Working Paper.

**UNITED NATIONS**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Workshop on Statistical Data Editing**
(Neuchâtel, Switzerland, 18-20 September 2018)

# A proposal of an evaluation framework for processes based on the use of administrative data

Prepared by [Fabiana Rocci, Roberta Varriale, Orietta Luzi, Italian National Institute of statistics, Italy]

# I. Introduction

1. The production of Official Statistics based on a combination of data from different sources has spread out in recent years and all the National Statistical Institutes (NSIs) are developing new strategies in producing the required outputs. The challenge is to move towards processes where the combination of the available administrative data (AD hereafter) should represent as far as possible the primary source, delivering strong and extensive information about the phenomena under study. In this view, the traditional processes are planned to be rethought and many theoretical analyses are outgoing in order to define proper guidelines. 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, useful to: (i) monitor the development and the final quality of new processes, (ii) help practical decisions about statistical design and monitoring.

2. The starting point to be taken as reference in literature are the adaptation of the two-phase life-cycle paradigm proposed by Zhang (2012) and applied by Zabala *et al.* (2013), and the subsequent Total Survey Error proposed by Reid *et al.* (2017) in the context of the use of AD (hereafter TSE*adm*) supplemented by survey data. As in a life-cycle approach (Groves et al., 2004), the TSE*adm* aims at providing a systematic outlook on the potential error sources starting from the conception, collection and processing till the final production of estimates. According to this classification, the *editing and imputation* (hereafter E&I) phase can cause mainly processing errors - represented by the difference between the observed and the edited responses - in the measurement process.

3. As case study, the Istat statistical register Frame-SBS (Luzi *et al.*, 2016a; Luzi *et al.*, 2016b; Luzi *et al.*, 2014), supporting the annual estimation of Structural Business Statistics (hereafter SBS) on enterprises' profit and loss accounts, has been chosen. The critical application of TSE*adm* to the register Frame-SBS has highlighted some interesting issues, mainly related to the case where different methodological solutions to produce the register could be adopted. As a major result, we propose to enhance the TSE*adm* with another phase, in order to take into consideration the need to measure the effect of every potential step of a new process. In this paper, we describe the entire Frame-SBS process according to the proposed quality framework, focusing on the two phases where E&I is performed.

4. The paper is structured as follows. Section 2 describes the Frame SBS characteristics and production process. In section 3 the proposed quality evaluation framework is presented and applied to the Frame SBS production process. Section 5 concludes the work.

# II. Frame-SBS case study

5. In this Section the characteristics and the production process of Italian register Frame-SBS are described. As introduced, our aim is to exploit the production process of Frame SBS in order to make a link between it and the quality evaluation framework we propose (see section 3).

## A. Frame SBS: a short description

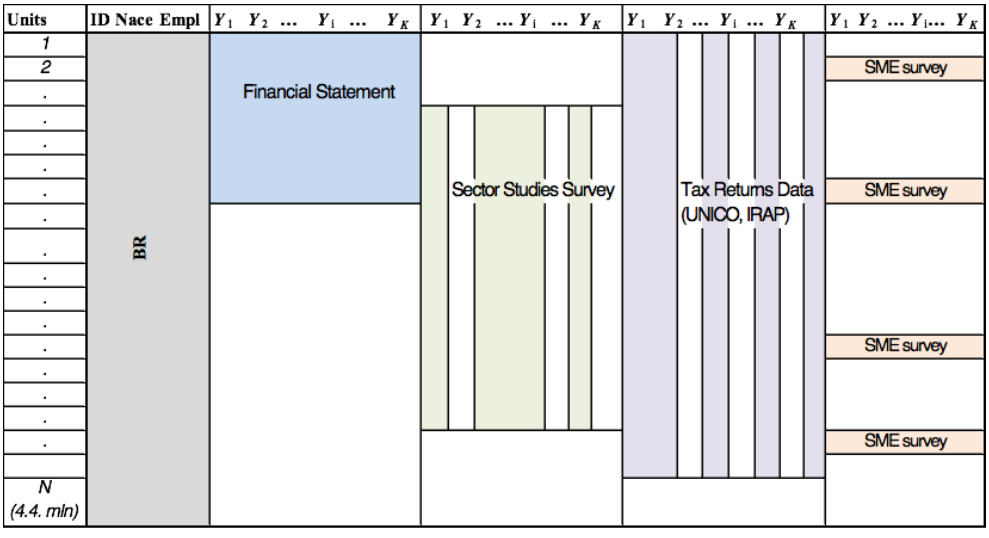
6. The statistical register Frame-SBS, designed to satisfy the European SBS regulation, is built for the annual release of statistics on loss and accounts of enterprises. It is designed with respect to the given international agreement on enterprises accountability and covers industry, construction, distributive trades and services, broken down to a very detailed sectoral level. In Italy, SBS variables are covered by a number of administrative sources, which can provide information on the enterprises' accountability variables at micro level: 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.
In the following the production process of the Frame-SBS is described, starting from the design issues that have been faced during the initial phase to the output release.

## B. Frame SBS: production process

7. In this paragraph, the Frame-SBS process is described along the steps that were tackled during the implementation design.

**Step 1.** At first, a quality assessment process on each candidate data source has been performed (Curatolo et al. 2016), in terms of quality w.r.t. the specific AD purposes, to evaluate to which extent they could ensure coverage, both from the units and variables side, and in terms of harmonization to the statistical definitions. A pre-treatment is needed to eliminate possible "unacceptable" information (e.g. formal inconsistencies, duplicated objects, etc.).
**Step 2.** For each source, the quality assessment process has been based on a set of quality criteria such as relevance and coverage (in terms of SBS target population), completeness and validity (in terms of target SBS), accuracy, timeliness. As a result, a final mapping of the overall coverage has been pictured for the whole system (Figure 1). The presence of the K SBS variables was assessed on every source and quality indicators have been computed for each variable and available sources.

**Figure 1. Mapping of the coverage of AD for the SBS variables and population**



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 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:

(a) Strategy A: for each statistical unit, all available information coming from the AD sources is integrated and subsequently data are treated to ensure internal consistency. Strategy A maximizes the overall quantity of information.
(b) Strategy B: a "priority" is assigned to every source (FS, SS and Unico-Irap). 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.

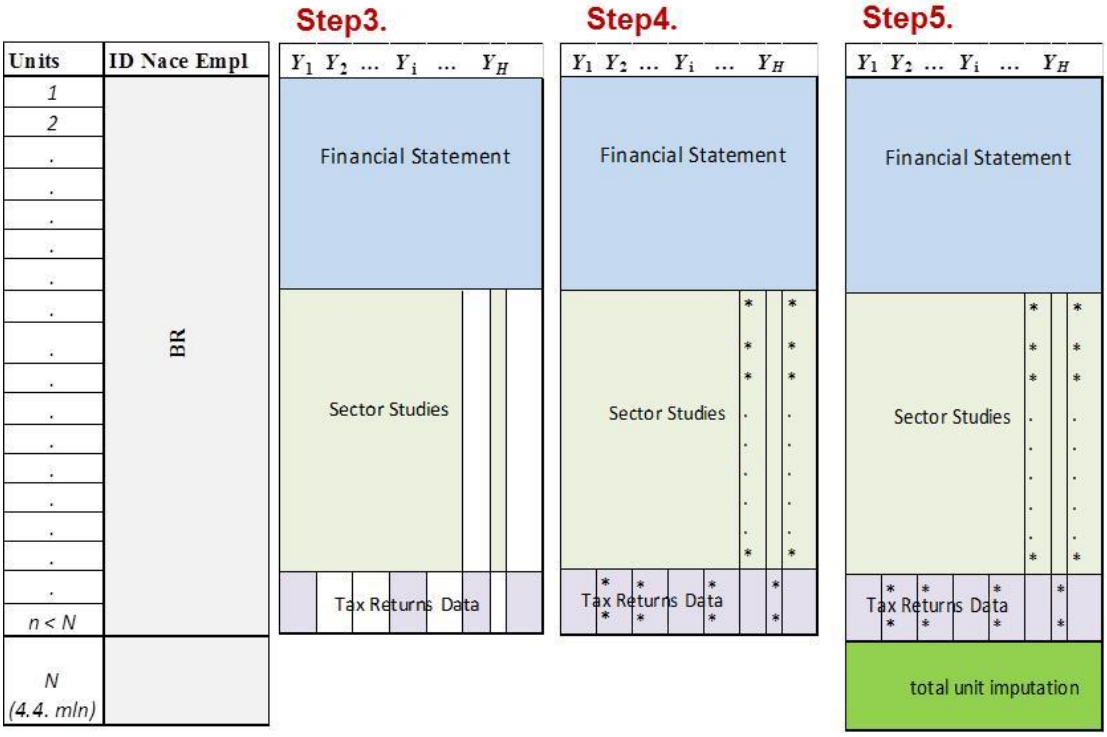
In Frame-SBS production process, Strategy B has been chosen, resulting in different coverage rates w.r.t. the various sub-populations of enterprises (Figure 2).
After the first integration, for each variable the coverage has been measured w.r.t. the whole target population and different groups of variable have been identified:

- Set of BR variables. A set of variables coinciding to those of the Businesses Register (BR): economic activity (Nace) and Employment (Emp) of each enterprise
- Set of core variables. The set of core variables Yh (h=1,...,H; H < K) that are the variables "highly" covered by AD, so that the integration of different AD covers 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 Yj (j= H+1,…, K) components of the core variables, which are not properly represented by AD (Step 6).

**Step 4.** Prediction/imputation of missing values of core variables treated as partial nonresponse for the n<N units covered by AD (Di Zio et al., 2016). In Figure 2, the integrated dataset for each unit i (i=1,..,n) of each core variable h (h=1, …,H) is represented.

**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 2) 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 2. BR and core variables – Step 3, 4, 5**




**Step 6.** Estimation of the components variables jointly using the SME sample survey data and the information on the core variables achieved at the previous step (as auxiliary information). For every variable Yj (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.

# III. The proposed quality evaluation framework

## A. Theoretical proposal

8. Starting from the analysis on how to apply the TSEadm to the Frame-SBS production process, a number of issues to be taken into account emerged, and some issues that need 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 TSEadm 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.

9. Therefore, we propose a general quality framework composed by two main Phases, which second Phase is composed by two sub-phases:

1. Assessment of AD w.r.t. administrative purposes
2. Combination/re-use/integration of AD for statistical purpose:
    2a. Assessment of AD w.r.t. statistical purposes

2b. Assessment of the combined AD for statistical purposes.

A third phase to evaluate the final outputs is needed. It will be introduced and exploited once the initial phases are settled. In the following, we briefly describe phases 1 and 2.

Phase 1. Assessment of AD w.r.t. administrative purposes.

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, and 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. Therefore, an E&I process may be performed separately for every source. This phase coincides with Zhang's phase 1 (Zhang, 2012).

Phase 2. Combination/re-use/integration of AD for statistical purpose.

The reference point corresponds to the statistical population and to the statistical concepts to be measured.

Phase 2a. Assessment of AD w.r.t. statistical purposes.

Each AD 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. Assessment of the combined AD for statistical purposes.

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 (Zhang, 2012). 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.

## IV. Case study: Frame SBS

10. In this section, we describe the application of the general quality evaluation framework to the production process of the Frame-SBS register, by describing the link between the steps of its production process (see par. 2.2) and the phases of the evaluation framework (see par. 3.1).

This schema is useful to show where the decisions on the process are taken, and to monitor the entire process. As introduced, the whole system of indicators should help as guideline to identify potential sources of errors, to measure their effect on the output and to prevent them, in order to progressively improve the production process.

**Table 1. Frame SBS process, steps and phases**

| **Steps** | **Phase** | | |
|---|---|---|---|
| | **1. Assessment of AD w.r.t. administrative purposes** | **2 . Combination/re-use/integration of AD for statistical purpose** | |
| | | **2a. Assessment of AD w.r.t. statistical purposes** | **2b. Assessment of the combined AD for statistical purposes** |
| **1** | Quality assessment of each candidate AD source (*FS, SS, Unico, Irap*) | | |

| | | | |
|---|---|---|---|
| **2** | | Quality assessment of each AD source (*FS, SS, Unico, Irap*) in terms of SBS purposes | |
| **3** | | | Integration of AD sources (*FS, SS, Unico, Irap)* |
| **4** | | | Prediction/imputation of the missing values of the *core* variables for partially uncovered units |
| **5** | | | Prediction/imputation of the *core* variables for totally uncovered units |
| **6** | | | Estimation of the *components* variables |

## V. Case study: Frame SBS, E&I indicators

11. In this section we focus on the indicators relative to the E&I actions that are performed during the Frame-SBS production process. More in details, as described in Section B, some E&I actions are performed during the first step of the entire process (Phase 1 in the quality framework) and during the fourth and fifth step (Phase 2b in the quality framework).
Table 2 and 3 report the (absolute) values of some selected indicators obtained from 2012 to 2014. Obviously, beside the evaluation of the indicators separately for each year, it is interesting to evaluate the changes over time in order to monitor the quality of the production process. Furthermore, when the changes are small, it is necessary to perform a transformation of the absolute values of the indicators to normalized values (i.e. change/range).

12. An interesting indicator for the Frame-SBS if the *Percentage of multiple records, by source*. We observe an increase of the quality from this perspective for both Unico and Irap, while the indicator does not change for SS and has a small decrease for FS.
It is interesting to note that, for FS, we register a quality increase. In fact, the *Proportion of units in FS w.r.t. the FS theoretical population in the BR* increases, and the *Proportion of units failing edit checks, by source* decreases. Furthermore, the *Proportion of imputed values, by source* also slightly increases, probably due to the fact that an higher percentage of inconsistencies could be solved by imputation.
This analysis can be repeated for each AD.

**Table 2. Frame SBS process, Phase 1, E&I indicators**

| Indicator | | Year | | |
|---|---|---|---|---|
| | | *2012* | *2013* | *2014* |
| ***Objects. Selection error*** | | | | |
| Proportion of units in FS w.r.t. the FS theoretical population in the BR | | 8.43 | 10.55 | 11.39 |
| Proportion of units in the source w.r.t. the BR population, by source | | | | |
| | *SS* | 20.71 | 19.98 | 20.75 |
| | *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 |

13. In Phase 2b, we the E&I process is applied to the integrated dataset.
It is interesting to observed that, while in 2012 and 2013 almost all indicators are quite stable, in 2014 they show a general worsening. In fact, we observe that both the *Proportion of missing units of the SBS population in the integrated dataset (under-coverage)* and the *Proportion of units with at least one imputed value* increase. Furthermore, the E&I process in 2014 causes an higher *Modification rate* and *Net imputation rate* for the most important variables of Frame-SBS, and has an higher *Impact on aggregates.* Since the indicator *Proportion of units of the SBS population in the integrated dataset, by source* is quite stable over time for both FS and SS (representing the two most important sources in terms of number of units in the integrated dataset), the overall quality decrease can be due to a decrease of the quality of SS (see Table 2).

**Table 3. Frame SBS process, Phase 2b, E&I indicators**

| 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 |

## VI. Conclusions and open issues

14. In this paper a comprehensive framework for the quality assessment for statistical processes using administrative data is proposed. 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 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.

15. 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 TSEadm 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 of external data time by time.

17. Different E&I actions can be performed at different stages of a complex production process, for example when different AD and survey data have to be combined in order to produce a register. The proposed quality framework allows to disentangle the effect of the E&I actions at each step. This is useful to: (i) monitor the development and the final quality of new processes, (ii) help practical decisions about statistical design and monitoring.

16. 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.

## 5. References

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 Varriale R. (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.

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

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), Wien, 2-5 June.

Luzi O., Rocci F., Sanzo R., Varriale R. and Brancato G. (2016), Quality evaluation for statistical register: the Italian FRAME-SBS, European Conference on Quality in Official Statistics (Q2016), Madrid 31 May – 2 June.

Reid G., Zabala F. and 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.

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.