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 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**
- Eligibility 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)

**Macrodada level**
- Calibrated estimates, Small Area Estimation, Bayesian Models
Ensuring consistency between ISSR and Survey Production system

Integration between surveys and the ISSR

ISSR supporting surveys
- Direct estimation
- Sampling Frames
- Auxiliary variables for Calibration
- Auxiliary variables for imputation
- ...

Surveys supporting ISSR
- Estimation of variables not in the ISSR
- Model-based prediction of missing information
- Correction of measurement errors
- ...
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 \( L \):

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

- \( g \): 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
- \( d_k = \frac{\hat{P}_{g,L|R}}{\hat{P}_{g,R|L}} \) 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:

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

\[ \begin{align*}
B - D & \quad \text{Natural Increase} = \text{Births} - \text{Deaths} \\
I - E & \quad \text{Net Migration} = \text{Immigrants} - \text{Emigrants}
\end{align*} \]

- 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**.

<table>
<thead>
<tr>
<th>Unit code in $R$</th>
<th>True but <strong>UNKNOWN</strong> Value</th>
<th><strong>PREDICTED</strong> Value</th>
<th>Auxiliary Variable With no uncertainty</th>
<th>Domain membership (0,1) variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_1$</td>
<td>$\hat{y}_1$</td>
<td>$x_1$</td>
<td>1</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$k$</td>
<td>$y_k$</td>
<td>$\hat{y}_k$</td>
<td>$x_k$</td>
<td>0</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$N$</td>
<td>$y_N$</td>
<td>$\hat{y}_N$</td>
<td>$x_N$</td>
<td>1</td>
</tr>
</tbody>
</table>

**External estimates**
- $\hat{N}(U)$ Population Size
- $\hat{N}_R,E$ N. of Erroneous inclusions in $R$

\[
\hat{y}_k = \text{«Observed» value } y_k \text{ (for a subset)}
\]

Value built by an **explicit** or **implicit statistical model** or **algorithm**

---

*Illustration of the data structure in register $R$*
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 $R_d$

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

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

$$AV(\hat{Y}_{R_d}) = EP EM (\hat{Y}_{R_d} - Y_{U_d})^2$$

The $AV$ neutralizes variability due to the pure model variability $V_M(Y_{U_d})$ 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

* $EP, V_P = \text{Expectation and variance with respect to sampling}$
* $EM, V_M = \text{Expectation and variance with respect to statistical models used for the prediction}$
ISSR Architectural design

Requirements:
• Need to integrate concepts belonging to different thematic areas

Main advantages:
• Industrialization of processes
• Semantic data integration
• Data quality control
• Flexibility in adding new sources or modifying existing ones

Adoption of the Ontology-Based Data Management (OBDM) approach for accessing, integrating and managing heterogeneous data sources
Ontology enables users to access to integrated and harmonized data using «their own language».

Both human-based access and machine-access are possible with several channels.

Data virtualization hides the complexity of organization of physical data -both at micro and macro level.
Conclusions and Future work

• The ISSR strongly supports the consistency of the statistical production through the harmonization 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


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


Thank you for your attention