

Overview of the Istat activities and open problems

Marco Di Zio, Stefano Falorsi, Silvia Loriga

5 December 2022

- 1 The new Istat data production process
- 2 Design and implementation of quality by design
- 3 Choice of primary and secondary sources
- 4 Assessing uncertainty of register based statistics
- 5 Accuracy and coherence for the CSSIS
- 6 Other main lines of research for the next few years

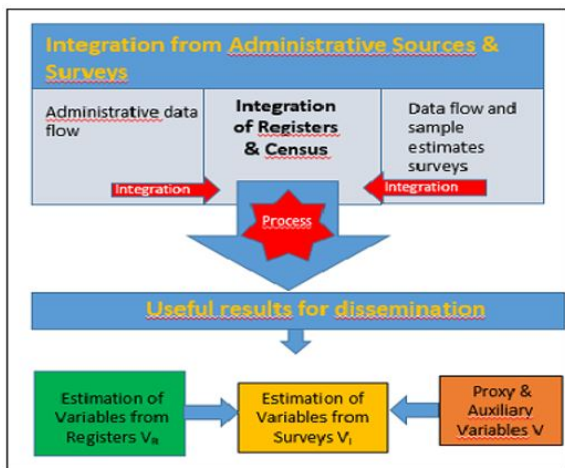
The new Istat data production process

The new production strategy, started up in 2016 following a deep methodological review of the statistical processes, and is primarily based on a multi-source approach:

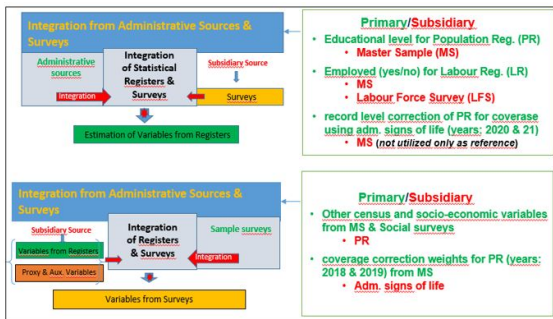
- The Integrated System of Registers (ISR)
- The new Population Census System (PCS), based on the integration of Registers with Census Master sample surveys
- The Census and Social Surveys Integrated System (CSSIS)

The goal of CSSIS is to obtain **more accurate and more coherent** estimates, exploiting the different available sources, in an integrated in an efficient manner

The new Istat data production process 1



The new Istat data production process 2



Design and implementation of quality by design

A **quality by design** framework consists in a holistic approach in which **quality** represents the goal to be pursued in **all the phases** of the statistical process, from the design to the output

Single-source approach: the accuracy is the result of choices (sample design, estimation method...) and events (measurement errors, non-responses, under-coverage...) referred to the single statistical process.

Being the different sources/processes independent of each other, coherence issues may arise

Multi-source approach: the accuracy is taken into account since the design of the processes and coherence issues may be faced in a coordinated way. The quality should be assessed taking into account the quality of the different sources.

Design and implementation of quality by design


Two issues will be treated in the following:

- The choice of primary and secondary sources and the evaluation of the uncertainty of register based statistics within ISR;
- The improvement of accuracy and coherence for the ICSSS exploiting unit level auxiliary information coming from both Registers and Census sample surveys based on a master sample design

CHOICE OF PRIMARY AND SECONDARY SOURCES

Choice of primary and secondary sources 1

- An important question is whether admin data values can be directly used for computing statistics or a random sample is needed.
- Two main issues to consider: How much the variable in the register represent the target variable, and to which extent the observed part of the data is representative of the population.
- Useful ideas are in Meng (2018)¹. He deals with the question: “Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?”. In fact, this is one of the questions of a survey manager to a statistician when asking for the use of administrative data.

¹ Meng X-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential elections. *The Annals of Applied Statistics*, Vol. 12, No. 2, 685–726. 

Choice of primary and secondary sources 2

In case of estimate of the mean value of a finite population \bar{G}_N with the mean computed on subset of observed data

$\bar{G}_n = \frac{1}{n} \sum_{j=1}^n G_j = \frac{\sum_{i=1}^N R_i G_i}{\sum_{i=1}^N R_i}$, $R_i = 1$ if the i -th is observed (0 otherwise). R represents the selection/observation mechanism

$$MSE(G_n) = E_R[\rho_{(R,G)}^2] \times ((1-f)/f) \times \sigma_G^2 = D_I \times D_O \times D_U$$

Error depends on three elements: 'data quantity' $(1-f)/f$, 'problem difficulty' σ_G^2 , and 'data quality' caught from the correlation $\rho_{(R,G)}$ (not measurable from the sample).

Choice of primary and secondary sources 3

The question is how much we gain or lose with respect to a simple random sample taken as benchmark. We can compute

$$Deff = \frac{E_R[\bar{G}_n - \bar{G}_N]^2}{V_{SRS}(\bar{G}_n)} = (N - 1)D_I$$

To ensure MSE of the sample mean can enjoy of the usual converge rate of the n^{-1} order, D_I should be controlled with a rate of N^{-1} or equivalently the 'data defect correlation' $\rho_{(R,G)}$ with a rate of $N^{-1/2}$.

For a population of 60,000,000 of individuals (approximately the Italian residents), $\rho_{(R,G)} \approx 0.0001$ that is a very low value.

Choice of primary and secondary sources 4

On the other hand, we may inspect the damage of $\rho_{(R,G)}$ even in case of a small value. The effective sample size n_{eff} of the subset of non-probabilistic data to have the same MSE of a SRS:

$$n_{eff} \leq \frac{n}{1-f} \frac{1}{ND_I}$$

ND_I increases rapidly with N causing an important reduction of n_{eff} .

For instance, with $E_R[\rho_{(R,G)}] = 0.05$, $n_{eff} \leq 400 \frac{f}{(1-f)}$, then for a subset of data with $f = 1/2$ (50% of the population) to have a behaviour similar to a SRS the effective sample size cannot be greater than 400 units, and with a population of 60 million units, there is a reduction of the sample size from 30 million to 400, that is a decrease of 99.9999%, or equivalently a loss of efficiency.

Choice of primary and secondary sources 5

- Hence, even if D_I for a non-probabilistic sample is small, the effect is magnified by N .
- It is necessary to think about those issues because we notice that if we do not control the R-mechanism, the fact that n (units observed in the register) is big can be even counterproductive, because we put more trust on biased data:

“The bigger the data, the surer we fool ourselves”

ASSESSING UNCERTAINTY OF REGISTER BASED STATISTICS

Assessing uncertainty of register based statistics 1

- Assessing uncertainty of statistics with multi-source data is particularly relevant and even more important when the output of inference is a statistical register, i.e., a set of data estimated at unit level.
- With statistical registers, the possibilities for analysis increase, so agile tools to assess their quality must be provided to the users.
- While in sample surveys the classical randomization or design-based approach is generally adopted, in case of multi-source inference accuracy evaluation needs to be developed in order to deal with non-probabilistic and probabilistic data sources, often jointly used.

Assessing uncertainty of register based statistics 2

Some methodological issues have been recently discussed in literature;

- Alleva et al. consider model and sampling errors jointly, and address the issue of providing statistical information about the quality reported for each unit such that an appropriate combination of it allows the evaluation of accuracy of linear estimators.
- Scholtus et al. 2021 study resampling techniques for accuracy evaluation in a similar context.
- Boeschoten et al. (2021) propose multiple imputation for data integration model based on hidden Markov models.
- Other papers related to this topic can be found among the works on mass imputation, see for instance Chen et al. (2022), Kim et al., (2021), Yang et al., (2021).

ACCURACY AND COHERENCE FOR THE CSSIS

Accuracy

The **use of the 1st phase information in the 2nd phase**, (sampling design and estimation), helps to improve the accuracy.

In this framework **the gain in efficiency** should be evaluated comparing different **sampling designs**, in which information from the 1st phase may be used, for instance in a **balanced sampling**.

As for the **estimation process**, information from the 1st phase may be used in a **design based approach**, (calibration) or even in a **model assisted approach**, such as SAE.

Accuracy

Alternatively **register information** may be used with the same goal. Alternative approaches exploiting surveys or register information should be evaluated considering **nature and quality of the different sources**

Usually information from **surveys** are more **timely** and the **target variables** may be **harmonized** between different surveys; but **measurement errors** may affect the variables, **estimates** are affected by **sampling errors** and may be **biased** due to non-response

On the other hand **register information** are usually available with a certain **delay** and the definitions of the variables follow **administrative criteria** and are not aligned to the statistical definitions; they are not affected by sampling errors, but there could be **coverage issues**.

Accuracy

The study on more efficient sampling designs and estimation processes should consider even the **output strategies** of the two surveys: if the same variable is collected by both the surveys, **which estimates are produced by the 1st phase survey, which by the 2nd one?** Which is the **periodicity** and the **detail** of the respective estimates? And then which estimates are produced by **direct estimates** and which by **model estimates**?

These topics are currently under study for the **Aspects of Daily Life**, which is already designed as a 2nd phase survey. **Labour force survey and Eu-Silc** are currently only partially integrated, simulation studies are being conducted to evaluate the opportunity to complete a full integration with Population Census Master sample

Coherence

As it was mentioned earlier a production and output dissemination strategy should be established if the 1st and 2nd phase surveys collect the same target variable

In this case **coherence issues** may arise and should be faced in an multi-source approach in a coordinated way, designing a priori the use of the available sources in the different phases of the statistical processes.

Generally speaking we may dispose of:

- register information, affected by coverage issues and containing administrative variables
- a big survey conducted at the 1st phase, collecting the target variable
- a small survey at the 2n phase, focused on a specific subject and collecting more precisely the target variable

Coherence

A multi-source approach should exploit the three kind of sources in a coordinated way

Register information could be used in both the surveys improving the accuracy of the estimates (both in design based or model based estimation processes)

The integration of the samples of the two surveys should allow the study of **measurement errors**

Such strategies, in addition to improving accuracy, also improve the **consistency** of the whole process and the **coherence** of final outputs

This is the case for instance of the **employment estimates**, for which we have:

- **Labour register** built integrating administrative data on regular employment
- **Population Census Master sample**, collecting employment according to ILO definition
- **Labour force survey**, collecting employment according to ILO definition as well, through more detailed variables

In this multi-source framework, the methodological approaches to improve the coherence of the employment estimates produced by Labour force survey, Population Census (and even National Accounts) are currently the subject of extensive studies

CONCLUSIONS: MAIN LINES OF RESEARCH FOR THE NEXT FEW YEARS

Other lines of research for the next few years 1

The main research activities and challenges for the near future are:

- Completing the methodological framework and implementation of the existing and new registers;
- Defining the final asset of the CSSIS with reference to ADL, LFS and EUSILC;
- Study for the population census master sample of adaptive survey designs using a CAWI contact scheme with CAPI sub-sample of CAWI non-respondents.

Other lines of research for the next few years 2

- To improve (out of sample domains) SAE methods for producing statistics at the level of territorial and structural detail exploiting spatial correlations and correlations among different variables ;
- Joint exploitation, even over time, of data from different sources to produce new information with greater timeliness and/or strong granularity:
 - e.g. for Register/Census estimates for 2023 year using administrative and sampling information available for 2023 and leveraging information coming from previous years 2018-2022.

MANY THANKS FOR YOUR ATTENTION!