An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data

Filippini R.,
Di Zio M., Rocchetti G.,
Istat, Italian National Institute of Statistics

UN/ECE Workshop on Statistical Data Editing
(31 August - 4 September 2020)

Istat | Istituto Nazionale di Statistica

OUTLINE

- Background: The Italian Base Register of Individuals
- Informative context:
  - Data sources description
  - Coverage and sub-population segments
  - Some critical issues
- The imputation procedure
- Some results to evaluate the predictions
- Final remarks

BACKGROUND

The Italian Base Register of Individuals (BRI): a comprehensive statistical register storing data gathered from various data sources.

The subset of resident people is the basis of the next Italian census that will be as much as possible register-based.

- core variables like place and date of birth, gender, citizenship are associated to each unit.
- given the high amount of available administrative information, a prediction of the attained level of education (ALE) for the individuals in BRI is proposed.

INFORMATIVE CONTEXT: Data source for ALE (t=2018)

2018 sample survey (permanent Italian census).
People are asked about their educational level.

- Available at reference time t
  - ✓ 17 levels of classification.

5% of reference population

Administrative information, provided by the Ministry of Education, University and Research (MIUR) and processed in ISTAT (BIT): data on ALE and educational paths of enrolled students.

- Available from 2012 until t-1.
  - ✓ 16 levels of classification.

2011 Census: Data on ALE, of 2011 resident population.

- Available at reference time of October 2011 (t-7)
  - ✓ 12 levels of classification.

Auxiliary information from the registration and cancellation forms (APR): self-declared ALE.

- Available from 2012 until t-1.
  - ✓ 4 levels of classification.

INFORMATIVE CONTEXT: Tabular representation of available information

White = available inf.

Grey = missing data

X_{BRI}			X_{miur}				Sample	Prediction	Group
G	E	C^t	$I^{(t-1)}$	$F^{(t)}$	$L^{(t)}$	I^{apr}	I^t_S	I^t	
									A 22%
									B 73%
									C 5%

G: Gender; E: Age classes;

C^t: Citizenship at time *t* (It./not It.)

I^{t-12}: ALE at time *t-1*; L^t: type of school

F^t: year of attendance [*t-1*, *t*]

I^{apr}: auxiliary information on ALE

Group A:
ALE from MIUR

Group B:
ALE from 2011 Census

Group C:
auxiliary information from APR

INFORMATIVE CONTEXT: Critical issues

Homogeneity of classification -> 8 dissemination classes

Timeline of information: lag between the moment in which data are available and BRI reference

Informative gaps in administrative sources: does not include all qualification courses

necessary to implement different procedures for the prediction of ALE in time t in the 3 groups

Group A: young people attending a course with information on ALE in t-1	Group B: older people with information on ALE in t-7	Group C: middle age people mainly not Italian with no information on ALE

THE IMPUTATION PROCEDURE: Group A (22%)

The course attended during the school year [t-1/ t] is known.

The conditional probabilities of obtaining a new educational level is estimated by **using only administrative data.**

1. Deterministic rule for determining people that cannot attain a new level of education in one year (e.g., children attending the first four years of the primary school)
2. Estimation of conditional probabilities on previous year data - $h(I^{t-1} \mid X)$: predict ALE at t-1 by using data available in the interval time [t-2/ t-1], via hierarchical log-linear models:

   $[C^t, E^t, F^t, L^t, I^t]$
3. Imputation with random draw from estimated conditional probabilities: predict ALE at the reference time t by using course attended in [t-1/ t] .

THE IMPUTATION PROCEDURE: Group B and C

People not attending any course covered by MIUR since 2011 characterize these groups.

The conditional probabilities of each ALE value are estimated
by considering the **observed values in the 2018 sample as target variable**: $I^t = I^t_s$.

- Group B: ALE from 2011 Census is known.
  Because of the MIUR under-coverage, it is necessary to resort to sample survey data**.**
  Conditional probabilities $h(I^t \mid X)$ are estimated by region through hierarchical log-linear models
  $[Prov, I^t]\ [G, C^t, E^t, I^{t-1}, I^t]$

- Group C: No ALE is known.
  Most critical population because of their peculiarity and the limited amount of administrative information.
  The model selected for the first step through cross-validation is
  $[Prov, I^t]\ [G]\ [C^t, I^t]\ [E^t, I^t]\ [I^{apr}, I^t]\ [D, I^t]$

- For the units observed in the sample, the observed values I^t_s is used as prediction in BRI.

SOME RESULTS to evaluate the predictions

Model and sample estimates of 2018 ALE (absolute values in thousands) and absolute (*D*) and relative (*Dr*) differences between model and sample percentages

		Model		Sample		Model – Sample	
	ALE 2018	***a.v.***	***%***	***a.v.***	***%***	$D_i^{(*)}$	$Dr_i^{(*)}$
1	Illiterate	354	0.6	330	0.6	0.04	6.23
2	Literate but no ed. Attainment	2,298	4.1	2,073	3.7	0.37	9.78
3	Primary education	9,297	16.6	9,139	16.5	0.12	0.73
4	Lower secondary education	16,509	29.5	16,169	29.2	0.33	1.11
5	Upper secondary education	19,716	35.3	19,874	35.9	-0.63	-1.76
6	Bachelor's degree	1,977	3.5	1,962	3.5	-0.01	-0.19
7	Master's degree	5,532	9.9	5,599	10.1	-0.22	-2.15
8	PhD	233	0.4	227	0.4	0.01	1.66
Total		**55,917**	**100.0**	**55,373**	**100.0**	***AD*=0.21**	***RD*=2.95**

$$AD = \frac{\sum_{i=1}^{8} |D_i|}{8} = \frac{1}{8}\sum_{i=1}^{8} |fr(\hat{I}^t)_i - fr(\mathrm{I}_s^t)_i|$$

$$RD = \frac{\sum_{i=1}^{8} |Dr_i|}{8} = \frac{1}{8}\sum_{i=1}^{8} \frac{|fr(\hat{I}^t)_i - fr(\mathrm{I}_s^t)_i|}{fr(\mathrm{I}_s^t)} * 100$$

FINAL REMARKS

- ✓ The procedure combines different data sources: Administrative data, sample survey data, and Census data.
- ✓ The imputation models are based on log-linear models, which have the advantage over the traditional hot-deck procedures to be more parsimonious.
- ✓ Methods to estimate the variance (based on resampling techniques) of register-based statistics built by using administrative and sampling data are being tested
- ✓ Istat has planned to produce BRI on a yearly basis, hence the imputation model proposed in the paper should be modified in order to include sampling information referring to each year.
- ✓ An important issue is related to the production of 2021 Census figures. Further studies are needed to produce predictions at a finer classification, as required by Eurostat.

Thank you!

filippini@istat.it

Filippini R., Di Zio M., Rocchetti G.,
Istat, Italian National Institute of Statistics

UN/ECE Workshop on Statistical Data Editing
(31 August - 4 September 2020)