3-7 October 2022, (virtual)
EXPERT MEETING ON STATISTICAL DATA EDITING

VARIANCE ESTIMATION FOR THE VARIABLE “ATTAINED LEVEL OF EDUCATION” IN THE ITALIAN BASE REGISTER OF INDIVIDUALS: A COMPARISON BETWEEN ANALYTICAL AND MONTECARLO ESTIMATES

MARCO DI ZIO, **ROMINA FILIPPINI**, SIMONA TOTI
Istat | DCME

Outline

- Introduction
- The informative context
- The ALE estimation procedure
- Variance estimation of ALE
- Experimentation
- Results
- Conclusion

Introduction

- The Attained Level of Education (ALE) of the Permanent Italian Census relies on a high amount of **administrative information**. Nevertheless, it is necessary to resort to sample survey data to cope with delay of information and coverage gaps.
- Istat adopted a **mass imputation** approach integrating administrative and survey data for the ALE estimation of the Italian resident population, based on a sequence of log-linear imputations.

Register-based statistics easily allow to compute estimates for many different **domains of interest** → **An agile measure of uncertainty is needed.**

GOAL: identify a "simple" analytical approach for the variance estimation of imputed ALE

The informative context

- The procedure for the ALE prediction is obtained by integrating different data: Administrative (BRI and MIUR), 2011 traditional Census and sample survey (CS)
- Different pattern of information determine the partition of the population of interest into **three subgroups**.

Source:	**BRI**	**MIUR**		**2011 Census**	**CS**	
Available information:	Core information *t*	ALE*t-2*	Course attendance *(t-2,t-1)*	ALE*t-2*	ALE*t*	**Sub population**
Coverage	■	■	■		■	**A**
	■	■		■	■	**B**
	■				■	**C**

The main difference is between group A and the others:

- Group A is composed by "**Active**" people: attending a course in academic year (t-2, t-1)
- Groups B and C are "**Inactive**" people

The ALE estimation procedure (1/2)

- **«Active» people**: administrative data provide longitudinal information on school enrolment.
  - "**No-Change**": zero probability of changing the educational level, from t-2 to t

    -> deterministic imputation.

    $ALE_t = ALE_{t-2}$

  - "**Change**": non-zero probability of obtaining a higher qualification

    -> the estimate is based on individual and schooling characteristics

    $\Pr(ALE_t | ALE_{t-2}, \text{age, citizenship, school attendance})$

    The model is estimated using only administrative information

The ALE estimation procedure (2/2)

- **«Inactive» people**: the model is estimated on units interviewed in CS_t considering the observed ALE as target variable.

  Pr (ALE_t | ALE_{t-2}, age, citizenship, province of residence, gender)

  Observed in CS_t

- **CS_{2018}:** people interviewed in the t-1 Permanent Census Survey (and not in CS_t) to which the imputation process assigns a lower ALE than that observed in CS_{t-1}

  ALE_t = ALE observed in CS_{t-1}

  to maintain the consistency between CS information

Variance estimation of ALE (1/2)

- Only the set characterized by probabilistic imputation (*Pr)* contributes to the variability of the estimator:

$$V(\hat{\bar{Y}}) = \left(\frac{N_{Pr}}{N}\right)^2 V\left(\hat{\bar{Y}}_{Pr}\right)$$

  - composed by different subpopulations characterized by specific imputation models:

    (simplifying)

    - Active people: target variable from administrative data known for all individuals
    - Inactive people: target variable from CS_t known for individual in the sample

$$V\left(\hat{\bar{Y}}_{pr}\right) = \left(\frac{N_{Active}}{N_{Pr}}\right)^2 V\left(\hat{\bar{Y}}_{Active}\right) + \left(\frac{N_{Inactive}}{N_{Pr}}\right)^2 V\left(\hat{\bar{Y}}_{Inactive}\right)$$

Most of the imputations of $\bar{Y}$ are carried out by means of a saturated log-linear model: the basic formula for the variance estimation in presence of donor imputation is considered (Wolter, 2007, Brick *et al.*, 2004)

Variance estimation of ALE (2/2)

- Inactive people: the variance formula is adapted taking into account the imputation classes

$$V\left(\hat{\bar{Y}}_{Inactive}\right) = \underbrace{\left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2}_{\text{Sampling}} + \underbrace{\frac{1}{(N)^2}\sum_k \left(N_{x_k} - n_{x_k}\right)\left(1 - \frac{1}{n_{x_k}}\right)\sigma_{y|x_k}^2}_{\text{Imputation}}$$

x_k for k=1,…, K are the imputation cell

- Active people: (1) individuals in the sample, (2) individuals not in the sample characterized by model imputations

$$V\left(\hat{\bar{Y}}_{Active}\right) = \underbrace{\left(\frac{n}{N}\right)^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2}_{\text{Sampling}} + \underbrace{\left(\frac{N-n}{N}\right)^2\left(1 - \frac{(N-n)}{N}\right)\sigma_{y_{ADMIN}}^2}_{\text{Imputation}}$$

estimated using the frequencies computed on the sample

estimated only on administrative data

- CS2018: for simplicity, in this experimentation we just consider the sample variance for this subpopulation.

Experimentation

- **Dataset**: people with age >= 9, resident in the Emilia Romagna region in year 2019 (4 mln individuals). For this population the official imputed ALE is available and it is considered as it was the true value.
- **Simulations**: 200 simple random samples (4.7% of total population); for each sample, the ALE mass imputation procedure is applied and the resulting frequency distribution of ALE ($\hat{\bar{y}}$) is calculated.
- **Variance estimation** of $\hat{\bar{y}}$ is computed by following two approaches:
  - Monte Carlo approach (MC): applying the classical variance formula (benchmark)
  - Analytical approach (AN): applying the formulas described in the previous slide for each simulation.
- The processing time for the simulation procedure takes about 2 days.
- The variance estimates on the simulated datasets are obtained using ad-hoc R code.

Results (1/2)

Table 1: Variance estimates for the mass-imputation of ALE

		$\hat{V}(\hat{\bar{Y}})$		Ratio (AN/MC)	N pop (.000)
		MC	AN		
1	Illiterate	1 .42E-08	1.53E-08	1.08	17
2	Literate but no formal att.	6 .62E-08	7.34E-08	1.11	162
3	Primary education	1 .82E-07	1.92E-07	1.06	660
4	Lower secondary ed.	3 .78E-07	3.84E-07	1.01	1,152
5	Upper secondary ed.	3 .69E-07	3.66E-07	0.99	1,514
6	Bachelor's degree	7 .49E-08	7.21E-08	0.96	166
7	Master's degree	9 .81E-08	1.17E-07	1.19	451
8	PhD level	1 .16E-08	1.09E-08	0.94	20

- The imputation procedure gives origin to very stable results for each of the 8 ALE modality
- The analytical approach shows a good approximation of the estimates if compared with the MC results
- The best results are obtained for the ALE's modalities where the frequency of the population is high.

Results (2/2)

Table 2: Variance estimates for the mass-imputation of ALE by subpopulation

	Active		Inactive B		Inactive C		Inactive CS_{2018}	
	MC	AN	MC	AN	MC	AN	MC	AN
1	1 .39E-12	1 .36E-12	1 .16E-08	1 .68E-08	2 .08E-06	1 .61E-06	9 .83E-08	2 .40E-07
2	8 .26E-11	8 .00E-09	7 .14E-08	9 .25E-08	5 .70E-06	5 .14E-06	3 .85E-07	1 .21E-06
3	1 .14E-09	2.03E-08	2 .79E-07	2 .72E-07	5 .70E-06	7 .34E-06	2 .66E-06	2 .98E-06
4	4 .35E-09	2.71E-08	5 .03E-07	5 .21E-07	2 .01E-05	1 .90E-05	5 .49E-06	4 .52E-06
5	1 .37E-08	3.15E-08	5 .14E-07	4 .92E-07	1 .88E-05	1 .91E-05	4 .10E-06	4 .47E-06
6	9 .02E-09	1.46E-08	9 .72E-08	9 .76E-08	4 .56E-06	3 .58E-06	5 .46E-07	1 .11E-06
7	4 .22E-09	3.56E-09	1 .37E-07	1 .47E-07	6 .96E-06	8 .29E-06	2 .57E-06	1 .83E-06
8	1 .22E-10	9.05E-11	1 .40E-08	1 .50E-08	8 .32E-07	4 .93E-07	5 .83E-08	1 .67E-07

- Good approximation of the AN approach for the Inactive subpopulations
- In the Active subpopulation the difference between MC and AN is higher for modalities 3 (Primary ed.) and 4 (Lower secondary ed.)
- Modality 1 (Illiterate) and 2 (Literate but no formal att.) ambiguous and difficult to attribute with certainty

Conclusions and future developments

- The results of the experiments are encouraging and suggest that the proposed method for the variance estimation of ALE for the permanent Census can be taken into account for further studies.
- The analytical formula is easy to implement and it is not time consuming as a bootstrap approach. This is a particularly important aspect for a complex case like that under evaluation.
- Further experiments should be performed to put the proposed method into the process, more studies should be referred to the variance estimation of ALE within domains composed of a moderate/small number of units as for instance Municipalities.

Thank you

MARCO DI ZIO | dizio@istat.it

ROMINA FILIPPINI| filippini@istat.it

SIMONA TOTI | toti@istat.it

Istat | Istituto Nazionale di Statistica