A new Total Error Framework for multi-source processes

Fabiana Rocci - Roberta Varriale - Orietta Luzi
Istat

General context

Administrative Data (AD) and other External Sources (ES) contain much information related to many target phenomena

NSIs are moving towards *processes* where (integrated) AD represent as far as possible the primary source of information for a new production system based on the combination of different data sources

Furthermore, new processes based on the combined use of different type of ES are under evaluation

A new framework to assess the quality of the Official Statistics *multi-source processes* and their outputs is required

Motivating example: the Istat register Frame-SBS

The **statistical register Frame-SBS** is built for the annual release of statistics on loss and accounts of enterprises to satisfy the Eurostat SBS regulation aimed at describing the structure and performance of businesses across the European Union

Different AD sources provide SBS variables at micro level:

- the Financial Statements - FS
- the Sector Studies survey - SS
- the Tax returns - Unico
- the Regional Tax on Productive Activities - Irap

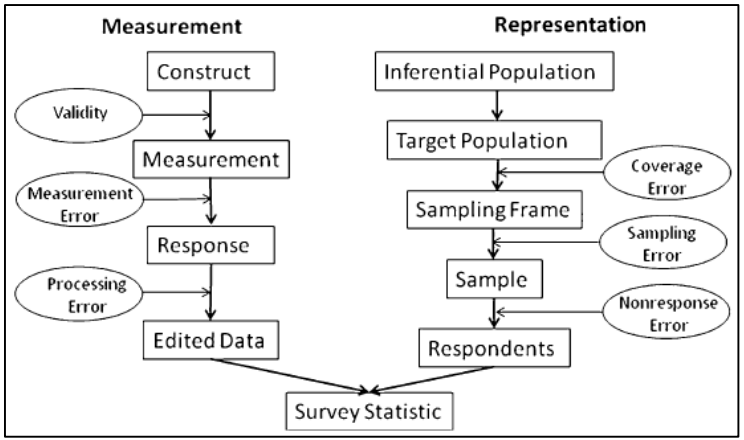
AIM: assess the quality of the **statistical register Frame-SBS**

References in literature

Starting point: Life-cycle of a survey

This approach aims at identifying the potential error sources *along the phases* of the survey process: conception, collection and processing till the final production of estimates (Groves *et al.*, 2004)

→ Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process



References in literature

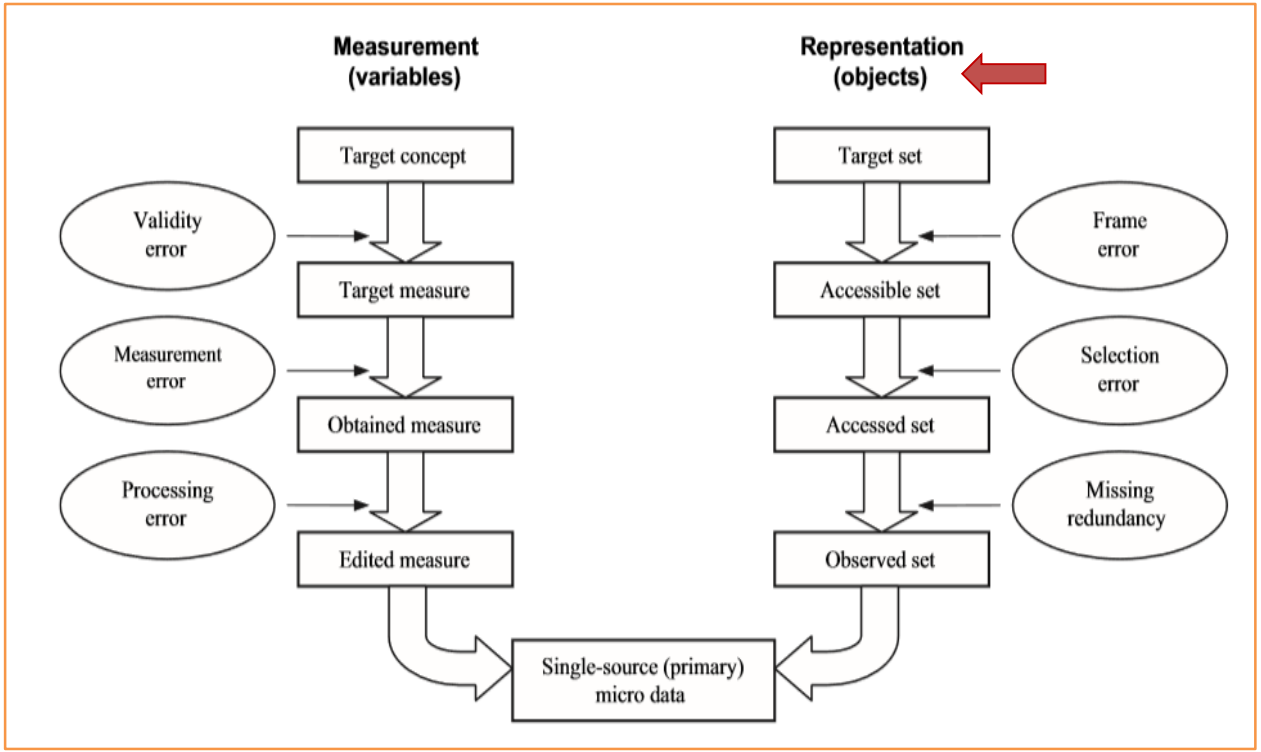
Two-phases life-cycle framework (for processes based on integrated AD)

proposed by Zhang (2012), applying a similar reasoning as the life-cycle for identifying errors, developing the idea in two different phases, each of them dealing with its specific target

1. each AD source is assessed w.r.t. its original target to measure its quality
2. the integrated AD sources are assessed w.r.t. the specific statistical purpose

Two-phases life-cycle of processes based on integrated micro-data

Phase 1

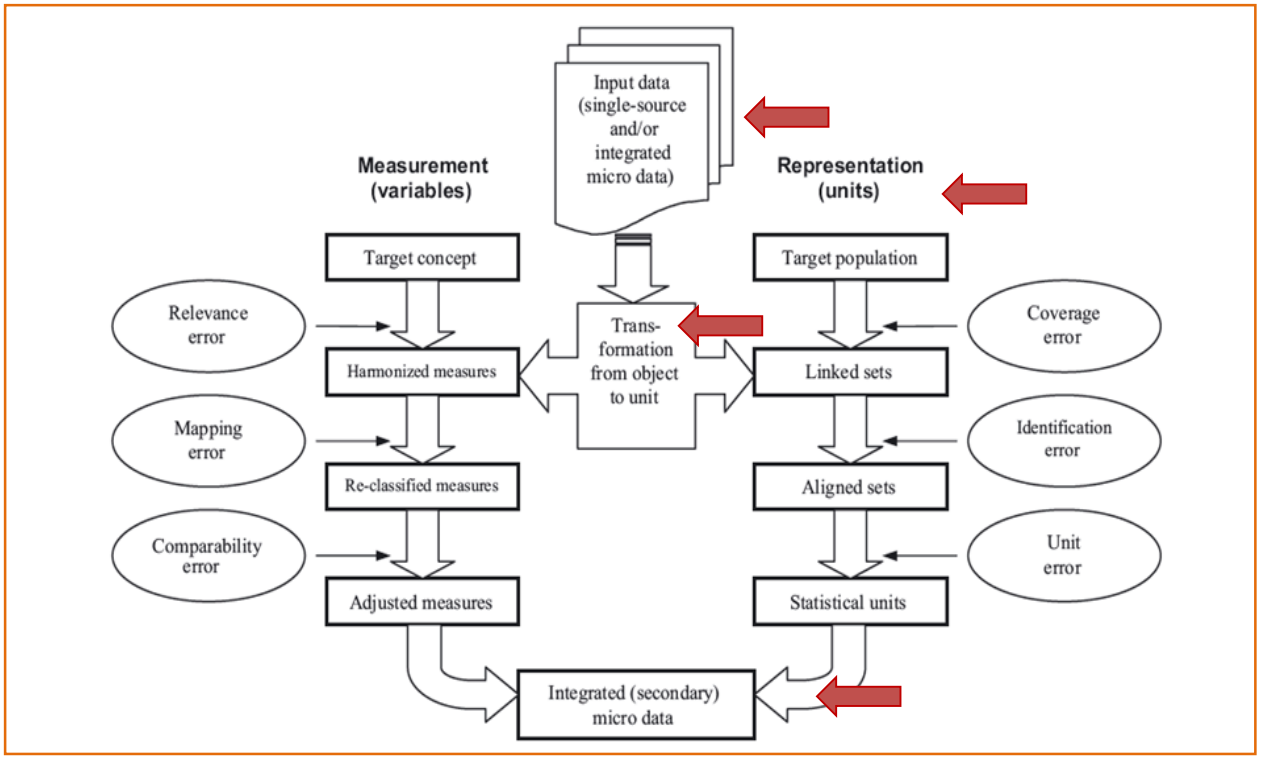


Step

Potential error

Two-phases life-cycle of processes based on integrated micro-data

Phase 2



Step

Potential error

References in literature

Three-phase life-cycle framework

Reid, Zabala and Homberg (2017) propose a three-phase framework applying the life-cycle paradigm to the new system of statistical production

A third phase is introduced to take into account errors that can arise in the creation of the final output → Total Survey Error in the context of the combined use of AD supplemented by survey data (*TSEmultisource - TSEms*)

Motivating example: the Istat register Frame-SBS

Through the application of the TSE*ms* approach,

we got lost in some part of the two-life cycle phases!!!

Therefore:

- first, we started back from describing every step of the process in order to clarify how to apply the quality evaluation framework
- then, we made some proposals to “modify” the *TSEms*
- nowadays, we think we need some enhancement to “re-think” the *TSEms*

Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

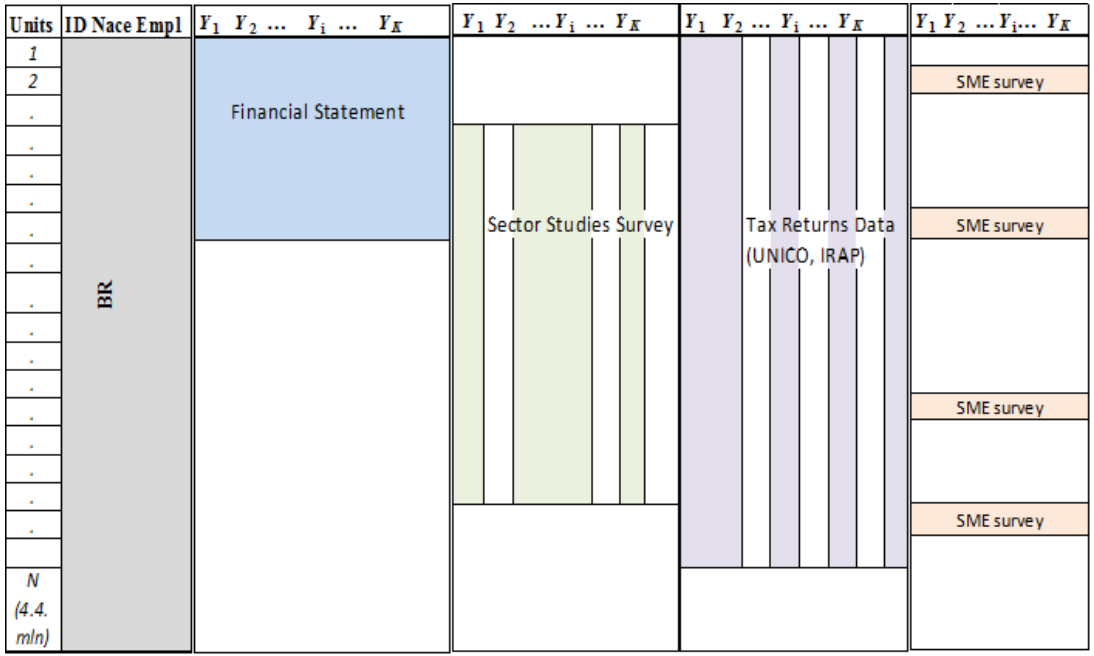
Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

Frame-SBS: steps

Step 1 - Step 2



Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

Step 3. Main decisions are taken about how to integrate AD sources

Two different alternative strategies of integration could be applied:

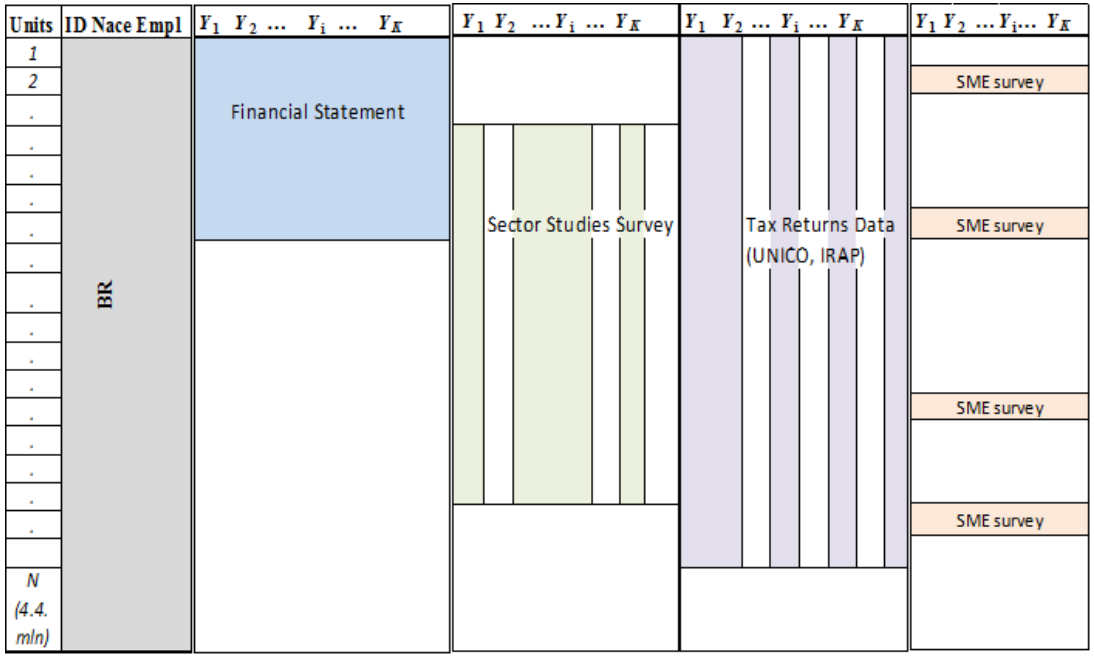
- **Strategy A**:
  For every unit, the final record is achieved by an integration of all the AD sources
  Strategy A maximizes the exploitation of the overall quantity of information

- **Strategy B**:
  For every unit, the final record is achieved by the use of only one AD source, with a "priority" assigned to every AD source
  Strategy B maximizes the internal coherence of the integrated dataset

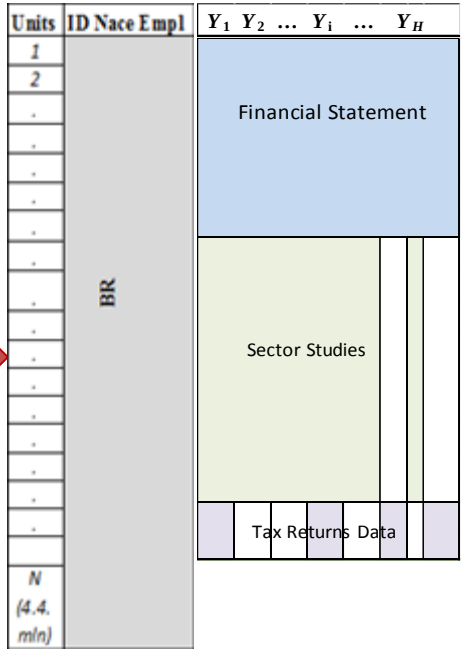
In Frame-SBS production process: Strategy B has been chosen

Frame-SBS: steps

Step 1 - Step 2



Step 3



Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

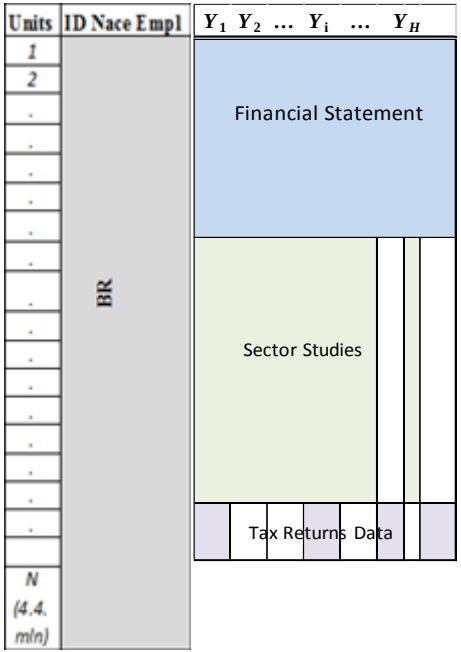
Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

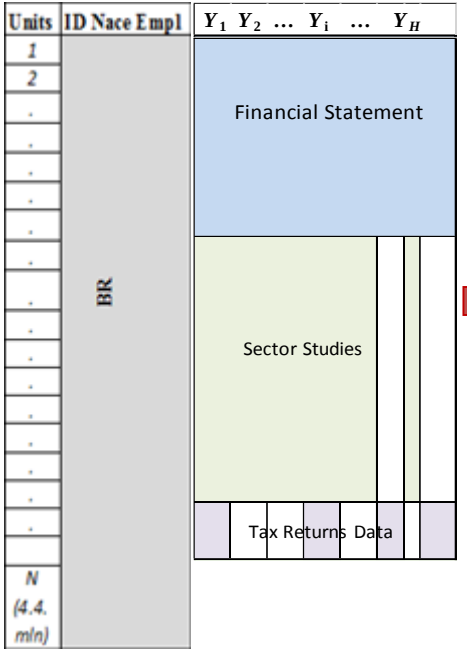
Frame-SBS: steps

Step 3

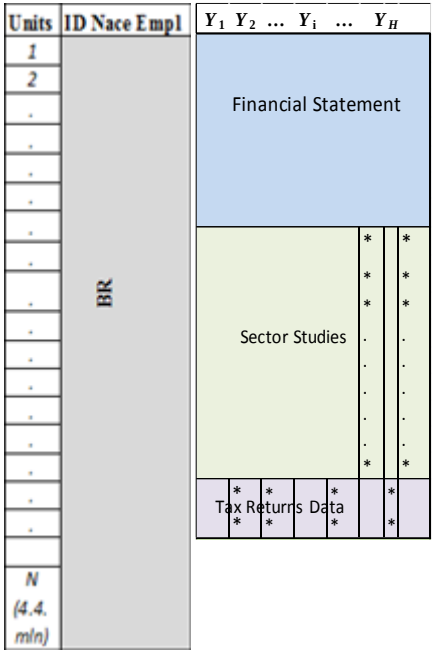


Frame-SBS: steps

Step 3

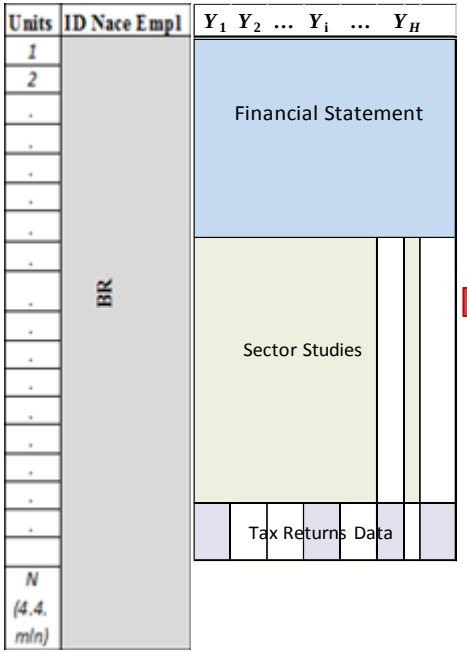


Step 4

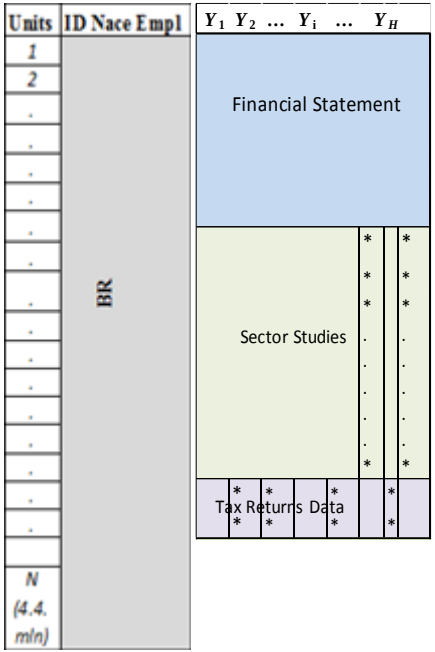


Frame-SBS: steps

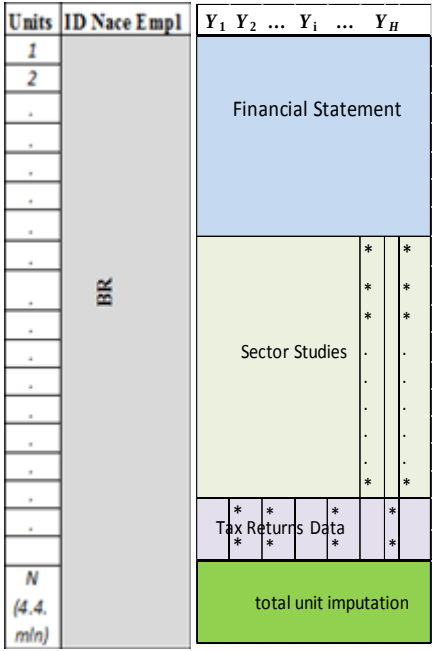
Step 3



Step 4



Step 5 - statistical register



Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

Frame-SBS: steps

Step 1. A quality assessment process on each candidate AD source

Step 2. A mapping of the coverage for every source for the whole system w.r.t. :

- the *K* required variables, grouped in *core* and *component* variables
- the target population

Step 3. Main decisions are taken about how to *integrate* AD sources

Step 4. Imputation of the partial missing data of *core* variables on the integrated AD

Step 5. Imputation of totally missing units to cover the total SBS target population

Step 6. Estimation of the *components* variables (using sampling information on Small and Medium Enterprises)

Step 7. Computation of SBS

Issues highlighted, Frame-SBS

- Two main different statistical processes can be distinguished, one for the *core* variables and one for the *components* variables
- About the integration of the AD sources: alternative strategies could be theoretically adopted
- It is completely different to evaluate the Frame-SBS in terms of different outputs:
  - ✓ The statistical register obtained only by the integration of the AD (Step3.)
  - ✓ The statistical register obtained also through the imputation of microdata of all the core variables (Step5.)
  - ✓ SBS final estimates using different methodologies for each group of variables (and, in some cases, for each variable) (Step7.)

Issues highlighted, general context (1)

- The second phase of TSE*ms* should be further enhanced to trace the actual assessment/integration/treatment process and better assess quality

The introduction of an explicit phase of integration

Phase 1. Assessment of each AD source w.r.t. the administrative purposes

This phase corresponds to phase one of Zhang (2012)

Phase 2a. Assessment of each AD source w.r.t. the statistical 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

This phase releases the input of the phase two of Zhang (2012)

Phase 2b. *Integration* of the AD sources

In this phase, the integrated database 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

Issues highlighted, general context (2)

- We developed an operative tool: a matrix that cross-classifies the process steps with the framework phases

This matrix provides a tool in order to gather information on the exact step of the process where the errors (potentially) originate; this also allows to evaluate the effect of different process strategies

Thus, the matrix can be considered as a "dashboard" associated to the process highlighting its critical aspects

Operative tool

Steps of the statistical process	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
Phase1			
Phase2			
…			
PhaseN			

This instrument can help us to understand if there are still some issues to be assessed…

Frame-SBS. Step and phases: a matrix representation

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		
2		Quality assessment of each AD source in terms of SBS purposes	
3			Integration of AD sources
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

?

Future work (1)

- We need to improve the vocabulary to better distinguish which kind of *data, processes and outputs* are involved in each phase

We believe it is important to face the problem of the lack of a comprehensive and clear terminology that would help in:

- classifying which type of *output* can be assessed,
- defining at which *stage of the process* an overall measure of quality can be delivered
- deciding what kind of *methodology* has to be used

This is an important step to allow to generalize this framework to processes using different types of multi sources

Future work (2)

- There is a need to define and to distinguish different kinds of statistical outputs that can be obtained based on the use of multi sources: this is necessary in order to identify the most appropriate quality indicators in the different contexts

In the future we will study:

- an output classification (such as: statistical register, estimates based on a register, etc.)
- how to assess the process outputs (third phase)

Future work (3)

- The final result should be a comprehensive framework including a set of indicators following the whole production process

  In addition, we will evaluate the possibility to identify suitable synthetic indicators for each phase and for each output

What do you think about that?

- Split of phase two
- Matrix cross-classifying the process steps with the framework phases
- Modify the name of the proposed quality evaluation framework from TSE (Total Survey Error) to **TPE** (Total Process Error) to underline that we need to consider that different kinds of errors can affect a process based on a combination of different sources

Thank you for your attention!

Fabiana Rocci, *rocci@istat.it*

Roberta Varriale, *varriale@istat.it*

Orietta Luzi, *luzi@istat.it*