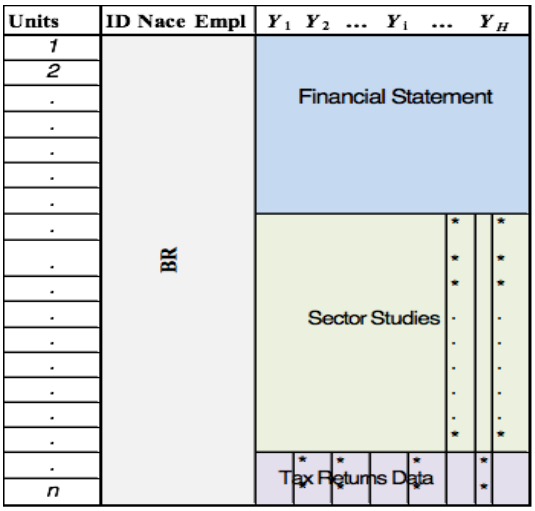
Set of *components* variables. The set of variables Y_j ($j= H+1,..., K$) components of the *core* variables, which are not properly represented by AD (see Step 6).

Figure 5 *BR* and *core* variables – Initial integrated dataset for the covered units



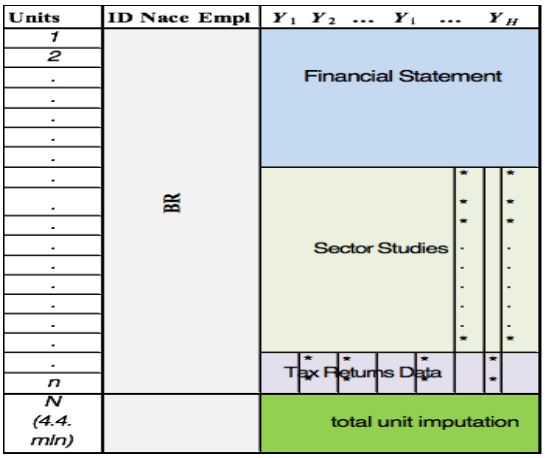
Step 4. Prediction/imputation of the missing values of the *core* variables treated as partial non response for the $n<N$ unit covered by AD (Di Zio *et al*., 2016). In Figure 6, the integrated dataset for each unit i ($i=1,..,n$) of each *core* variable h ($h=1, ...,H$) is represented.

Figure 6. BR and *core* variables - Final integrated and imputed dataset for the covered units



Step 5. Prediction/imputation of the *core* variables for totally uncovered units (Di Zio *et al*., 2016). The output of this step is a "census" database (Figure 7) containing information on the *core* variables at micro-data level for *all* the units in the SBS population, as identified by the Italian BR – Asia.

Figure 7. *BR* and *core* variables - Final dataset of the whole SBS target population



Step 6. Estimation of the *components* variables jointly using sample survey data (SME survey) and the information on the *core* variables reached at the previous step (as auxiliary information). For every variable Y_j (j = H+1, …, K), domain estimates at the required levels are obtained based on the use of a *projection estimator* (Righi, 2016).

Step 7. SBS are properly computed from the register and released.

4. The proposed quality evaluation framework

Starting from the analysis on how to apply the TSE*adm* proposed by Zhang (2012) and Reid *et al.* (2017) to the Frame-SBS production process, a number of issues to be taken into account emerged.

Above all, two main different statistical processes can be distinguished, one for *core* variables and one for *components* variables. Furthermore, the two processes are sequential, so that actions performed in steps 1-4 will effect actions in step 5. An important issue arose about step 3, concerning how to integrate administrative sources. As briefly described, alternative strategies could be theoretically adopted. The final integration method contains peculiar choices, based on the evaluation of the content and the internal consistency of the integrated sources.

Finally, it is important to observe how it is completely different to evaluate Frame-SBS in terms of: (i) register as released at step 3, coming only by the combination of administrative data; (ii) the statistical register released at step 4, obtained also through a micro imputation process of all the *core* variables; (ii) register including *core* and *components* variables, and (iii) SBS estimates (step 6), using different methodologies for each group of variables (and, in some cases, for each variable).

Starting from these considerations, some issues that needed to be further addressed were identified:

- there is a lack of a well-defined vocabulary to better distinguish which kind of data, processes and outputs are involved in each phase. This is necessary in order to give a clear definition of the general framework of analysis;
- there is a need to define and to distinguish different kinds of statistical outputs that can be obtained based on the use of AD and to develop methods to ensure coherence among estimates. This is necessary in order to identify the most appropriate quality indicators in the different contexts;
- the second phase of TSE*adm* should be further enhanced to trace the actual assessment/integration/treatment process and better assess quality. In fact, the dataset resulting from this phase can be obtained by using different integration strategies and treatments: as a consequence in phase two it should be allowed to evaluate the effects of different alternative choices.

Hence, starting from TSE*adm*, we propose to split the second phase into two sub-phases to better identify the specific steps of the "transformation" process the original data have to go through and create a system of indicators to evaluate each of them.

The general quality framework we propose can be represented in the following way:

Phase 1. Pre-treatment of the administrative sources.

The first phase of a production process based on AD consists in the pre-treatment of each external source's data. This phase is carried out separately for every source, covering different population and characterized by a peculiar structure and contents. This phase coincides with Zhang's phase 1. As a consequence, the potential error types are the ones reported in Figure 2.

Phase 2a. Assessment of the administrative sources, taking into account the administrative purposes.

Each administrative source is evaluated separately, in order to assess its quality with respect to the specific statistical targets (statistical units/variables). This phase provides useful elements to define the data selection and the integration strategy, e.g. when multiple sources are available for same target variables and/or sub-populations.

Phase 2b. Integration of the sources.

In this phase, the integrated dataset is generated, and a further quality assessment is performed. This phase partly corresponds to the Zhang's phase 2. Additional actions should be taken into account in order to allow the evaluation of the complete production process.

Actually, the integrated dataset is usually treated in order to resolve possible statistical inconsistencies (e.g. outliers), or to impute partially or totally missing information (deriving from the sources incompleteness w.r.t. target variables and under-coverage w.r.t. target population, respectively), etc.

For each phase and each potential error, specific indicators can be proposed for quality assessment. It is worthwhile to note that some type of errors (and the corresponding quality indicators) may appear in more than one phase (e.g coverage error).

4.1 The proposed quality evaluation framework: Frame-SBS as a case study

In this paragraph we present an application of the proposed *TSEadm* to the *Frame-SBS* production process.

Referring to error classification, we adopt the error typologies and the definitions proposed by Zhang (2012) and Zabala (2013). Nevertheless, the error types identified by Zhang as arising in Phase 2 (Figure 3) are assigned to phases 2a and 2b of the new TSE*adm*, according to the treatments that are performed during each of them.

A first set of suitable indicators are proposed by phase, subject (variables, objects and units), process step and error type. Both quantitative and qualitative measures are considered.

In Table 1, quality indicators for the assessment of the first phase of the Frame-SBS production process are suggested. Tables 2 and 3 contain a draft proposal of quality indicators for phases 2a and 2b according to the new TSE*adm*.

Table 1: Phase 1 quality indicators by subject, phase and error type

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	$Modification\ rate: \frac{Tot.no.values\ changed\ from\ a\ code\ to\ another\ code\ in\ sour}{Total\ no.\ imputed\ values\ in\ source\ S}$ $Net\ imput.rate: \frac{Tot.no.values\ changed\ from\ missing\ or\ zero\ to\ a\ code\ in\ sour}{Total\ no.\ imputed\ values\ in\ source\ S}$ $Cancellation\ rate: \frac{Tot.no.values\ changed\ from\ a\ code\ to\ zero\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \times 100$

Table 2: Phase 2a quality indicators by subject, phase and error type

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 harmonized 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: Phase 2b quality indicators by subject, phase and error type

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
	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	$\textbf{Modification rate}: \frac{N.\ of\ values\ of\ the\ variable\ Y\ changed\ from\ a\ code\ to\ a\ different\ code}{Total\ n.\ of\ imputed\ values\ of\ variable\ Y}$ x 100 $\textbf{Net imputation rate}: \frac{N.\ of\ values\ of\ variable\ Y\ changed\ from\ missing\ or\ zero\ to\ a\ code}{Total\ n.\ of\ imputed\ values\ of\ variable\ Y}$ x 100 $\textbf{Cancellation rate}: \frac{N.\ of\ values\ of\ the\ variable\ Y\ changed\ from\ a\ code\ to\ zero}{Total\ n.\ of\ imputed\ values\ of\ variable\ Y}$ 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

5. Conclusions and future work

In this paper a comprehensive framework for the quality assessment for statistical processes using administrative data is proposed, starting from the scheme proposed by Zhang (2012), applied by Zabala *et al.* (2013) and further developed by Reid *et al.* (2017). Actually, the identification of error sources in the production process of a register represents the basis for the systematic and continuous improvement of the quality of both the register and the derived outputs, through the prevention/elimination (or at least reduction) of such errors in the subsequent replications of the production process itself. The availability of quality indicators for different reference years will also allow the analysis of both data and process quality in a longitudinal perspective. In addition, based on the quality framework, a complete quality report could be developed for documentation and dissemination purposes.

An in depth analysis of the proposed framework in terms of life-cycle of a multi-source process and the corresponding phases, where specific errors can occur, has showed at this stage some lacks. A critical application of the TSE*adm* to a case study, the Frame-SBS production process, has highlighted how different decisions can be taken in integrating and combining different data sources.

We propose to introduce a distinction of the second phase of Zhang's framework into two sub-phases, to better identify the different patterns along which the process can go through, taken into account all the features the external data can present time by time.

This proposal has to be considered as an initial step of a complex project. The definition of a complete framework with a *final* phase, the classification of possible outputs of multi-source statistical processes, and the development of proper quality measures for the final outputs will be the future goals.

Aknowlegments
We sincerely thank Dr. Giovanna Brancato for having supported the study and for her precious suggestions, which have worked out as a valid and solid guideline for the structure and the contents of the publication.

Methodological issues submitted to the attention of the Committee

- Has the critical analysis of the TSE*adm* to the Frame-SBS driven to significant reflections?
- Is it worth it to underline the importance to start again in determining which are the possible ways of obtaining estimates based on the a multi-source process?
- Phase 2b comprehends the transformation process that data have to go through from the integrated dataset to the final outputs. Does it worth to separate some phases (i.e. imputation, estimation)? Are the indicators proposed apt to distinguish between errors due to different parts of the transformation process?

- Concerning the set of indicators presented in this paper, do you have recommendations/suggestions?
- It would be interesting to propose some synthetic indicators for the single phases and sub-phases of the framework. Do you have any suggestion on that?
- Which kind of longitudinal indicators can be appropriated?
- A categorization of all possible outputs of multi-source statistical processes (third phase) is under study. Consequently, specific indicators are proposed as future study. In particular, how is it possible to evaluate the accuracy of the final estimates in this context?

References

Bakker, B. 2012. Estimating the Validity of Administrative Variables. *Statistica Neerlandica* 66: 8–17. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00504.x.

Biemer P.P. de Leeuw E., Eckman S., Edwards B. Kreuter F., Lyberg L.E., Tucker N.C. West B.T. (editors) (2017). Total Survey Error in Practice. John Wiley & Sons, Hoboken, New Jersey.

Brancato G., Boggia A., et al. 2016. Guidelines for the quality of statistical processes that use administrative data Version 1.1 http://www.istat.it/en/files/2013/04/Linee-Guida-v1.1-Versione-inglese.pdf

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

Di Zio M., Guarnera U. 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.

Eurostat. 2000. *Assessment of the Quality in Statistics*. Eurostat General/Standard Report, Luxembourg, April 4–5

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Rourrangeau. 2004. Survey Methodology, Wiley, New York

Groves R. M., Lyberg L.E. 2010. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, Volume 74, Issue 5, 1 January 2010, Pages 849–879, Doi: https://doi.org/10.1093/poq/nfq065

Kish L. 1962. Studies of Interviewer Variance for Attitudinal Variables. *Journal of the American Statistical Association.* pg.92-115.

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 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)*. Wien, 2-5 June.

OECD. 2003. *Quality Framework and Guidelines for Statistical Activities*. Version 2003/1. Doi: http://www.oecd.org/dataoecd/26/42/21688835.pdf

Reid G., Zabala F., Holmberg A. 2017. Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics*, Vol. 33, No. 2, 2017, pp. 477–511. Doi: http://dx.doi.org/10.1515/JOS-2017-0023

Righi P. 2016. Estimation procedure and inference for component totals of the economic aggregates in the "Frame SBS". *Rivista di Statistica Ufficiale*. N.1/2016.

UNECE (2013) Generic Statistical Business Process Model (GSBPM) (Ver 5.0., December 2013) http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

Weisberg H.F. (2005). The total survey error approach. The University of Chicago.

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: 41-63.