Building the Italian Integrated System of Statistical Registers: Methodological and Architectural Solutions

Giorgio Alleva , Piero Demetrio Falorsi**, Orietta Luzi**, Monica Scannapieco***

*** University of Rome La Sapienza**
**** Istat**

Outline

- The Integrated System of Statistical Registers (ISSR)
- ISSR Methodological design
- ISSR and the Population Census
- Accuracy of ISSR statistics
- ISSR Architectural design
- Conclusions and future work

The Italian Integrated System of Statistical Registers (ISSR)

On January 2016 *Istat approved the Modernisation Programme*, in accordance with ESS commitment to *Vision 2020*

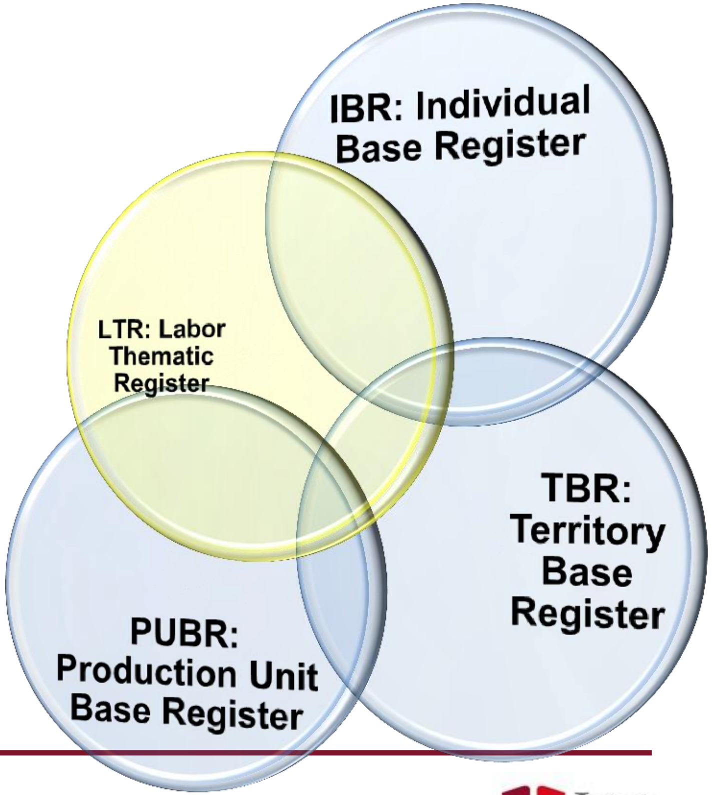
- Deep revision of the production processes for official statistics
- ISSR created by massive integration of administrative and survey micro-data - *Integrated System of Statistical Registers (ISSR)*

Base Statistical Registers (BSR)

- statistical units belonging to populations relevant for official statistics
- "core" variables from admin sources, *highly identifiable* and *stable in time*

Satellite Registers (SR)

- "thematic" variables from admin sources or direct surveys



TBR: Territory Base Register
IBR: Individual Base Register
PUBR: Production Unit Base Register
LTR: Labor Thematic Register

Single logical environment to support the consistency of statistical production processes

The Italian Integrated System of Statistical Registers (ISSR)

Main benefits

- Consistency between statistics in the ISSR and the statistical system of surveys
- Increase of the of available information on populations and variables as compared to each source individually taken
- Reduction of costs and statistical burden

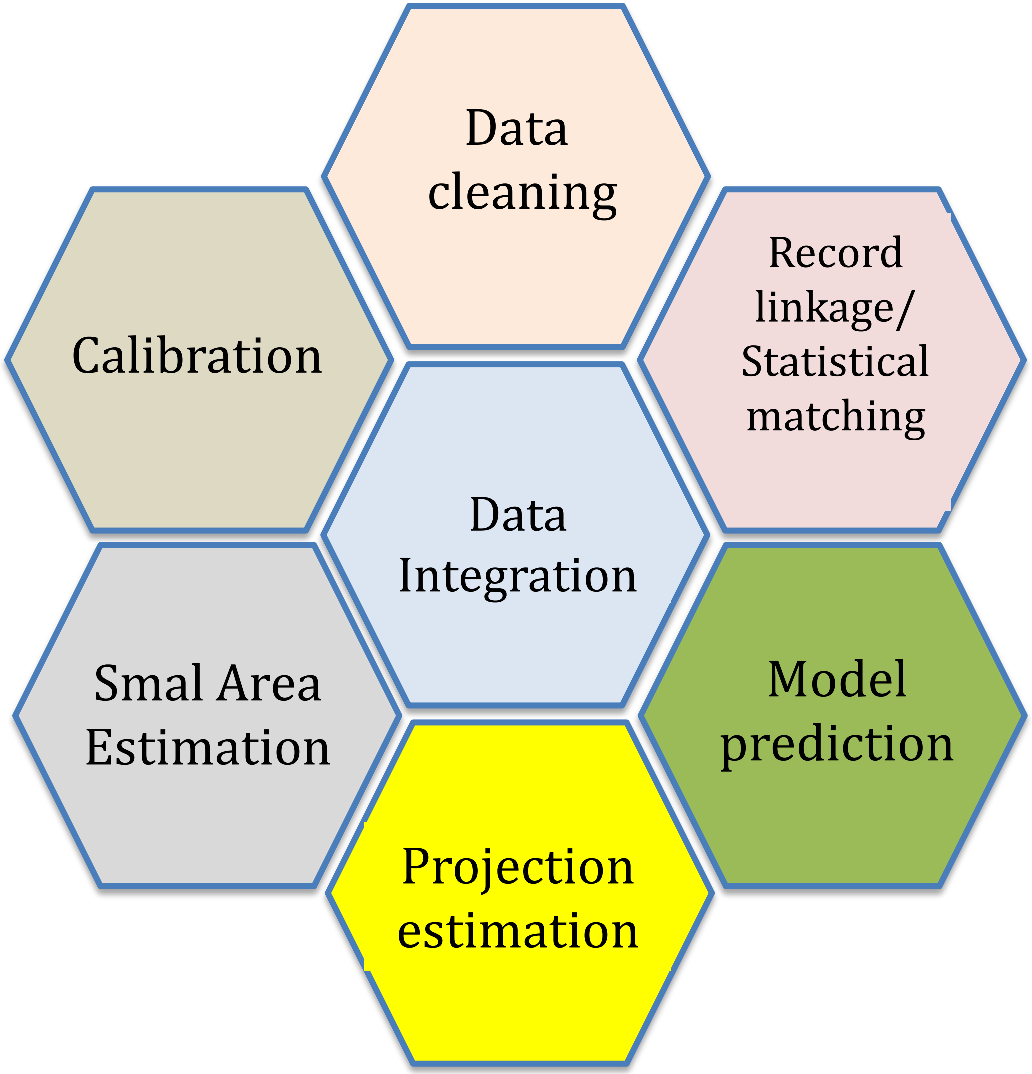
Methodological design of the ISSR

Microdata level

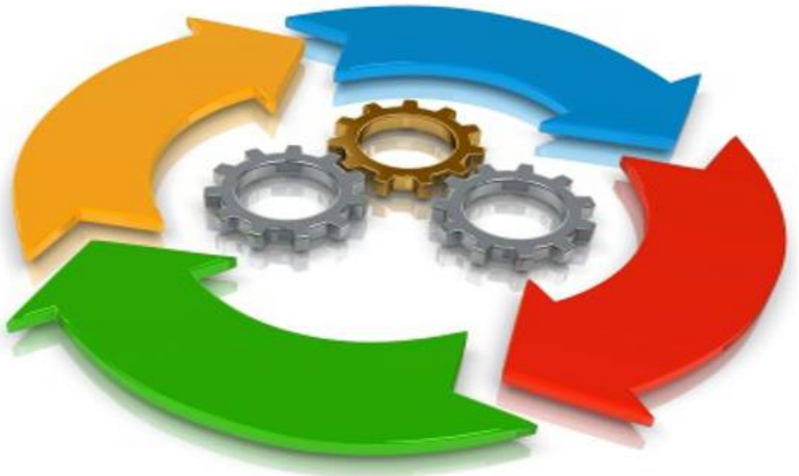
- Elegibilty of units
- Statistical integration of administrative and survey data (Record linkage/ Statistical matching)
- Model predictions for single units using both admin and survey data (e.g. Projection estimators)

Macrodata level

- Calibrated estimates, Small Area Estimation, Bayesian Models



Integration between surveys and the ISSR



Surveys supporting ISSR

- Estimation of varibales not in the ISSR
- Model-based prediction of missing information
- Correction of measurement errors
- …

ISSR supporting surveys

- Direct estimation
- Sampling Frames
- Auxiliary variables for Calibration
- Auxiliary variables for imputation
- …

Towards Register-Driven Production: the new Italian Population Census

- No longer a complete enumeration survey, but the result from the integration of administrative and survey data:
  - "Permanent" Census: the Census surveys (i.e. the Master Sample) are carried out every year
  - Each Census round is register-based borrowing strength from the ISSR
  - The ISSR is integrated by information from Census surveys (Master Sample)
  - A pivotal role played by the Individual Base Register (IBR)

Towards Register-Driven Production: the Population Census and the Individual Base Register

Every year:

For units and variables:

- Integration of IBR with the Master Sample
  - Pop. counts adjusted for coverage errors affecting IBR
  - Collect information not included in the IBR

For units:

- Adjust IBR estimates of population counts (*stocks*) based on other admin archives
  - Births, Deaths, Migrations *(flows)*

Adjust population counts for coverage errors in the IBR

Corrected estimates of population counts arise from an Extended Dual System Estimation based on the linkage between the IBR (*first capture*) and an Area-sampling survey - Population coverage survey - (*second capture*)

Population coverage Survey (PCS)

- Based on an area-sampling drawn from the Territory Base Register (TBR)
- 2.850 Municipalities each year (1.143 AR) covering all the Italian Municipalities in 4 years
- About 400 thousands households each year

Number of usual residents in a given municipality L:

$$\widehat{N}_L = \sum_g N_{g,R} \frac{\hat{P}_{g,L|R}}{\hat{P}_{g,R|L}} = \sum_g \sum_{k \in g} d_k \text{ (1)}$$

- g : post-stratification cell
- $N_{g,R}$: register total of the number of people in the post-stratum g
- $\hat{P}_{g,L|R}$: PCS estimate of proportion of usual residents among those registered, being $(1 - \hat{P}_{g,L|R})$ the estimated over-coverage proportion in the register
- $\hat{P}_{g,R|L}$: PCS estimated proportion of people registered among those usual residents, being $(1 - \hat{P}_{g,R|L})$ the estimated under-coverage proportion in the register

- $d_k = \hat{P}_{g,L|R} / \hat{P}_{g,R|L}$ for $k \in g$: individual weight measuring the misalignment between the place of residence and the place of *usual* residence

Consistency of Population Stock and Flow Estimates

- Estimates of population counts (stocks) should be consistent with information about demographic events (flows) available from civil registries (independent estimates)
- The Demographic Balancing Equation (DBE) to be fulfilled

$$\hat{N}_L^{(t+1)} = \hat{N}_L^{(t)} + (B - D) + (I - E)$$

$$\begin{cases} B - D & \textit{Natural Increase} = \textit{Births} - \textit{Deaths} \\ I - E & \textit{Net Migration} = \textit{Immigrants} - \textit{Emigrants} \end{cases}$$

- Constrained optimization problem
- Stone-Byron balancing method – commonly adopted for balancing large systems of National Accounts
- Time and space consistency of estimates of population counts and demographic figures ensured

Accuracy of ISSR statistics: Data structure in register R

As the register values are the output of statistical processes, they are subject to statistical uncertainty in terms of both units and variables

Erroneous Inclusions

Register

Unit code in R	True **but UNKNOWN Value**	**PREDICTED Value**	**Auxiliary Variable With no uncertainty**	**Domain membership (0,1) variable**
1	y_1	$\hat{y}_1$	$\boldsymbol{x}_1$	1
⋮	⋮	⋮	⋮	⋮
k	y_k	$\hat{y}_k$	$\boldsymbol{x}_k$	0
⋮	⋮	⋮	⋮	⋮
N	y_N	$\hat{y}_N$	$\boldsymbol{x}_N$	1

External estimates

$\widehat{N}_{(U)}$ Population Size

$\widehat{N}_{R,E}$ N. of Erroneous inclusions in R

$\hat{y}_k =$
- «Observed» value y_k (for a subset)
- Value built by an **explicit** or **implicit statistical model** or **algorithm**

Istat

Accuracy of ISSR statistics

Let the **target parameter** be the **total** of the variable y in the domain U_d

$$Y_{U_d} = \sum_{k \in U_d} y_k$$

Let R be a statistical register built at micro-level for the target population U

Let R_d be a subset *of* R which should represent the target domain U_d

Let $\hat{y}_k$ be the value in the register that predicts the value y_k of the unit k.

The proposed measure of accuracy

Statistical estimate of Y_{U_d} over R_d

$$\hat{Y}_{R_d} = \sum_{k \in R_d} \hat{y}_k \quad (\textit{not a true value})$$

Anticipated Variance (Isaki and Fuller, 1982; Sarndäl *et al*., 1992; Nedyalkova and Tillé, 2008; Nirel, and Glickman, 2009; Falorsi and Righi, 2015) useful for **official statistics**:

$$AV\left(\hat{Y}_{R_d}\right) = E_P E_M \left(\hat{Y}_{R_d} - Y_{U_d}\right)^2$$

The AV neutralizes variability due to the pure model variability $V_M\left(Y_{U_d}\right)$ of the **finite population value** Y_{U_d}.

It allows us to easily take into account the different sources of variability resulting from various approaches to inference

* E_P , V_P = Expectation and variance with respect to sampling
* E_M , V_M = Expectation and variance with respect to statistical models used for the prediction

ISSR Architectural design

Requirements:

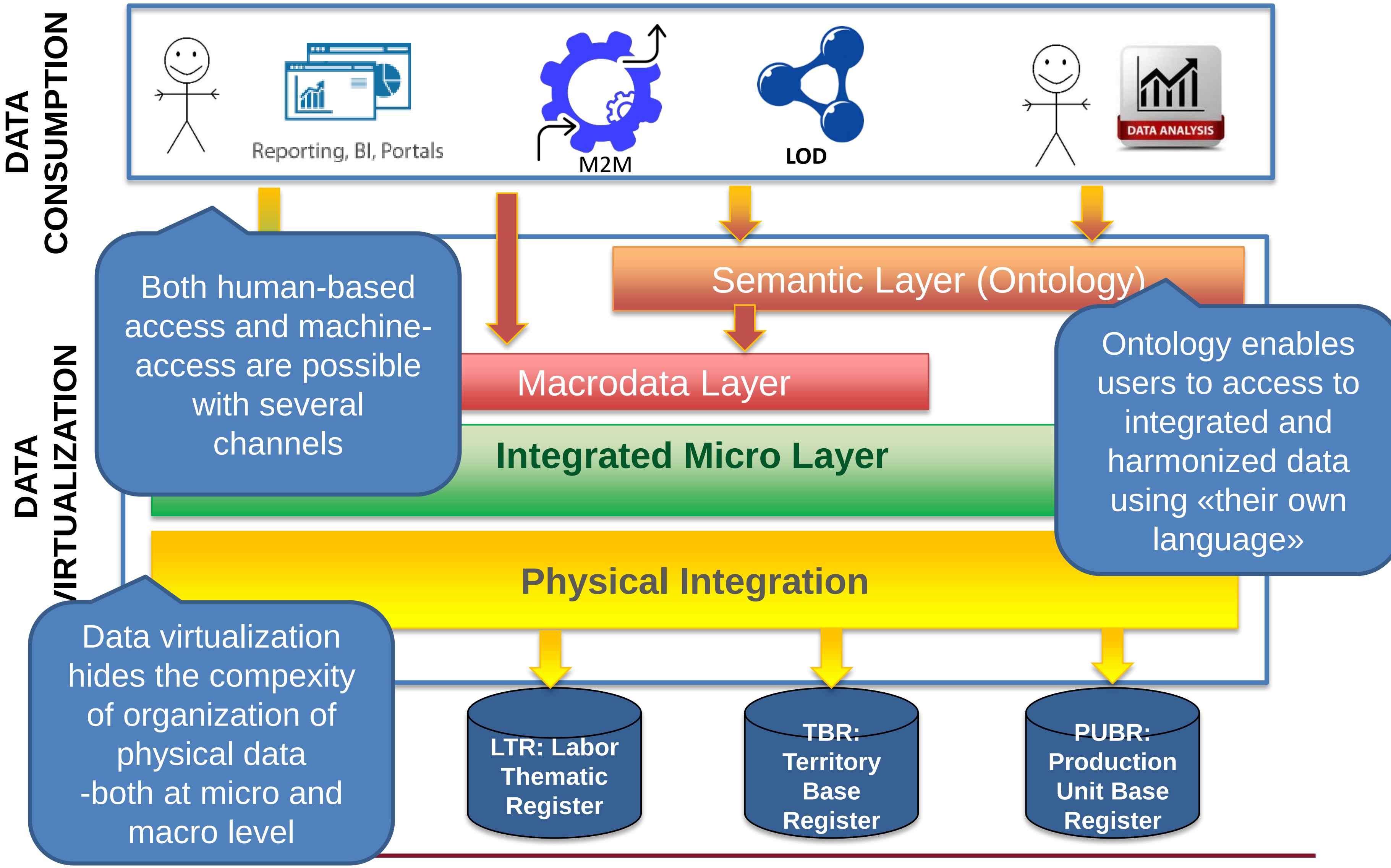
- Need to integrate concepts belonging to different thematic areas

Adoption of the Ontology-Based Data Management (OBDM) approach for accessing, integrating and managing heterogenous data sources

Main advantages:

- Industrialization of processes
- Semantic data integration
- Data quality control
- Flexibility in adding new sources or modifying existing ones

ISSR Architectural Design



Istat

Conclusions and Future work

- The ISSR strongly supports the consistency of the statistical production through the hamonization and integration of administrative sources and direct surveys

- Methodological and architectural design have been developed according to standards and with quality as first concept

- Next steps:
  - Continue the ongoing development efforts
  - Build on the current experience in a quality improvement feedback loop

References

Alleva G., (2017). *The new role of sample surveys in official statistics*, ITACOSM 2017, The 5th Italian Conference on Survey Methodology,14 giugno 2017, Bologna, https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf

Byron, R. (1978), The Estimation of Large Social Account Matrices, Journal of the Royal Statistical Society A, vol. 141(3), pp. 359-367.

Di Zio M, Fortini M., Zardetto D., (2018), Achieving Consistency between Estimates of Demographic Stocks and Flows through Balancing Methods, Itacosm, https://events.unibo.it/itacosm2017/abtracts-of-invited-papers

EARF Enterprise Architecture Reference Framework (2015), Available at: https://ec.europa.eu/eurostat/cros/content/ess-ea-rf_en

Falorsi, P.D., Righi P. (2015). Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey Methodology*, 41, 215-236.

Isaki C.T., Fuller W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89–96.

Istat, (2016), Istat's Modernisation Programme, https://intranet.istat.it/News/Modernizzazione/Istat%27sModernistionProgramme_EN.pdf

Nedyalkova, D. , Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 521-537.

Nirel, R. , Glickman, H. (2009). Chapter 21 - Sample Surveys and Censuses. In: Rao, C.R. (ed.) *Handbook of Statistic*, Elsevier.

Pfeffermann, D., Eltinge, J. L., & Brown, L. D. (2015). Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture. Journal of Survey Statistics and Methodology, 3(4), 425-483.

A. Wallgreen; B.Wallgreen, (2014). Register Based Statistical methods for Administrative Data. Wiley,: New York.

Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81, 338 - 346.

Istat

Thank you for your attention