Computational and Methodological Statistics - ERCIM 2018
Pisa, 16 December 2018

Quality evaluation of statistical processes based on administrative data

Roberta Varriale - **Fabiana Rocci** - **Orietta Luzi**

Istat

General context

Administrative Data (AD) and other external sources contain much information related to many target phenomena

National Statistical Institutes (NSIs) are moving towards processes where (integrated) AD represent as far as possible the *primary* source of information

AIM:

A new framework to assess the quality of the Official Statistics multi-source processes and their outputs is required. The quality framework should support both the *design*, *monitoring* and the *documentation* of the statistical processes.

Outline

- Presentation of the main literature
- The Istat register Frame-SBS as case study
- Issues highlighted by the analysis of the case study
- Proposal for change to the TSE*admin* → the TPE
- Conclusions and open issues

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* until the *final production of estimates* (Groves, Fowler, Couper, Lepkowski, 2004)

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

proposed by Zhang and applied by Zabala, 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

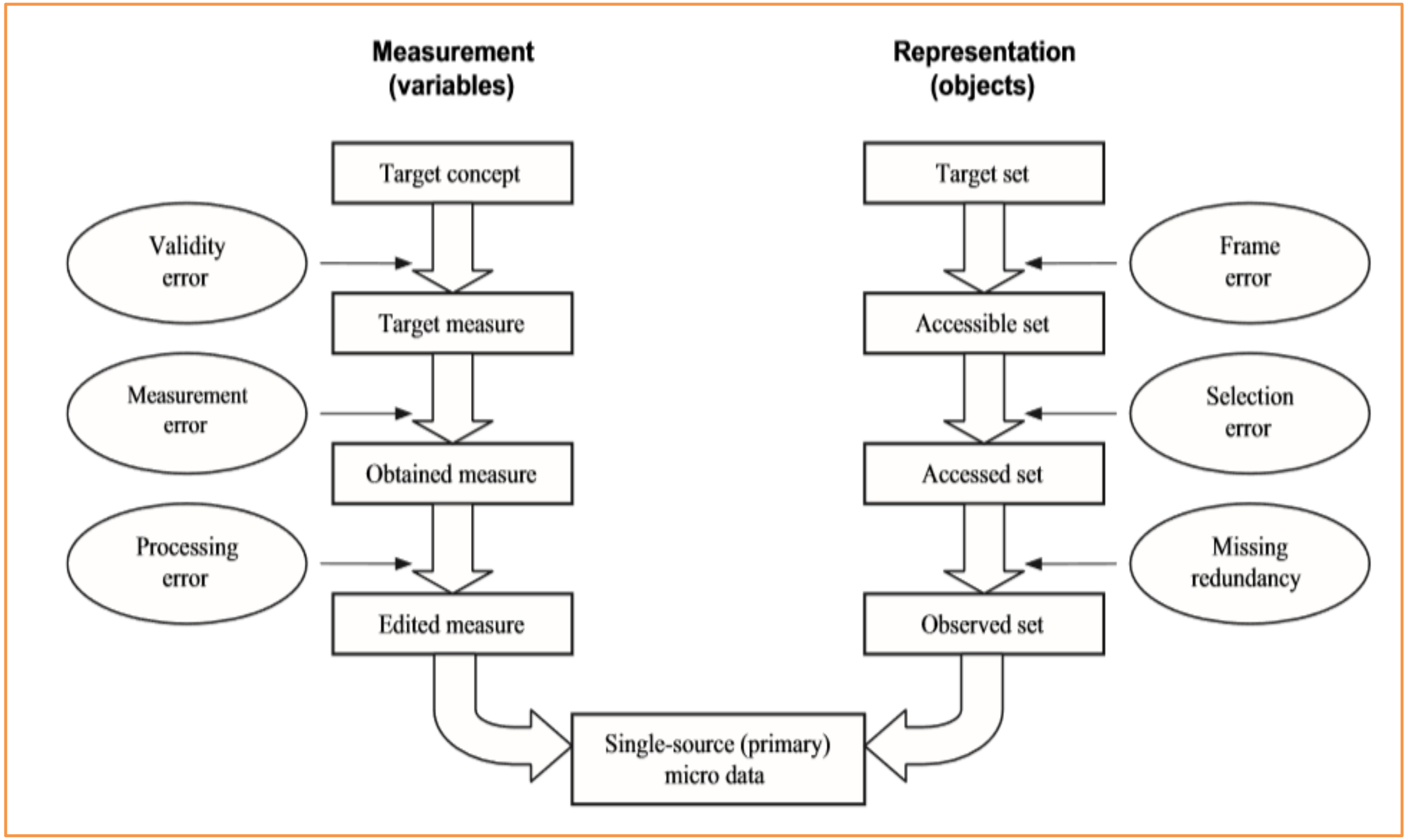
1. the quality of each AD source is assessed w.r.t. its original (administrative) target
2. the integrated AD sources are assessed w.r.t. the specific statistical purpose

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

Phase 1

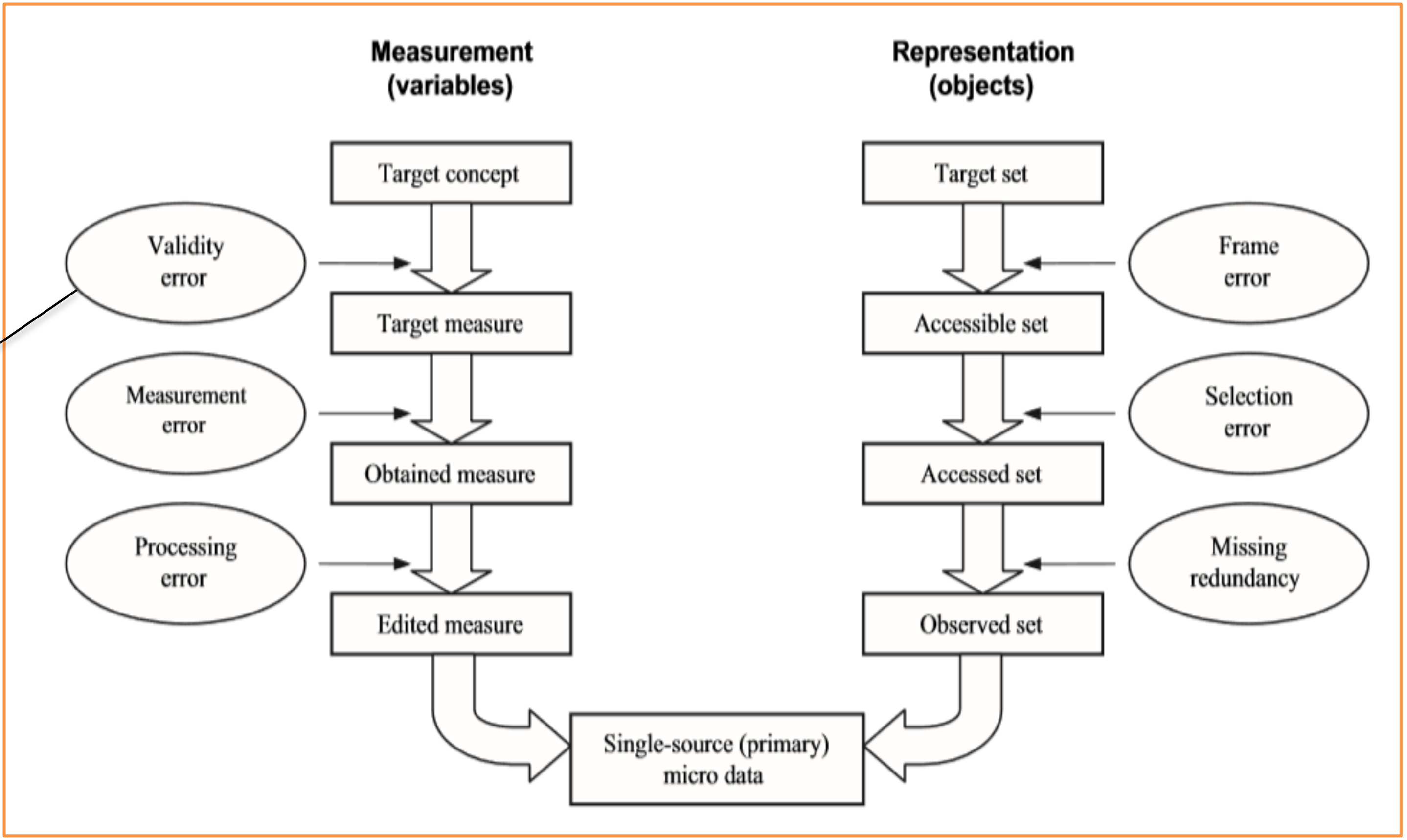
Step

Potential error



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

Phase 1



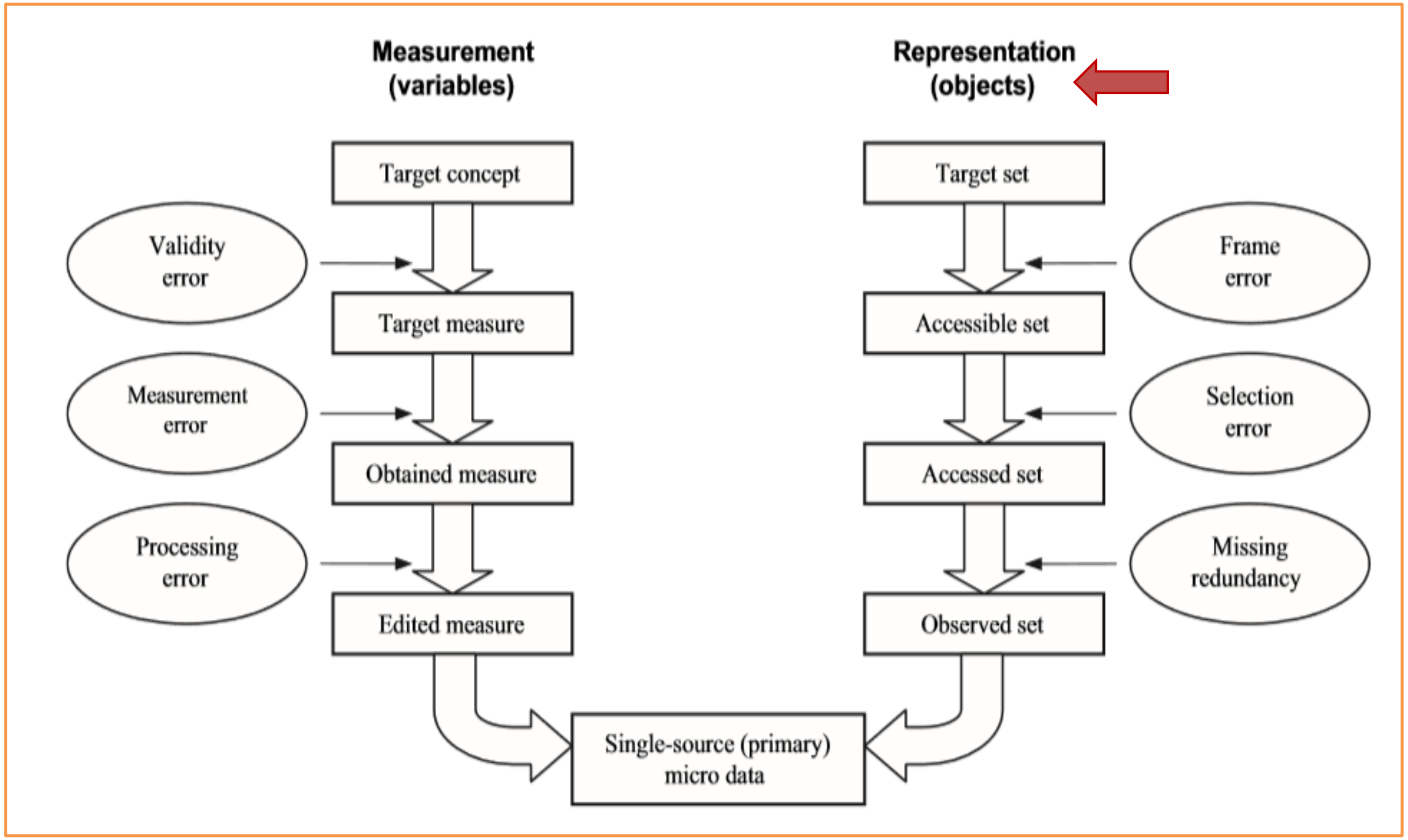
Connected to each potential error, there are indicators

Step

Potential error

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

Phase 1

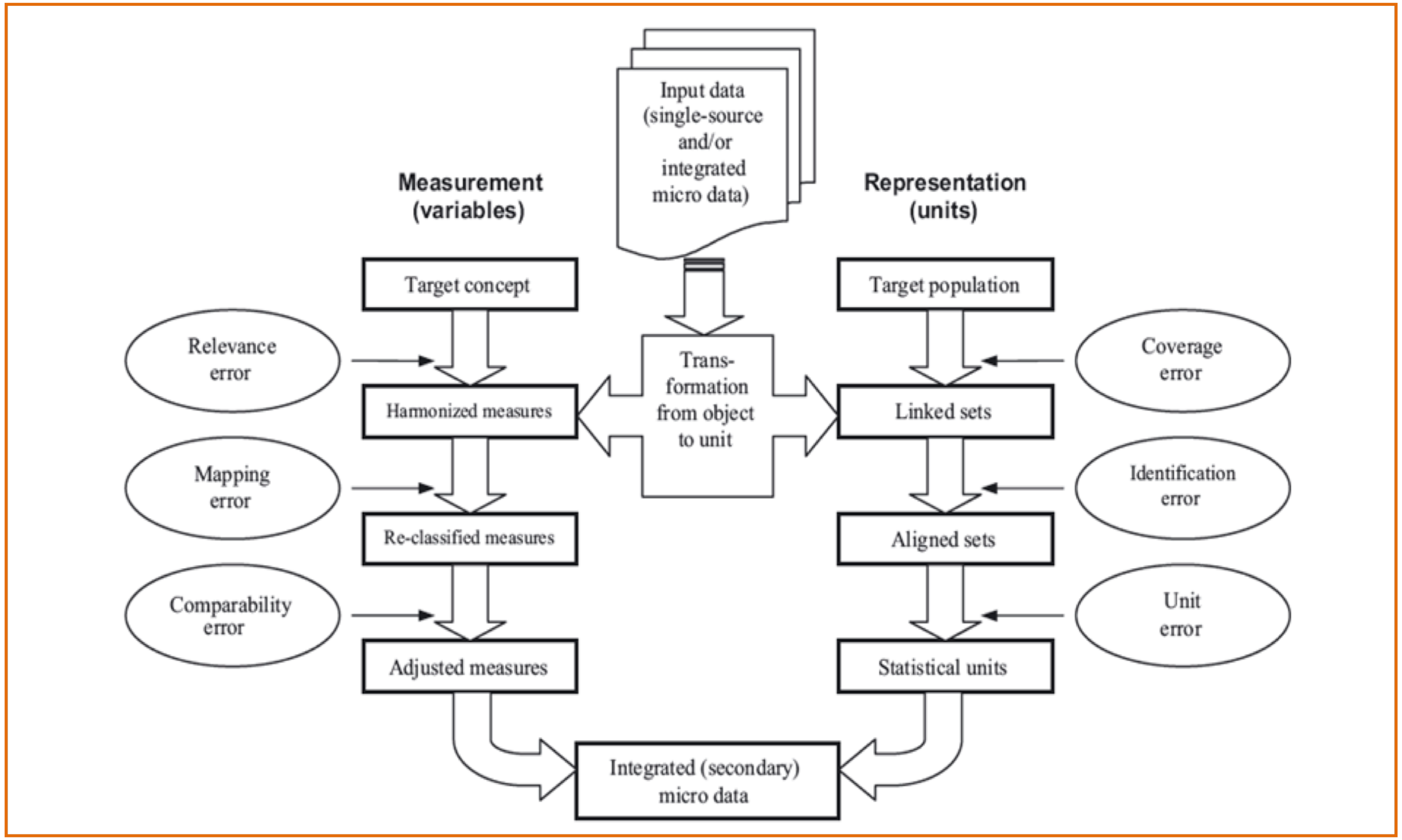


Step

Potential error

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

Phase 2

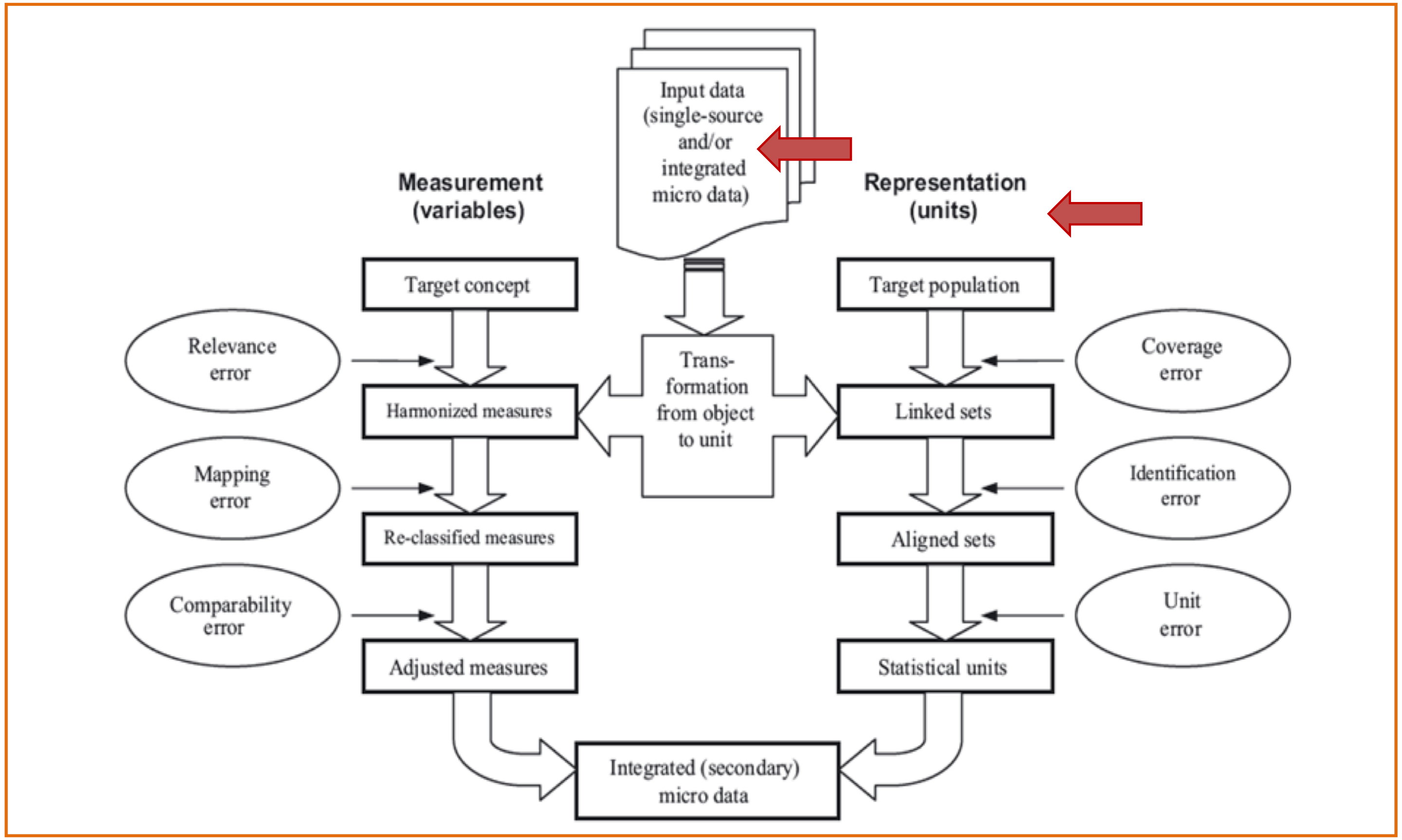


Step

Potential error

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 introducing the expression *TSEadmin* (Total Survey Error in the context of the use of AD supplemented by survey data)

Case study: 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: evaluate the quality of the **statistical register Frame-SBS**

… how? what does it mean?

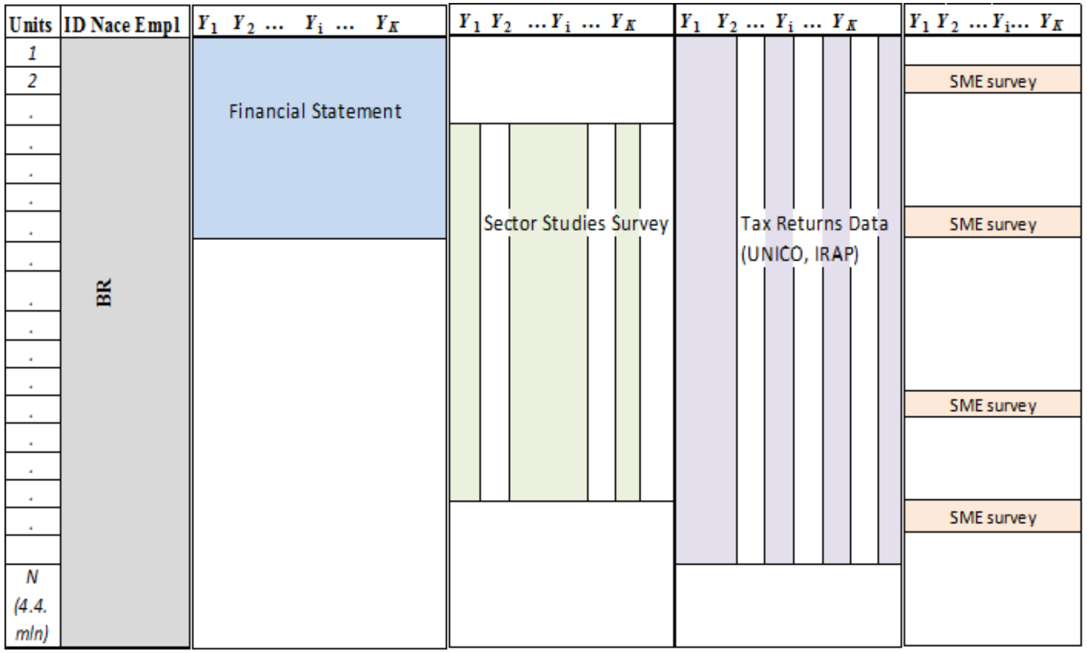
Frame-SBS: steps 1-7

Step 1. A quality assessment process on each candidate AD source w.r.t. *administrative* purposes

Frame-SBS: steps 1-7

Step 2. A mapping of the coverage for the whole system w.r.t. statistical purposes:

- the *K* required variables, grouped in *core* (*H<K*) and *component* variables
- the *target* (SBS) population

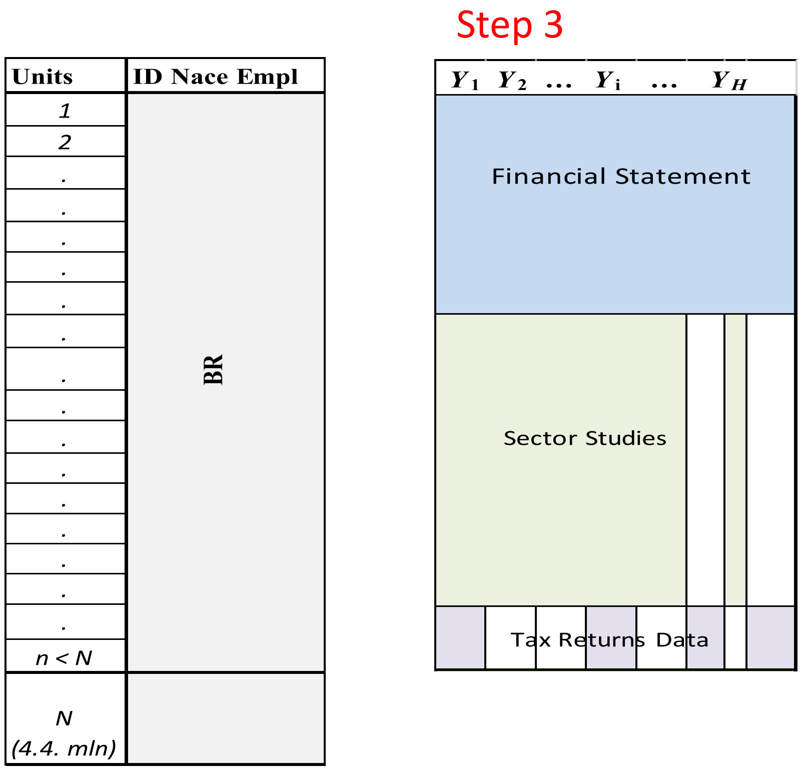


Frame-SBS: steps 1-7

Units	ID Nace Empl
1	BR
2	
.	
.	
.	
.	
.	
.	
.	
.	
.	
.	
.	
.	
.	
.	
.	
n < N	
N *(4.4. mln)*	

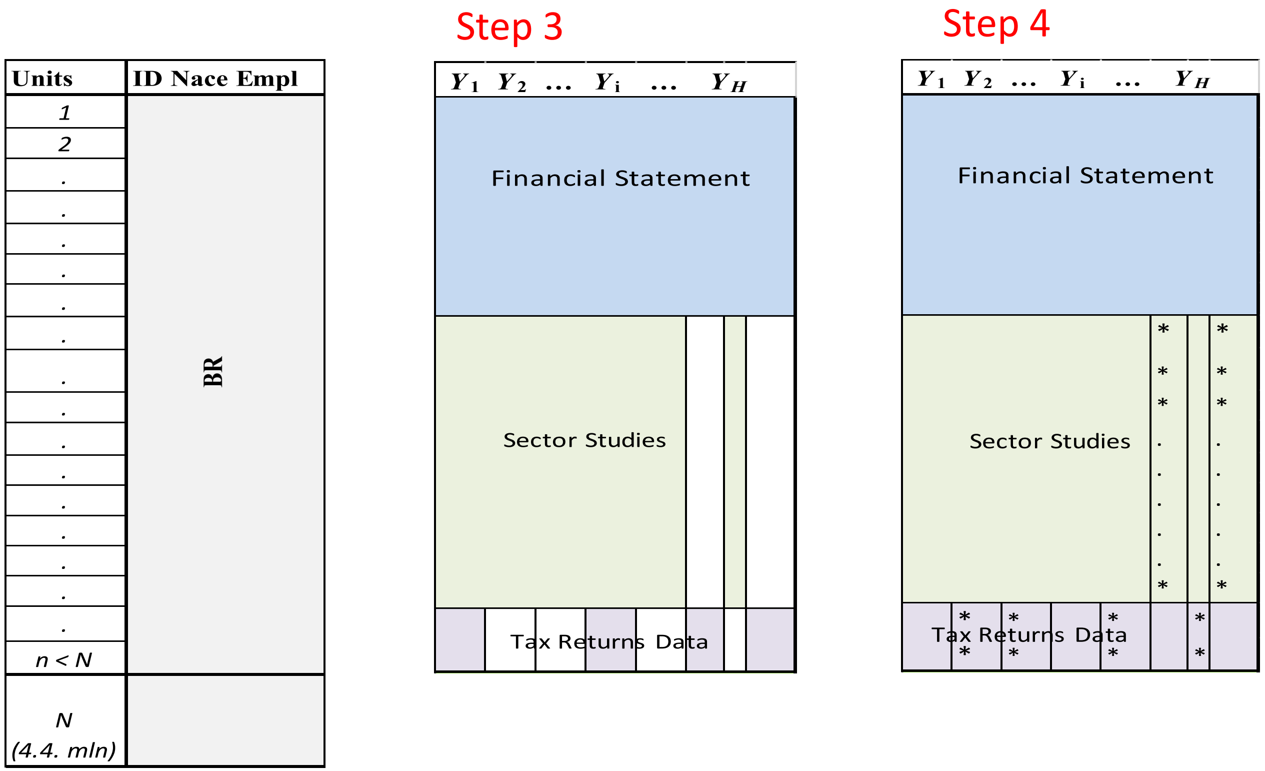
Frame-SBS: steps 1-7

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



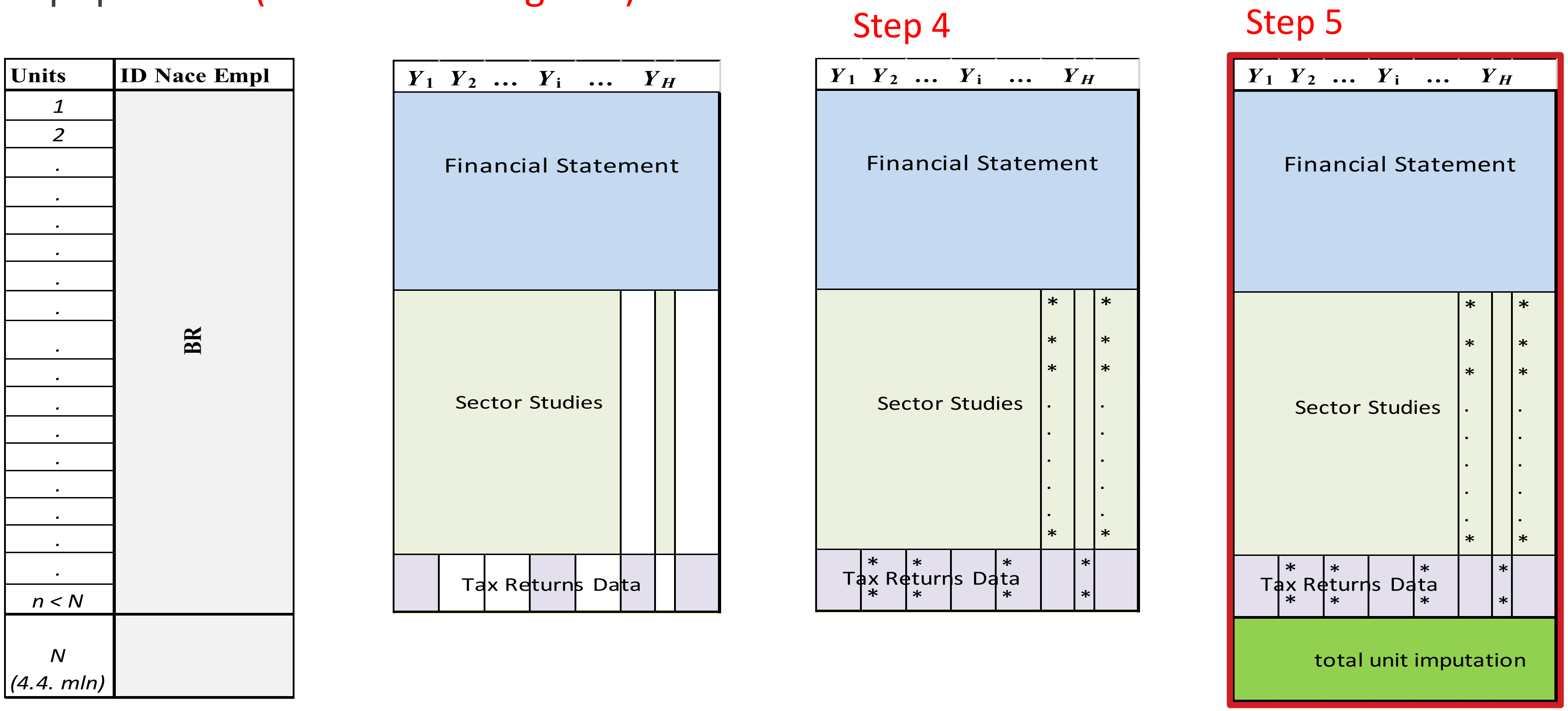
Frame-SBS: steps 1-7

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



Frame-SBS: steps 1-7

Step 5. Imputation of totally missing units of *core variables* to cover the total SBS target population (Frame - SBS register)



Frame-SBS: steps

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

Frame-SBS: steps

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

Step 7. Computation of SBS estimates

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 process in terms of several outputs:
  - ✓ The statistical register obtained at Step4. (covering a subset of the target population)
  - ✓ The statistical register obtained at Step5. - Frame SBS (covering all the SBS target population)
  - ✓ SBS final estimates using different methodologies for each group of variables (and, in some cases, for each variable) (Step7.)

Issues highlighted, general context

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

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

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

Preliminary results and future work

- Proposed name: 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

Preliminary results and future work

- Proposed name: 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

- We split the second phase of TSE*admin* into two sub-phases ➡

The introduction of an explicit phase of integration

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

This phase corresponds to phase one of Zhang (2012)

Phase 2a. Assessment of each administrative 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 sources

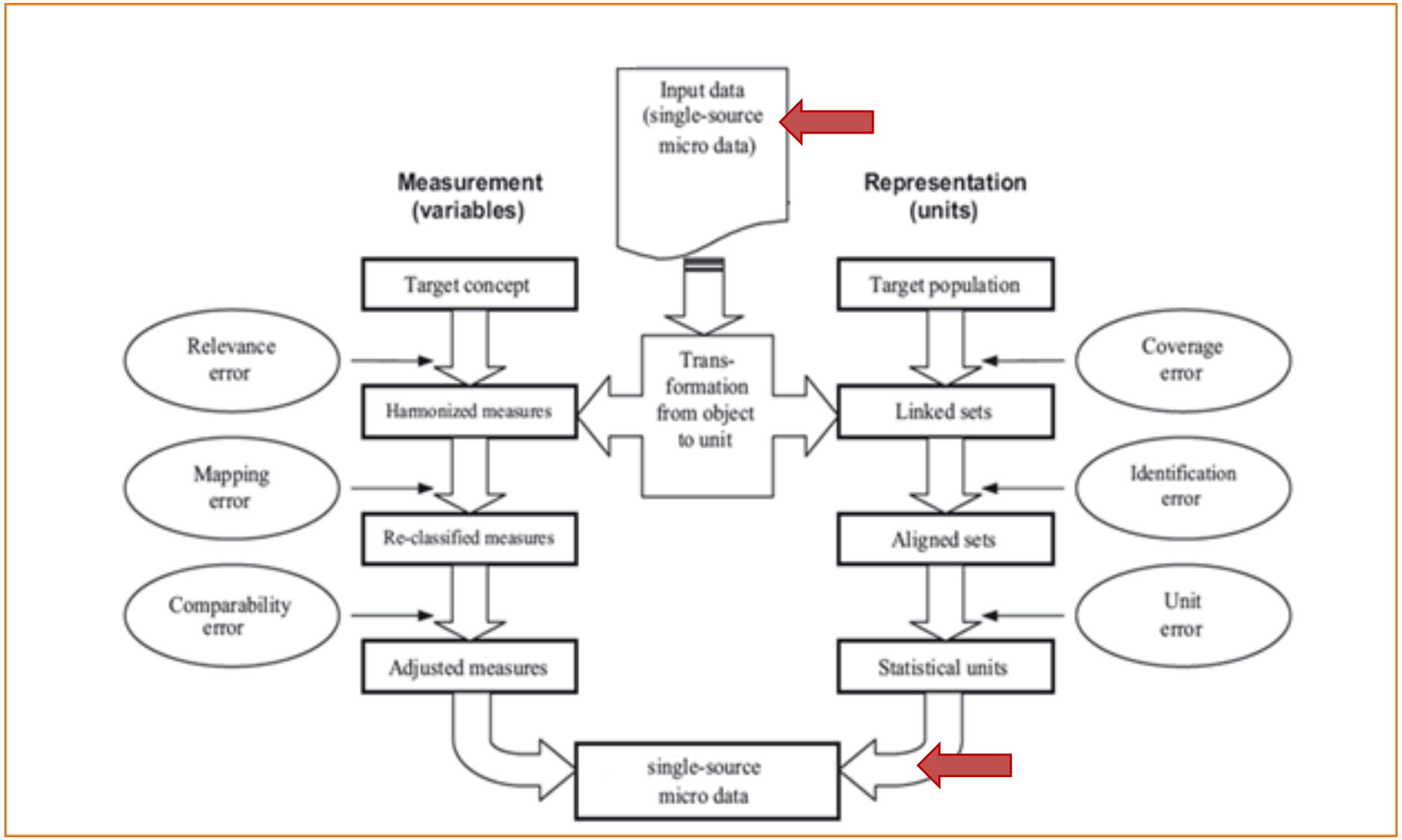
- In this phase, the integrated dataset is generated, and a further quality assessment is performed
- This phase partly corresponds to the Zhang's phase 2 (Zhang, 2012)
- Additional actions should be taken into account in order to allow the evaluation of the complete production process

The introduction of an explicit phase of integration

Phase 2a

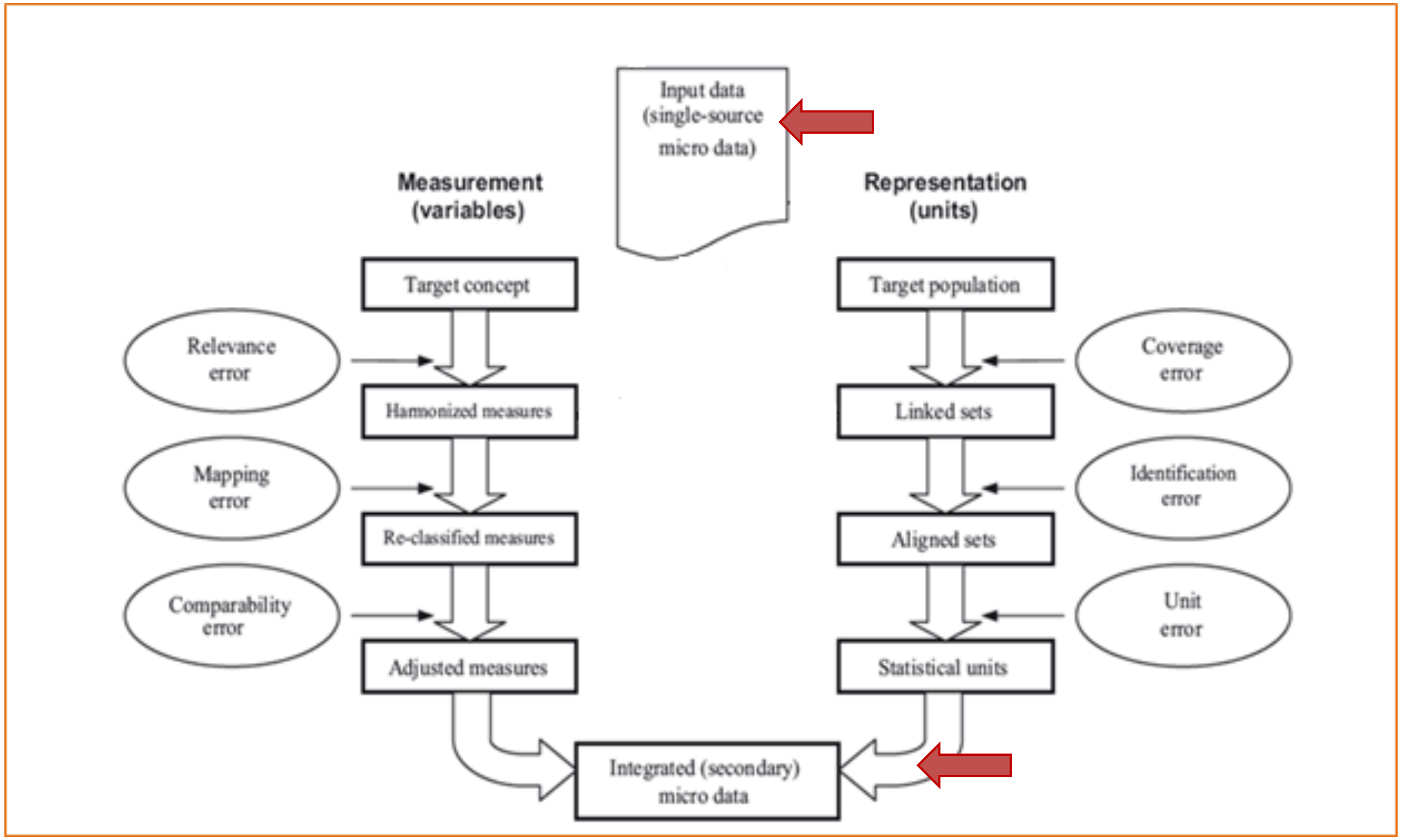
Step

Potential error



The introduction of an explicit phase of integration

Phase 2b



Step

Potential error

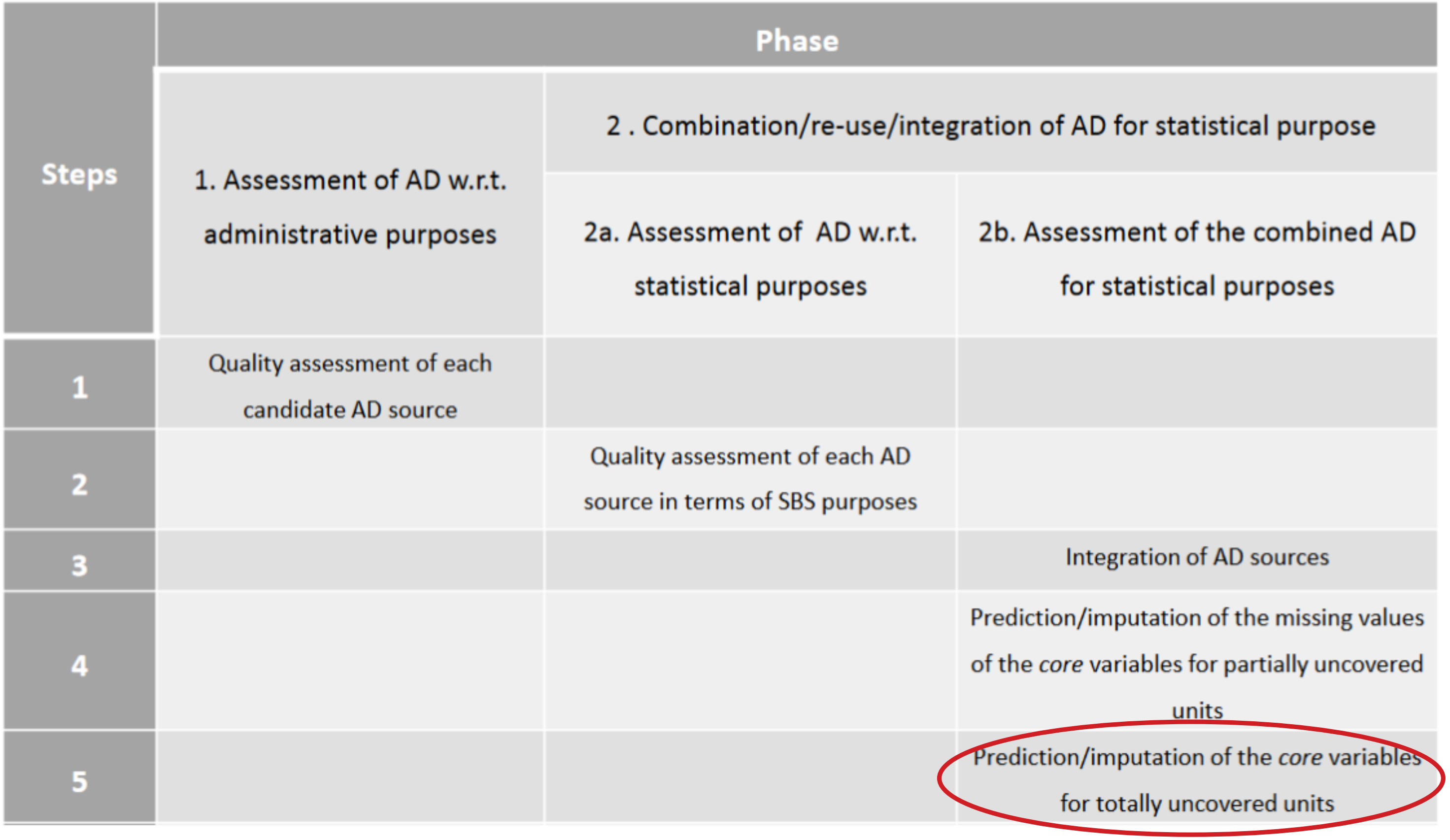
Preliminary results and future work

- 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 with the process highlighting its critical aspects

Frame-SBS. Steps 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

Preliminary results and future work

- 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

- In the future we will study an output classification, such as:

  statistical register, *estimates based on a register*, etc.

- After a clear taxonomy of the possible outputs is completed, the development of quality measures for the final outputs will be studied (for example the accuracy of final derived estimates)

Preliminary results and future work

- The final result should be a comprehensive framework including a set of indicators following the whole production process, with the aim of *monitoring* the process

- We will evaluate the possibility to identify suitable synthetic indicators for each phase and for each output

- The longitudinal monitoring of the process through the quality evaluation framework will allow defining proper thresholds for each simple and synthetic indicator

Thank you for your attention!

Fabiana Rocci, *rocci@istat.it*

Roberta Varriale, *varriale@istat.it*

Orietta Luzi, *luzi@istat.it*