Assessing multi-source processes: the new Total Process Error framework

**Keywords:** multi-source processes, quality assessment, data integration

# Introduction

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.

Many new experiences have delivered new important results, that can be considered at the basis of the modernization that NSIs are facing. In this view, new processes are taking place and new methodological issues are arising.

Among others, a 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 including AD. This paper focuses on this issue, with the final aim to propose an evaluation framework, based on a system of indicators, useful to: (i) help practical decisions about statistical design together with monitoring its development (ii) define quality measure/measures of the new processes and of their statistical outputs.

The starting point to be taken as reference in literature is the adaptation of the life-cycle paradigm (Groves and Lyberg [1]) to a multi-source process proposed by Zhang [2] and applied by Zabala *et al.* [3], resulting in a two-phases life-cycle where for every phase every source of specific errors that can occur is identified. Afterwards, Reid *et al.* [4] applied the Total Survey Error (TSE) paradigm (Groves and Lyberg [1]) to the new realm of statistical production, which involves integrating and combining data from various sources. It builds on Zhang [2] extension of this TSE thinking to administrative and integrated data, proposing a three-phases framework: 1) a single source assessment, 2) an integrated data set assessment, and 3) an estimation and output assessment (TSE*adm*).

As case study for a critical application of the TSE*adm*, we chose the Istat statistical register Frame for Structural Business Statistics (Frame-SBS) ([5],[6],[7]), supporting the annual estimation of Structural Business Statistics (hereafter SBS) on enterprises’ profit and loss accounts.

The analysis of the proposed framework in terms of life-cycle of a multi-source process based on integrated AD like the Frame-SBS one has led to some reflections. Nevertheless, as a surprise, the application of the scheme as it is fails to fully represent the production process of the Frame-SBS register. Hence, more motivation arose to enhance deeper studies of the reasoning under the proposed framework, thus confirming Zhang’s assertion that we should think to be in “a pre-Neymann era” and that 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” ([4]).

By applying the TSE*adm* to the Frame-SBS process, we realized that the entire production process would not be monitored in some points, that are considered instead to be delicate. Furthermore, trying to cover also the third phase, as Reid *et al.* did, we realized that the type of final output that can be defined from the Frame-SBS process is not unique (register of microdata based only on the use of different AD, register based on the integration of different AD with imputated missing data, estimates based on registers data, etc.).

Hence, further discussion and reasoning should be enhanced to achieve a more comprehensive framework of multisource processes and outputs. As first ideas the following issues could/should be tackled/highlighted (Rocci *et al.*[8]): (i) there is a lack of a well-defined vocabulary to distinguish which kind of data, processes and outputs are involved in each phase, (ii) there is a need to define and to distinguish different kinds of statistical outputs that can be obtained based on the integrated use of AD, (iii) to consider how the processes are not still standardized, so that a quality evaluation framework should be designed in order to be as much flexible and general as possible.

The point we started from is how to describe a process based on the use of external data (that come in many different varieties as can be combined in different ways) in a structured way. In this context, while the Generic Statistical Business Process Model (GSBPM) gives clear guidelines on the conduction of an entire survey, for these new processes we do not have a standardized and systematic way to act. In recent literature, [9] guidelines for multi-source statistics are provided.

In order to generalize the quality evaluation framework for multi-source processes using AD described in the current literature, we propose to: (i) take distance from the “survey” reasoning and name it Total Process Error framework (TPE hereafter), (ii) split the second phase of the two-phases life-cycle ([2]) into two sub-phases, in order to describe the process of combining different data sources, (iii) use a matrix connecting the steps of the entire production process and the phases of the quality evaluation framework in order to identify every potential error source. The final aim is to provide a flexible tool to evaluate the quality of complex processes.

In the following, the case study Frame-SBS is showed and the TPE framework is presented.

# Frame SBS: a short description

The statistical register Frame-SBS, designed to meet 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 *core* variables (such as Turnover, Costs, etc.) are covered by a number of AD sources, providing 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). SBS variables which are *components* of the *core* variables are not properly represented by AD and are estimated also exploiting survey data. The production process of the Frame-SBS is composed by six main steps, and it represents the starting point to produce SBS.

# The proposed quality evaluation framework: the TPE framework

Starting from the analysis on how to apply the TSE*adm* 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.

As a first step, we propose to split the second phase of the two-phases life-cycle from [2] into two sub-phases, in order to describe the process of combining different data sources. Therefore, the TPE results to be composed by two main Phases, which second Phase is composed by two sub-phases:

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 ([2]).

1. *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.

* 1. *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.

* 1. *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 ([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.

As second step, we propose to use a connection scheme between the steps of the entire register production process and the phases of the TPE framework as a tool to identify every source of specific errors. This schema is useful to show where the decisions on the process are taken, and to monitor the entire process. Table 1 represents the application of the general TPE framework to the production process of the Frame-SBS register, by describing the link between the steps of its production process and the phases of the evaluation framework. The schema can also be used as guideline to build a system of indicators useful 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.

This proposal has to be considered as an initial step of a complex project. In particular, a final phase has to be defined in order to complete the framework, consisting in the classification of the possible outputs of a multi-source statistical processes involving the use of AD, and in the development of a proper set of quality measures for assessing them.

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 |

# References

1. R. M. Groves and L.E. Lyberg, Total Survey Error: Past, Present, and Future, *Public Opinion Quarterl*y, 74(1), (2010), 849-879.
2. L.C. Zhang, Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012), 66(1), 41-63.
3. F. Zabala, G. Reid, J. Gudgeon and M. Feyen, Quality Measures for Statistical Outputs using Administrative Data, Statistical Methods, Statistics New Zealand (2013).
4. G. Reid, F. Zabala and A. Holmberg, Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ, Journal of Official Statistics, (2017), 33(2), 477–511.
5. O. Luzi and R. Monducci, The new statistical register Frame-SBS: overview and perspectives, Rivista di Statistica Ufficiale, N.1/2016 (2016), 5-14.
6. O. Luzi, F. Rocci, R. Sanzo, R. Varriale and G. Brancato, Quality evaluation for statistical register: the Italian FRAME-SBS, European Conference on Quality in Official Statistics Q2016 (2016).
7. O. Luzi, U. Guarnera and P. Righi, The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data, European Conference on Quality in Official Statistics Q2014 (2014).
8. F. Rocci, R. Varriale and O. Luzi, A proposal of an evaluation framework for processes based on the use of AD, Workshop on Statistical Data Editing UNECE 2018 (2018), 1-10.
9. T. de Waal, A. van Delden and S. Scholtus, Multi-source Statistics: Basic Situations and Methods, Discussion Paper CBS (2017), 2-38