2013 European Establishment Statistics Workshop

Integrating administrative and survey data in the new Italian system for SBS: quality issues

O. Luzi, F. Oropallo, A. Puggioni, M. Di Zio, R. Sanzo

Nurnberg, 9-11 September 2013

Istat

Outline

- The new Italian system for SBS
- The sources
- Quality issues
- Future activities

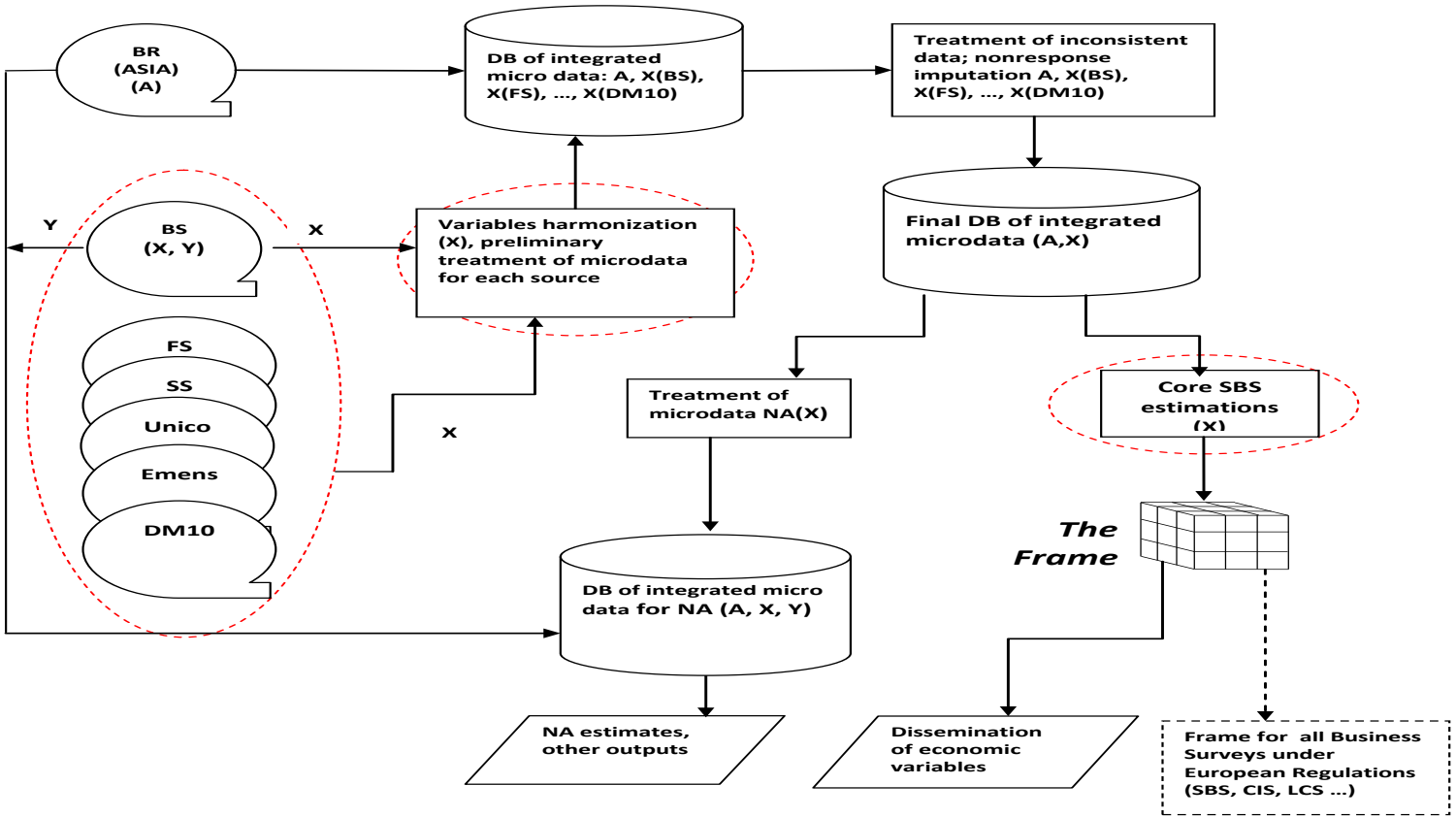
The new Italian system for SBS

New production process of a statistical database *(frame)* containing complete and coherent unit-level information to estimate key annual SBS (*production value, turnover, intermediate costs, value added, wages, labor cost, profits*)

- based on the integrated use of administrative and fiscal data, as primary source of information,…
- ….complemented by direct survey data (non covered sub-populations, or variables which are not directly available from external sources)

Release: October 2013

The overall strategy



Legend

X: Basic variables common to SBS and National Accounts (NA)
Y: Specific variables for National Accounts (NA)
A: Auxiliary variables from Business Register (BR)
BS: SME - Sample Survey on Small and Medium Enterprises; LE - Total survey on Largest Enterprises
FS: Financial Statements
SS: Fiscal survey on Sector Studies
Unico: Tax returns forms
Emens: Social Security Archive (individual level)
DM10: Social Security Archive (firm level)

The administrative and fiscal data sources

- **FS**: **Financial Statements** from Chambers of Commerce (about 700,000 units).
- **SS**: **Sector Studies survey** (includes each year about 3.5 million enterprises)
- **Unico**: **Tax data**, from the Ministry of Economy and Finance, based on a unified model of tax declarations by legal form
- **Social Security** data, from the National Security Institute (INPS), which includes firm level data and individual (employees) data on *wages and labor cost*.

The statistical sources

- The **Business Register (BR):** about 4.5 million enterprises. *Central role as reference list of the SBS target population*
- The **Business Surveys (BS)**
  - ✓ SME - Sample survey on Small/Medium Enterprises (1-99 persons employed)
  - ✓ LE - Total survey on Large Enterprises (100 and more persons employed)

The micro-data matrix X*

Estimation Domains	**SME survey**	**FS**	**SS**	**UNICO**	**S.Security**	**BR**
N° units	$X_1 \equiv X^*$	$X_2 \subset X^*$	$X_3 \subset X^*$	$X_4 \subset X^*$	$X_5 \subset X^*$	***ID codes,X*** $_{struct}$
N1						
N2						
N3						
N4						
N5						
N6						
N7						
….	…….	…….	…….	…….	…….	
Nk						
No source	?	?	?	?	?	

Istat

Sources' coverage

Administrative Source	Units (non overlapping)	%
Financial statements	718,239	16.2
Fiscal Purpose Survey	2,931,090	66.0
Tax Return Data	585,863	13.2
No source	208,688	**4.6**
Total	4,443,880	100.0

Assessing sources' quality and usability

(also based on recommendations from Essnet admin data and *BLUE-ETS)*

Coverage completeness of the source in terms of target population

Completeness degree to which the source includes information to estimate the target statistics

Consistency how much admin definitions of variables/units are close to the SBS ones

Accuracy: statistical adequacy of admin items for estimating target parameters

Punctuality: relates to the delivery of source data along time

Integrability: extent to which the source is capable of being integrated

Assessing sources' quality and usability: consistency and accuracy

Let X* be the target SBS and A_i the related information from the source *i*

For each enterprise, some of the X* variables may exist in more than one sources in different combinations, according to the dimension, the social security rules, the fiscal status etc.

The variables Ai in source ***i*** may coincide or may approximate the corresponding X*

Objective of the analyses

- Establish a hierarchy between the sources
- Identify the possible deterministic "corrections" from Ai to X*
- Identify the appropriate multivariate models so that $x^*_{ij} = f(a_{ij}.....)$

Assessing sources' quality and usability

(i) Assessing **consistency**: closeness between the A_i definitions in source *i* with respect to the X* definitions

→*variables harmonization*

(ii) Assessing **accuracy**: analysis (on overlapping admin and survey data) of A_i and X* distributions to assess possible "biases"

- at micro-data and distribution level
- at estimation level

and

- *selective editing*

→*further harmonization; data modeling*

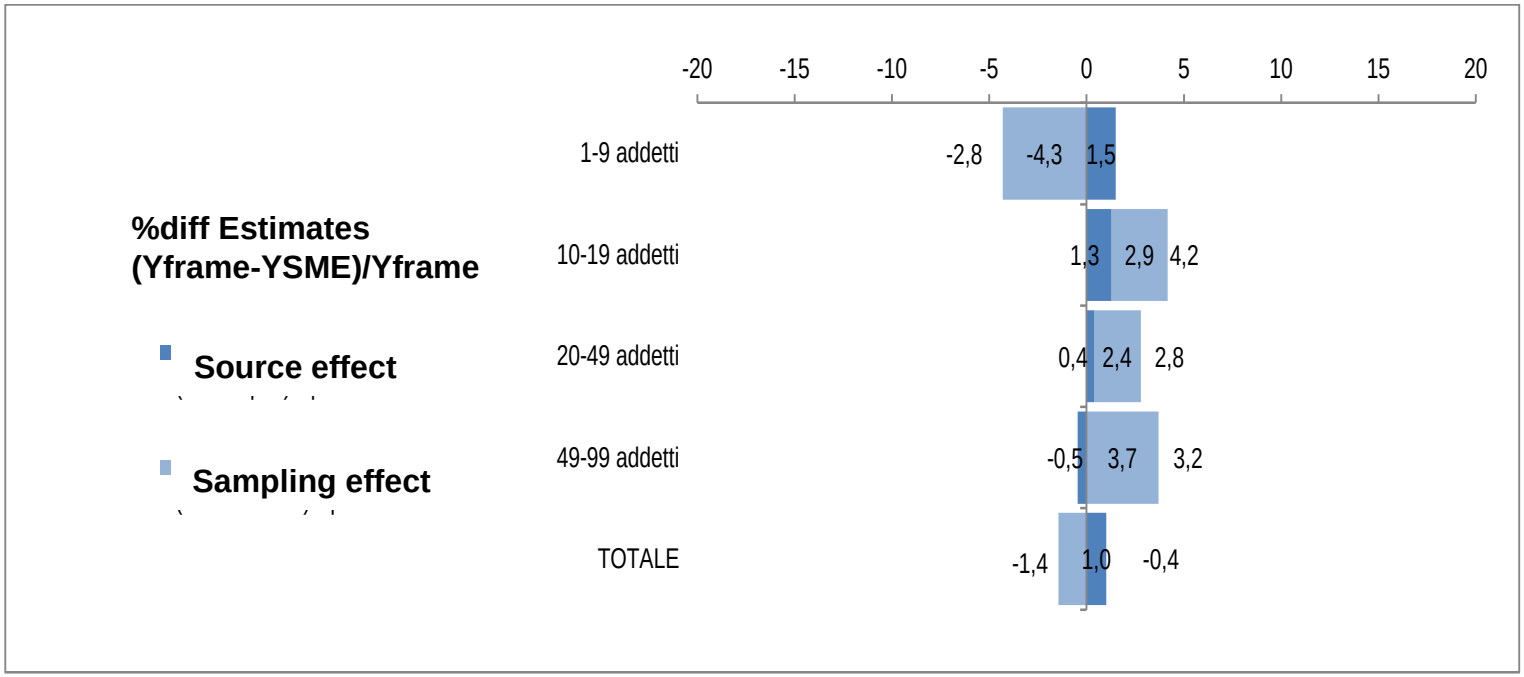
Assessing accuracy at micro-data and distribution level

Example: quality Indicators for the main SBS variables (survey vs SS)

Main economic variables	KS*	± 5% (%units)	± 5% (%value)	%diff.	value diff. (000€)	median diff. (000€)	IQR diff. (000€)	CV diff.
Current earnings excl. VAT, inclusive of indirect taxes	1.0	90.6	94.0	0.4	1.5	0.0	0.0	89.6
Increase in fixed assets for internal works	0.4	99.3	47.0	25.5	0.1	0.0	0.0	339.4
Change in work in progress	0.5	96.6	522.9	-936.1	0.9	0.0	0.0	88.2
Changes in inventories (finished goods, raw materials and goods for resale)	2.1	82.4	31.1	-86.8	-2.5	0.0	0.0	-38.2
Other revenues and income (non-financial, non-overtime)	9.6	61.5	25.1	-19.5	-1.5	0.0	0.0	-40.1
Purchases of goods (a)	10.1	60.7	76.8	-2.4	-5.1	0.0	2.8	-26.4
Purchases of services (b)	5.7	23.0	26.6	10.8	6.5	0.2	7.8	17.8
Purchases of goods and services (a+b)	1.5	52.4	76.9	0.5	1.4	-0.1	5.3	88.0
Tenure Leasehold	4.7	80.5	68.1	2.2	0.3	0.0	0.0	48.2
Other operating expenses	15.1	13.1	7.2	-11.1	-1.2	0.0	3.9	-26.8
Cost of labor	4.8	85.1	81.3	1.6	1.0	0.0	0.0	26.5
Depreciation and amortization	3.6	67.8	60.3	-8.3	-1.2	0.0	0.0	-21.9
Value Added	1.1	52.1	48.9	-1.6	-1.9	0.2	5.0	-46.4
EBITDA	1.4	45.6	38.2	-4.7	-2.9	0.1	4.5	-29.7
Net Operating Margin	2.1	42.8	36.4	-3.1	-1.5	0.0	5.1	-61.1

Assessing accuracy at estimation level

Example: **Value added**: source and sampling effects



Assessing accuracy: selective editing

Selective editing allows for the identification of influential units at a given domain level based on pre-defined influence criteria (*score functions*)

- **Identification of possible lacks**

  Domains characterized by the highest amount of influential data

  - *lacks in the harmonization process*
  - *Legal/administrative constraints*
  - *enterprises' behavior related to the specific source objectives*

  →*systematic discrepancies in sources' information*

- **Other sources of discrepancies**
  - *Measurement errors*
  - *Data variability*

  →*random discrepancies in sources' information*

Assessing accuracy: identify “anomalies” in administrative data

Absolute and % number of units with influential discrepancies between SS and SME data which need correction, by source. *k=0.05*

	Variable			
Source		Revenues	Purchases	Value Added
FS	n. units	4	42	73
	%units	0.02	0.025	0.44
SS	n. units	17	**1,002**	290
	%units	0.06	4.9	1.07

Current activities

- Improve the selective editing approach
- Improve *multivariate robust data modeling* for
  - Imputation of non covered units
  - *Modeling* of either non covered or not consistent source's variables
- Further analyses of data and economic indicators to improve accuracy for specific SME sub-populations
- Improving relationships with data providers for ensuring sources punctuality and quality
- Process documentation for future standardization

Expected benefits

- Reduction of statistical burden and costs (medium term)
- Data available at an extremely refined level of detail
- Higher accuracy and coherence within and across statistical domains

Some drawbacks

- Stability over time – data can be changed for internal decision of the owner administration
- Initial costs for assessing sources' quality and for ensuring sources' usability