The Italian Esperience in Population Census in The Year of Covid

Update

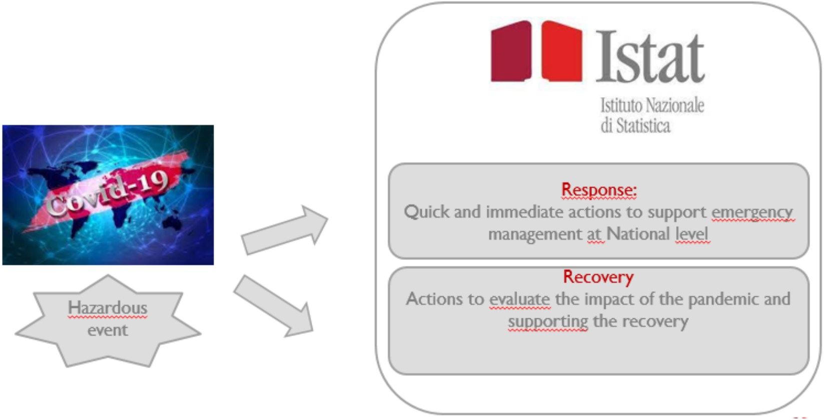
Marco Di Zio, Stefano Falorsi, Romina Filippini, Danila Filipponi, Silvia Loriga, Gaia Rocchetti

October 8, 2021

Overview

1. COVID-19 and general strategies to guarantee the production of statistics
2. The Census production process in 2018-2019
3. The Census output in 2020
4. Situation 1 Attained Level of Education
   Estimation: unit-level 2020 adm.data + *model-prediction*,
   Evaluation: total-level 2020 adm.data from Min. of Educ.
5. Case 2 Population Counts
   Estimation: unit-level 2020 adm.data + *determ. correction*,
   Evaluation: non-response LFS 2020 data
6. Final Considerations

COVID-19 and the reaction of Istat

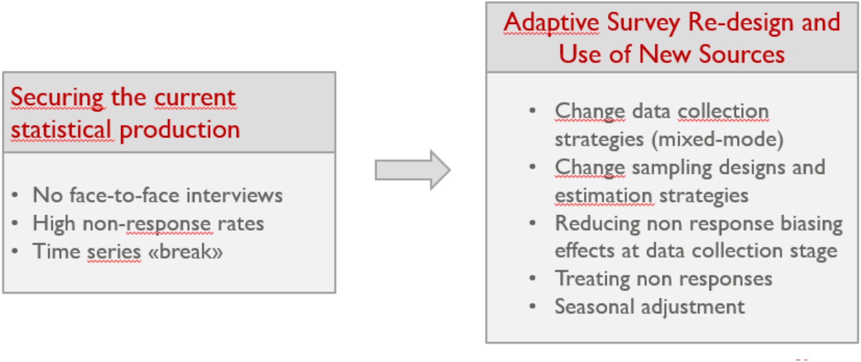


Istat
Istituto Nazionale di Statistica

Response:
Quick and immediate actions to support emergency management at National level

Recovery
Actions to evaluate the impact of the pandemic and supporting the recovery

Response Phase: Needs and Actions



Securing the current statistical production

- No face-to-face interviews
- High non-response rates
- Time series «break»

Adaptive Survey Re-design and Use of New Sources

- Change data collection strategies (mixed-mode)
- Change sampling designs and estimation strategies
- Reducing non response biasing effects at data collection stage
- Treating non responses
- Seasonal adjustment

The Census production process in 2018-2019

The first two annual (2018-2019) waves of the Permanent Census, were based on two pillars:

- Registers that organize, process, correct and complete, by means of appropriate statistical models, the target *available information* (i.e. r-variables) from adm. sources;
- Surveys updating information on the r-variables and adding the target *missing information* (i.e. nr-variables) through their direct observation on respondents .
- E.g. ALE and employment status are r-variables, while commuting and other modalities of employment status are nr-variables.

The Census surveys in 2018-2019

Two sampling surveys *List* and *Area* are carried out yearly collecting almost the same variables by means of different sampling schemes. *Area* survey has the additional objective of providing information for correcting the population counts by *under-coverage* on the basis of a dual frame estimator. *List* and *Area* surveys provide an estimation of *over-coverage* rates of the PR.

- *List* survey is selected by PR. Each year, about 950,000 households are planned to be sampled
- *Area* survey is selected by *AR* (Address Register) and is based on a door to door enumeration. Each year the expected number of households is about 450,000 households.
- Out of about 7,900 municipalities, about 2,850 are surveyed every year.

The Census output in 2020

In 2020, due to the covid-19 epidemic, Census sample surveys were not conducted and it was decided to publish a Reduced set of tables (for which administrative data were available) by Municipal level x Demographic profiles (*Gender* x *Age Classes* x *Citizenship* *Italian* / *Foreign*) from PR (Pop. Register):

1. ALE - Attained Level of Education
   - *2020 predicted ALE*
   - *2019 fitted model* using 2019 admin. and surv.data
2. PCO - Population COunts
   - *2020 deterministic correction of PR*
   - using 2020 admin. "life signals", available for each individual - adding and deleting records from PR - in order to correct PR for over/under-coverage

Output validation criteria

The quality of 2020 census output was evaluated through 2018-2019 census survey data + 2020 reference data:

1. the 2019 estimates released by the census were compared with the corresponding estimates obtained using the 2020 production process applied to 2019 data + the 2020 estimates were evaluated using external sources
2. 2020 ALE estimates were assessed through the administrative totals referring to 2020 provided by the Ministry of Education
3. over-coverage rates underlying 2020 PCO estimates have been compared too, with those calculated from time series of non-response indicators from LFS starting from 2020 quarters till 2014.

Situation 1 - Estimation of ALE

The estimation of ALE is based on modelling the probability of attaining a certain level of education given a profile (e.g. gender by age) based on Multisource data:

- Administrative data concerning 2011-2018 (Attending courses 2018/19). Population of active students is included here.
- 2011 Census data covering years before 2011.
- Survey data for courses not in administrative data and for people entered Italy between 2011-2020.

Remark. Probabilities of attaining a higher LE are stable from time to time. For active students, they are estimated on admin data referring to years 2016-2018 and applied to 2018 to obtain the 2020 ALE.

Covid 19 consequences on estimation of 2020 ALE

- Sample survey is not carried out.
- The stability assumption of probabilities is not necessarily maintained given the shock of Covid on education.

What can we do in practice?

- Using the 2019 sample survey.
- Using admin Data 2011-2018 by assuming the stability of probabilities.

Measuring the impact of Covid 19 on quality

- The impact is on active students attending the last year of a course, representing a small part of the population.
- To evaluate the impact on estimates, Istat asked the Ministry for any kind of available data that provided timely but provisional information, that is different from the information used in the model, in terms of: (a) population; (b) time reference; (c) geographical detail.

Information is on

- Percentage of students admitted to final examination that obtained a qualification for middle school and high school at time t (2020) - Covid era - by Region
- School dropout rate by Region

We need to transform this information to fit - as much as possible - our modeling.

Evaluation

- We compared estimates obtained by using 2019 survey data and admin data with the stability assumption (practical solution proposed) vs estimates obtained by updating data with the transformed information.
- Negligible differences for middle school (0.49%) and high school (0,35%).
- To evaluate the impact on the distribution of ALE on the total population (publication figures), we remind the main impact is only on active students attending the last year of a course.
- For resident population the impact is small, and it is about an underestimation of 0.013% at national level.

Situation 2 - Estimation of PCO

- for the coverage 2018-2019 estimation and the production of population counts at the municipality level we have studied the possibility to integrate non response LFS outcomes with outcomes of Census surveys jointly with "life signals" admin. data
- for 2020 estimation purposes we explored, too, the possibility to use LFS outcomes to produce an over-coverage estimation using only "life signals" admin. data.

Main features of the Italian LFS

The LFS is the largest social survey conducted by Istat, apart from the Census surveys.

The LFS sampling design is similar to the L survey:

- 2-stages: PSU municipalities, FSU households
- stratification of PSU
- a very large number of municipalities overlap with L
- the sample of households is selected from an administrative frame that is the main source of RBI (99.97%)

LFS features: continuous survey - time series of non-response indicators since 2004 - high quality profile - data collection rules clearly defined in contracts - professional interviewers network trained by Istat - standardized field indicators - monitoring.

LFS for the over-coverage estimation

LFS currently produces a set of standardized non-response indicators, some of which can be exploited in this context. Focusing on households of the first wave (by CAPI), we may distinguish non-contact reasons for the analysis of **household over-coverage**.

- Uncertain outcomes
- Outcomes due to the frame updating (moved to another municipality or abroad or death)
- Outcomes related to household over-coverage (vacant houses because holiday properties or second homes, or persons living in a collective household)

LFS non-contact outcomes

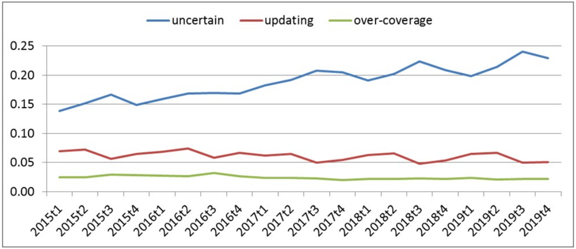


Figure: LFS non-contact outcomes (quarterly data years 2015-2019)

Outcomes related to over-coverage are **very stable** over time. Other outcomes show a **seasonal pattern** due to respectively higher non-contacts during summer and the time lag between the frame reference date and the data collection.

"life signals" admin data

admin.e data are used to complement LFS non-contacts outcomes.
By this way, over-coverage and updating outcomes may receive confirmation and the uncertain outcomes may be analysed to detect residual over-coverage:

- an **updated version of PR** is used to identify people who moved to another municipality or abroad or died
- **"life signals"** are used to identify people currently living in a different municipality (or abroad) than the one in which they reside in PR.

LFS for the 2020 over-coverage estimation

In 2020 LFS data collection was affected by Covid-19:

- in March data collection had a break: interviews could be conducted only by phone (only to households with available phone number, mainly from previous waves)
- in the 2nd quarter a new sample was selected composed only by households with available phone number.

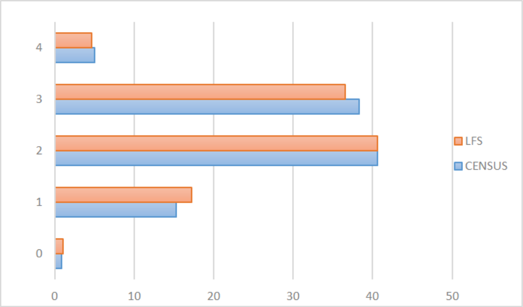
Even if 2020 LFS data collection was affected by Covid-19, due to the **stability of LFS outcomes** related to over-coverage, the past years results from the LFS on the over-coverage may be used, **complemented by updated administrative data** (updated PR and "life signals")

LFS for the 2020 over-coverage estimation

Table: Incidence rate of non-contact outcomes by the two surveys and by number of living signals

N° of living signals	CENSUS	LFS
4	5,29	22,33
3	6,08	23,73
2	8,74	28,23
1	16,28	34,00
0	30,64	47,60

Figure: Percentage of population coverd by the two surveys by number of living signals



LFS for the 2020 over-coverage estimation

- The living signals and the contacts of LFS are all indicators, measured with errors, of the coverage.
- We are evaluating the possibility of using a latent variable models to predict the over-coverage of RBI using the available indicators, together with additional covariates.

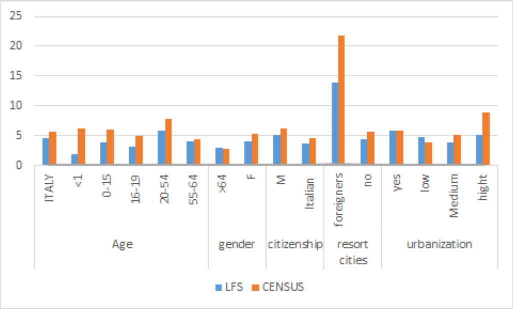


Figure: Estimated over-overage rate, by domains

Final considerations

- For both ALE and PCO problem is not only with the lack of sample data, but in this case even more with admin data. They have a lag and draw a picture before Covid.
- for ALE fortunately, the impact of Covid is essentially on a small subset of people (active students and potentially able to attain a higher level of education).
- An evaluation of estimates obtained by resorting to available data is carried out. However, they give approximated evaluation since data are partial and provisional.

Final considerations

- In more general terms, even if the two Census surveys were not conducted in 2020, the Permanent Census Process, based on the integration of Registers and Surveys, resulted rather "robust" to face a deep and sudden shock, such as Covid-19 pandemic, allowing the production of *ALE* and *PCO* estimates. The quality of such figures has been evaluated.
- In this context, strategies properly exploiting different (relevant) sources could represent valid solutions to improve robustness, even in case of unpredictable shocks. We're testing latent variable models to predict the over-coverage of PRI using the available sources (PC surveys and LFS), together with additional covariates (living signals)

Final considerations

- We're currently studying strategies based on the integration of Registers, Census surveys and also social surveys (for instance LFS)
- In this context the coordination of sampling design among Census and social surveys, introducing also an overlapping in terms of selected households, could be exploited in the estimation processes, enhancing robustness and coherence of the results