Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

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

Prepared by Di Zio M., Filippini R., Rocchetti G., Istat, Italy

I. Introduction

1. The Italian Base Register of Individuals (BRI) is a comprehensive statistical register storing data gathered from various data sources. In BRI, core variables like *place* and *date of birth, gender*, *citizenship* are associated to each unit. Moreover, a classification variable denoting people resident in Italy is introduced. The subset of resident people is the basis of the next Italian census that will be as much as possible register-based. According to this idea, given the high amount of available administrative information, a prediction of the *attained level of education* for the individuals in BRI is proposed.

2. The main sources containing administrative information originate from the Ministry of Education, Universities and Research (MIUR). MIUR provides information about the attained level of education and student's attendance to a course (e.g., attending first year of primary education). Administrative data refer to students from 2011 onwards. For the rest of the people not included in this period, we may resort to the 2011 Census information. Unfortunately, not all the people classified as resident after 2011 belong to these two data sources, as for instance immigrated people entered Italy after the Census and not attending any educational course. Another important source of information is the sample survey collected for the permanent Italian census starting from 2018. These data are particularly important to fill the informative gaps of MIUR and 2011 Census data. We remind that the so-called permanent census is a system of yearly surveys and administrative data organized in registers that once combined are supposed to provide each year the main Census figures.

3. The focus of the work is on the prediction or mass-imputation (in this application, we adopt both the terms as synonymous) of the attained level of education in the Base Register of Individuals. A mass-imputation is justified by the high amount of detailed available information. Similar studies are available in other NSIs, see for instance Scholtus and Pannekoek (2015), Daalmans (2017).

4. The procedure discussed in the paper follows the study by Di Cecco *et al*. (2018) where different methods were evaluated on preliminary data. The procedure chosen for the prediction of level of education is applied to the 2018 BRI and, although the 2018 sample survey is not yet completely cleaned, we expect survey data be close enough to the final ones that will be used for producing official estimates.

5. The paper is structured as follows. Section II depicts the informative context describing the data sources used for the prediction. The imputation procedure is presented in Section III, and the results of some analysis carried out in order to assess the results are reported in Section IV. Some final remarks are presented in Section V.

II. Informative context

A. Data sources description

6. In carrying out the prediction procedure, data of different nature are jointly used. In fact, it combines administrative, traditional Census and sample survey data. The preliminary estimates of population in BRI considered as usually resident in Italy at 31st of December 2018 amounts to 60,441,487 units. BRI also includes the main personal information, i.e. the core variables - place and date of birth, gender, citizenship - used in the present study. Core variables are obtained through an extensive utilization of administrative data, reconciled and stored yearly in the BRI.

7. Administrative information on the attained level of education (ALE hereinafter) is gathered making use of the information collected by the Ministry of Education, University and Research and processed in ISTAT with the purpose of creating a database on Education and Qualification, named BIT (see Runci *et al.*, 2017). BIT collects, checks and integrates data from different sources provided by MIUR, on a yearly basis, about the ALE and the attendance to a course (e.g., attending first year of primary education) of students. Data on the ALE is available at the reference time, set on 31st of December 2017; meanwhile, data on the attendance to a course refer to the academic year 2017/2018 (BIT 2017 hereinafter). Summarizing, BIT 2017 collects information on the ALE achieved between 2012 and 2017 for 13.966.581 units.

8. People that have not attended any course since 2011 are not in BIT. For our purposes, we turn to data from 2011 Census to fill the gap. The 2011 Census operations, whose reference date was October 31, 2011 (CENS 2011 hereinafter), surveyed 59,433,744 individuals. For our needs, data on educational attainment was collected for persons aged 9 or older, who were still living in Italy on the 31st of December 2018, for a total of 53,745,821 units.

9. Another important source of information is given by the 2018 sample survey conducted for the permanent Italian census. In this sample, units are asked about their educational level. More precisely, survey data used for the prediction are obtained by the integration of the list and the area samples. They approximately amount to the 5% of the total population.

10. In addition, auxiliary administrative data on ALE can be taken from the registration and cancellation forms for transfer of residence (APR4) gathered for the period 2012-2017. ALE on APR4 is self-declared by individuals that fill the form in order to apply for a new registration in Italy coming from abroad and/or when they change usual residence. In APR4, ALE comes with 4 levels of classification:

1 - Up to the elementary license corresponding to ISCED[1] 0, 1;
2 - Lower secondary education corresponding to ISCED 2;
3 - Secondary and short cycle tertiary education corresponding to ISCED 3, 4, 5;
4 - Tertiary and post tertiary education - ISCED 6, 7, 8.

B. Reconciling classifications and quality analysis

11. Both CENS 2011 and BIT 2017 use detailed and reciprocally consistent classifications of educational level; consequently, data were univocally reclassified according to a 8-items dissemination classification (named CDIFF) adopted by Istat for the purpose of disseminating permanent census data on the ALE. In particular, mapping operations are carried out such that items in the classifications adopted by CENS 2011 (12 items and a separate question for those having obtained a doctoral or equivalent level) and BIT 2017 (16 items) could be homogenously reclassified into the new one (17 items; ISTAT 2017 hereinafter). Furthermore, we univocally recode data into the CDIFF classification (Table 1).

[1] ISCED (International Standard Classification of Education) is a statistical framework created by UNESCO for organizing information on education (http://uis.unesco.org/).

Table 1. Correspondence table between ISTAT 2017 and CDIFF 2018 classifications on Attained Level of Education

CDIFF 2018 classification	BRI and Survey Sample 2018 for the permanent census 2020 classification
1 Illiterate	01) Illiterate
2 Literate but no formal educational attainment	02) Literate but no formal educational attainment
3 Primary education	03) Final assessment (Primary school)
4 Lower secondary education	04) Diploma of lower secondary education
5 Upper secondary education	05) Diploma of upper secondary education (2-3 years)
	06) IFP - Vocational training qualification (three-year courses)/ Professional diploma (fourth year)
	07) Diploma of upper secondary education (4-5 years)
	08) Certification of higher technical specialization (IFTS)
	10) Fine Arts, drama, Dance and Music Diploma (2-3 years)
6 Bachelor's degree or equivalent level	09) Diploma of Higher Technical (ITS)
	11) University diploma
	12) Fine Arts, Drama, Dance and Music First level academic diploma (Bachelor's
	13) *Laurea triennale* (I level , Bachelor's degree)
7 Master's degree or equivalent level	14) Fine Arts, Drama, Dance and Music Second level academic diploma (Master's
	15*) Laurea (4-6 years*, Master's degree)
	16) *Laurea biennale specialistica* (II level, Master's degree)
8 PhD level	17) Research Doctorate (PhD)/ Advanced research academic diploma

12. It is worth noticing that the choice of using both CENS 2011 and BIT 2017 comes from a comprehensive data quality analysis (see Di Cecco *et al.*, 2018). Here, for our purpose, we shortly present the principal results on data consistency, based on a cross-comparison at a micro level.

Table 2. Consistency of data on attained level of education in CENS 2011 and BIT 2017

	a.v.	*%*
BIT 2017 > CENS 2011	7,705,099	80.2
BIT 2017 = CENS 2011	1,763,050	18.3
BIT 2017 < CENS 2011	144,332	1.5
Total RBI 2018 population >8 years-old	**9,612,481**	**100.0**

13. Table 1 shows that out of about 9.6 millions of individuals co-present in the two datasets, 18.3% shows the same level of education in both sources. Moreover, 7.7 million (80.2%) gained a higher degree than observed in CENS 2011. The remaining 1.5% - almost 144 thousands population units – instead, reports inconsistent data, being the most recent level of education lower than the one assigned in CENS 2011.

14. The reasons of such inconsistencies are not easily identifiable. They are probably due to response errors. For instance, as far as cases in which BIT 2017 data are lower than data registered in CENS 2011 operations, the majority of cases concerns units reporting a "Diploma of upper secondary instead" of a "Diploma of lower secondary education", or a "Laurea triennale (I level , Bachelor's degree)" instead of a "Laurea (4-6 years, Master's degree)". To some extent, they may also be caused by linkage errors.

15. In order to reconcile the information, we update CENS 2011 data with BIT 2017 data. In fact, data that MIUR provided on yearly basis are usually reliable; furthermore, the process leading to the construction of BIT is a well-established one and is characterized by high quality standards (see Runci *et al*., 2017).

C. Coverage and subpopulation segments

16. Table 3 classifies target population in main subgroups categorized by presence or absence of information on educational attainment. Data from BIT 2017 covers 22.1% of the overall BRI 2018 population; instead, people observed not in BIT but in CENS 2011 provides the most consistent part of coverage (67.7%). As far as we consider information available for people being at least 9 years-old, the total coverage of administrative data is about 95%.

Table 3. Reference population by presence in CENSUS 2011 and BIT 2017

	Total population		**9 years old and more**	
	a.v.	***%***	***a.v.***	***%***
Present in BIT 2017	13,388,761	22.1	12,292,331	21.9
Present in CENS 2011 only	40,938,914	67.7	40,938,904	73.2
Records without information on ALE	6,113,812	10.1	2,686,016	4.8
Total BRI 2018 Population	**60,441,487**	**100.0**	**55,917,251**	**100.0**

17. Despite of the high coverage rate of administrative and Census data, there are still about 2.7 million of eligible units older than 9 years without data on ALE. These are either people entered Italy after 2011 that have not attended any course covered by MIUR, or people not caught by the 2011 Census. Concerning the latter, during post-Census operations, the collaboration with municipalities (named SIREA operation) allowed to identify 1,403,991 individuals who could not be found in CENS 2011 but that were resident: they were “detected” for the purpose of counting resident population but they didn’t answered the questionnaire.

18. Data sources have some informative gaps. It is worthwhile to remark that BIT does not include qualification courses like Fine Arts, Drama, Dance and Music academic diplomas and more relevantly training and vocational careers managed by Italian Regions that are not required to provide data to MIUR. Although this lack of information has an impact on under-coverage of units, we expect that the major pitfall be about the underestimation of the level of education of units in the subset reporting 2011 Census ALE. It is worth reminding that BIT is able to trace only students enrolled in an educational course between 2012 and 2017. For all population units for whom schooling or training is over, it has to be experimented the use of auxiliary information.

19. Another critical issue concerning BIT has to do with timeliness. The lag between the moment in which BIT data are available and BRI reference time makes it necessary to implement procedures for the prediction of the variable. As shown afterwards, the attained level of education at time *t* should be predicted by having available one-year lagged data; information of attendance of educational courses has instead a lower delay, being related to the academic year [*t-12* months, *t*].

20. We need to resort to additional data to both fill the informational and temporal gap. The core information comes from the survey data collected during the permanent census operation in October 2018. The sample survey gathers information for about 2.6 million of units (Table 4). As well as for BIT 2017, the sample survey ALE has been originally classified according to the ISTAT 2017 classification and has been recoded to the 8-items of dissemination for the purpose of prediction ALE for 2018.

Table 4. Sample 2018 and APR4 population by presence in CENSUS 2011 and BIT 2017

	Sample 2018		**APR4**	
	a.v.	***%***	***a.v.***	***%***
Present in BIT 2018	554,791	21.4	1,054,013	20.2
Present in CENS 2011 only	1,965,509	75.8	3,121,639	60.0
Records without information on ALE	71,403	2.7	1,031,026	19.8
Total	**2,591,703**	**100.0**	**5,206,678**	**100.0**

21. The additional auxiliary administrative information on ALE from APR4 forms allows collecting data on about 5.2 million of units. Mostly of them overlap data from CENS 2011. It is worth noticing that 19.8% of observations covers the segment of population without any administrative information on ALE.

III. Imputation of the *attained level of education*

A. The imputation procedure

22. In this section, we illustrate a procedure for the prediction of the attained level of education at the reference year *t* of the resident population in BRI. At time *t*, the BRI contains the following structural information:

- The resident population at 31/12/*t*
- Gender - (G)
- Date of birth - (D)
- Place of birth - (P)
- Country of citizenship at 31/12/*t*.

From the MIUR administrative data, we have used:

- the attained level of education at 31/12/*t*-12 months - (I^{t-12});
- the year attendance of educational courses in the time period [*t*-12, *t*], e.g. 1st year, 2nd year,.. - (F^t);
- the type of school (liceo, other) – (L).

We remind that the year of attendance of previous years as [*t*-24, *t*-12] and so on are available as well. From the APR4 administrative data, we have exploited the self-declared ALE (I^{apr}) with 4 levels of classification as detailed in Section II.

For the application of the procedure, the following transformed variables are also computed:

- Italian, not Italian citizenship – (C^t)
- Age in 8 classes - (E8).

23. As so far mentioned, in addition to administrative data, we may resort to information on the ALE from the 2011 Italian Census and from a sample survey referring to the target time *t*. We notice that for units not in MIUR but in the Census 2011, the ALE at time *t*-12 (I^{t-12}) is the one reported in the 2011 Census. Finally, we denote with I^t_S the ALE at time *t* observed in the sample survey.

24. Let I^t be the target variable, i.e. the ALE at time *t* that we would like to predict. We denote with A the subset of data for which information from MIUR is available, with B the set of units for which only information from Census 2011 is available and with C the subset of data observed neither in the Census nor in MIUR. Table 5 depicts the data/information scenario that we need to take into account when making predictions for I^t. We remark that the grey cells in the table represent missing data and the relative frequencies of groups with respect to the population with at least 9 years is reported in the last column.

Table 5. Tabular representation of the informative context for mass-imputation of the attained level of education at time *t*

X_{BRI}				X_{miur}				**Sample**	**Prediction**	**Group**
G	**E**	**P**	**C^t**	**$L^{(t)}$**	**$I^{(t-12)}$**	**$F^{(t)}$**	I^{apr}	**I^t_S**	**I^t**	
									▨	A 22%
								▨	▨	
							▨	▨	▨	
						▨			▨	B 73%
						▨		▨	▨	
						▨	▨	▨	▨	
				▨	▨	▨			▨	C 5%
				▨	▨	▨		▨	▨	
				▨	▨	▨	▨	▨	▨	

25. The general idea is to estimate a model for the prediction of I^t given the values of known covariates X. In particular, we estimate the conditional probabilities $h(I^t |X)$ and then impute I^t by taking at random a value such a distribution. The conditional probabilities $h(I^t |X)$ are estimated by means of log-linear models as follows. First, a log-linear model is applied to the contingency table obtained by cross-classifying the variables (I^t , X) to estimate their expected counts $\widehat{N}(I^t ,X)$ from which we can compute the counts $\widehat{N}(X)$. The estimated conditional probability distribution $\hat{h}(I^t | X)$ is easily obtained by computing $\widehat{N}(I^t , X)/ \widehat{N}(X)$. This approach includes as a special case the random hot-deck when a saturated log-linear model is assumed, but it has the advantage that allows the possibility of using a more parsimonious model. This is an important characteristic when the number of cells increases.

26. In order to take into account the missing data mechanism, sampling weights adjusted for non-response (that is indeed low in this survey) are used. It is adopted a pseudo-maximum likelihood approach that consists in estimating log-linear models on weighted count data (Thibaudeau *et al*., 2017, Skinner *et al*., 2010).

27. Similarly to hot-deck, it may happen that a missing observation is not imputed because their covariates have a pattern not observed in the sample. In order to overcome this problem, a sequence of log-linear models with increasing levels of aggregation of covariates are used to impute values. Models are chosen by means of cross-validation, in fact different covariates may induce the selection of different models.

28. For the units observed in the sample, the observed values I^t_S is used as prediction in BRI. This choice has the additional advantage of preserving the consistency of predicted ALE in BRI with the variables observed in the sample survey, and inference on those variables may use micro-data in BRI without any problems concerning micro-consistency.

29. Different log-linear models are used within groups A, B and C, mainly because of the different available information. The principal difference is between the group A and the others. In group A, a log-linear model is estimated by using only administrative data, while for the other groups log-linear models are estimated by using survey data as well. In the following, some details for each group are given.

B. Imputation in Group A.

30. This group is characterized by active people, which means people that are currently attending a course in the reference year. Administrative information is particularly important in this group, in fact the aim is essentially to predict the attainment of educational level given that is known which is the year of the course they are attending during the year [*t*-12, *t*].

31. We have decided to estimate the conditional probabilities of obtaining a new educational level by using only administrative data. The imputation method consists in the estimation of a model applied to data referring to 1 year before the time reference, and then by applying the estimated model to the year of reference. In the specific application, firstly we have estimated the conditional probabilities on data to predict the ALE at 2017 (known in the administrative data) by using data available in the interval time [2016, 2017]. Then, we have applied the model to predict the ALE at the reference time *t*=2018. The underlying idea is that there is no variation into the conditional probabilities in one year, and that the error introduced by this assumption is lower than the sampling error introduced by using sample survey data.

32. The log-linear model used in the first step of the sequence of imputations is the saturated model:

$$[C^t, E8^t, F^t, L^t, I^t] \qquad (1)$$

that is first estimated with *t*=2017 by region and then applied to *t*=2018.
Although, as previously declared, a sequence of models is used to impute data, most of the non-responses are imputed by using (1).

C. Imputation in Group B.

33. People not attending any course covered by MIUR since 2011 characterize this group. These are people that either have decided to stop their studies or that are still attending some courses unfortunately not covered by MIUR. Because of the MIUR under-coverage, it is necessary to resort to sample survey data. The conditional probabilities $h(I^t \,|X)$ are estimated by region through the log-linear model

$$[\mathrm{Prov}, \mathrm{I}^t]\ [\mathrm{G}, \mathrm{C}^t, \mathrm{E8}^t, \mathrm{I}^{t\text{-}12}, \mathrm{I}^t\ ] \qquad (2)$$

where Prov is the province of residence. The model is estimated on t=2018 by considering the observed values in the sample, i.e., $I^t = I^t_s$.

Also in this group, a sequence of imputation models are used, however almost all the units are imputed by using model (2).

D. Imputation in Group C.

34. This group is characterized by two types of units (denoted by the variable D):

- individuals resident on the Italian territory but not detected by the 2011 Census (D=1);
- individuals entered Italy after 2011 and that have not attended any training courses released by MIUR from 2011 onwards (D=2).

35. These are two populations with distinct socio-demographic characteristics (see Di Cecco *et al.*, 2018) and it is important to include the variable D in the model to distinguish them. Although affected by missing values, another important information is the self-declared ALE I^{apr} reported in APR4. This cannot be used directly as a value to assign to the I^t both because of its level of classification that is too much aggregated, and because of its level of quality being a self-declared variable. However, it results strongly correlated to I^t, therefore it is used as a covariate in the model. This is certainly the most critical population because of their peculiarity and the limited amount of administrative information. In order to fill the lack of knowledge, it is important to use survey data that report the ALE at time *t*.

36. The model selected for the first step through cross-validation is

$$[\mathrm{Prov}, \mathrm{I}^t]\ [\mathrm{E8}^t, \mathrm{I}^t]\ [\mathrm{G}]\ [\mathrm{C}^t, \mathrm{I}^t\ ]\ [\mathrm{I}^{apr}, \mathrm{I}^t\ ]\ [\mathrm{D}, \mathrm{I}^t] \qquad (3)$$

The model is estimated on t=2018 by considering the observed values in the sample, i.e., $I^t = I^t_s$. Also in this group, almost all the units are imputed according to this model.

IV. Some results to evaluate the predictions

37. In this section, we illustrate the results of some analysis carried out in order to assess the imputations. Analysis on micro-data and aggregates are performed. Results on the population are analysed and compared with the data collected in the sample of the 2018 permanent census, appropriately weighted, and with data from administrative sources and 2011 Census where available. In the micro level analysis, the transitions from 2017 observed ALE to 2018 estimated ALE are studied. In the macro level validation, comparisons between distributions of observed and estimated 2018 ALE are analysed.

38. For groups A and B, transitions between observed and estimated ALE provide a first evaluation of the results. In group A, the estimated 2018 ALE is in most cases consistent with the 2017 information from administrative sources (Table 6). This happens when the estimated 2018 ALE confirms the 2017 ALE or increases it by 1 degree. On the other side, inconsistencies arise when the 2018 estimated ALE is lower than the observed 2017 ALE or when the estimated 2018 ALE is more than one degree higher than the 2017 ALE. The inconsistencies regard the subset of people interviewed at the 2018 sample, for which the collected information is used as prediction (see Section III).

39. It is worthwhile to report that, data editing of sample data was performed mainly looking for the consistency within the sample. Administrative data were used in the sample data editing and imputation process for two main purposes: a macro level validation of the sample data on ALE and a micro level

comparison when the sample data on ALE was inconsistent within the sample. Only in this case the administrative ALE was considered in substitution of the ALE declared by respondents. In order to minimize inconsistencies between administrative and survey data, administrative information could be jointly used with survey in the data editing process.

Table 6. Group A: transition from ALE 2017 (administrative data) to ALE 2018 (estimated data) – row percentage and total absolute value in thousands

		ALE 2018 (estimate)								
	ALE 2017 (administrative)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total (*a.v.*)**
1	Illiterate	-	-	-	-	-	-	-	-	-
2	Literate but no ed. attainment	0.0	**65.8**	34.2	0.0	0.0	0.0	0.0	-	1,628
3	Primary education	0.0	0.0	**65.9**	33.5	0.6	0.0	0.0	-	1,791
4	Lower secondary education	0.0	0.0	0.0	**77.1**	22.7	0.1	0.1	0.0	3,565
5	Upper secondary education	0.0	0.0	0.0	0.0	**90.4**	7.1	2.4	0.0	3,449
6	Bachelor’s degree	0.0	-	0.0	0.0	0.2	**81.4**	18.2	0.2	923
7	Master's degree	0.0	0.0	0.0	0.0	0.1	0.1	**97.0**	2.8	892
8	PhD	-	-	-	0.0	0.0	0.0	0.8	**99.1**	45
	Total	**0.0**	**8.7**	**14.1**	**27.3**	**32.0**	**8.2**	**9.1**	**0.6**	**12,292**

40. In group B, the estimated 2018 ALE shows some inconsistencies with ALE in 2017 (Table 7). The information on ALE in 2017 derives from the 2011 Census and regards individuals who have not enrolled in any standard training course from 2011 to 2017 so the educational level is not changed until 2017. The basic hypothesis is that the information collected in 2011 is not error-free and that the administrative data on school attendance may be under-covered, therefore, for this sub-population, the information on the educational level in 2017 can be corrected based on the information from the 2018 sample. There is no restriction on the fact that the estimated ALE in 2018 should be higher than that of 2017.

Table 7. Group B: transition from ALE 2017 (CENS 2011) to ALE 2018 (estimated data) – row percentage and total absolute value in thousands

		ALE 2018 (estimate)								
	ALE 2017 (CENS 2011)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total (*a.v.*)**
1	Illiterate	**46.2**	22.1	19.7	8.7	2.6	0.2	0.5	0.0	371
2	Literate but no ed. attainment	5.3	**38.4**	42.2	10.1	3.3	0.1	0.6	0.0	1,182
3	Primary education	0.6	5.6	**78.7**	12.5	2.3	0.1	0.3	0.0	7,301
4	Lower secondary education	0.1	0.6	5.8	**78.9**	13.9	0.2	0.5	0.0	12,938
5	Upper secondary education	0.1	0.2	1.0	6.5	**89.3**	1.2	1.6	0.0	14,051
6	Bachelor’s degree	0.0	0.2	0.6	3.1	16.8	**61.8**	17.1	0.2	943
7	Master's degree	0.0	0.1	0.7	2.0	3.9	2.3	**90.0**	1.0	4,016
8	PhD	0.0	0.1	0.4	1.0	2.1	0.7	27.4	**68.4**	136
	Total	**0.7**	**2.6**	**17.7**	**30.0**	**36.4**	**2.1**	**10.1**	**0.4**	**40,939**

41. To evaluate the imputation procedure in a macro level approach, the estimated ALE in 2018 ($\widehat{I^t}$), obtained on the Italian resident population is compared with the data collected in the 2018 census sample, appropriately weighted (I_s^t). In particular, we focus on the differences between the frequency distributions of estimated 2018 ALE in BRI and the distribution computed on weighted sample data. A synthetic measure of the difference between distributions is given by the average of the absolute value of the differences between percentage of each item, in absolute (*AD*) and relative (*RD*) terms. Specifically:

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

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

where $fr(\widehat{I^t})_i$ is the relative frequency of ALE item *i* estimated in 2018 and $fr(I_s^t)_i$ is the relative frequency of ALE item *i* estimated with the 2018 weighted sample.

42. The macro level comparison between BRI and sample estimates shows that the two distributions are very similar (Table 8). The distribution of the estimated ALE differs from the weighted sample data by 0.21% points on average on each item; the differences are concentrated in level 5 “Upper secondary

education", which is the most frequent one. In relative terms (*Dr*), differences are concentrated in the extreme and less frequent levels. In particular level 1 "Illiterate" and level 2 "Literate but no formal educational attainment" are confused and difficult to be predicted.

Table 8. 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**

(*)Warning: the calculations from the table may give different numbers due to the approximation

43. The distribution of ALE 2018 will be published yearly by ISTAT taking into account for some other variables such as gender, age classes and citizenship so it is important to take into account the distributional accuracy in these specific subpopulations. Looking at the distribution of ALE 2018 by citizenship, differences between estimated and weighted sample data are evident especially on the sub-population of Not Italian people (Table 9) in which we observe an average difference of 0.38 points on each estimated item with respect to the weighted sample. The Not Italian subpopulation is small with respect to the total (9%) and is characterized by particular features and less information available determining a different fit of the model.

Table 9. Model and sample estimates of 2018 ALE (absolute values in thousands) and absolute (*D*) differences between model and sample percentages, by citizenship

	Model				**Sample**				**Model – Sample**	
	Italian		**Not Italian**		**Italian**		**Not Italian**		**Italian**	**Not Italian**
ALE 2018	***a.v.***	***%***	***a.v.***	***%***	***a.v.***	***%***	***a.v.***	***%***	$D_i^{(*)}$	$D_i^{(*)}$
1 Illiterate	271	0.5	84	1.8	258	0.5	72	1.7	0.02	0.15
2 Literate but no ed. attainment	1,979	3.9	320	6.9	1,801	3.5	273	6.2	0.33	0.66
3 Primary education	8,790	17.1	506	10.9	8,663	17.0	476	10.8	0.15	0.04
4 Lower secondary education	14,970	29.2	1,539	33.1	14,702	28.8	1,466	33.4	0.37	-0.33
5 Upper secondary education	18,048	35.2	1,669	35.8	18,248	35.8	1,626	37.0	-0.59	-1.19
6 Bachelor's degree	1,813	3.5	164	3.5	1,815	3.6	146	3.3	-0.02	0.20
7 Master's degree	5,176	10.1	356	7.6	5,282	10.4	317	7.2	-0.26	0.42
8 PhD	215	0.4	18	0.4	212	0.4	15	0.3	0.00	0.05
Total	**51,262**	**100.0**	**4,656**	**100.0**	**50,982**	**100.0**	**4,391**	**100.0**	***AD*=0.21**	***AD*=0.38**

(*)Warning: the calculations from the table may give different numbers due to the approximation

V. Final remarks and future developments

44. In this paper a mass-imputation procedure for the attained level of education is described. The procedure combines different data sources: Administrative data, sample survey data, and Census data.

45. The imputation models are based on log-linear models, which have the advantage over the traditional hot-deck procedures to be more parsimonious. This flexibility is an important issue since as noted in De Waal (2016) "mass imputation relies on the ability to capture all relevant variables and relevant relations between them in the imputation model, and to estimate the model parameters sufficiently accurately".

46. Methods to estimate the variance of register-based statistics built by using administrative and sampling data are being tested. They are based on resampling techniques for finite population, see Chen *et al*. (2019) for a general discussion and Di Consiglio *et al*. (2019) and Scholtus (2018) for the cases of integrated administrative data.

47. 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, that in the illustrated case means it should be designed a model based on sample data related to 2018 and 2019 to predict the ALE 2019.

48. Further analysis will be dedicated to the use of additional information to improve the predictions, for instance, the inclusion of family composition can be important to this aim.

49. An important issue is related to the production of 2021 Census figures. In this paper, ALE is predicted with a classification based on 8 categories, while for the 2021 Census a more detailed classification is required. Further studies are needed to produce predictions for the attained level of education at a finer classification.

References

Chen S., Haziza D., L_eger C., Mashreghi Z., (2019). Pseudo-population bootstrap methods for imputed survey data, *Biometrika*, 106 (2), pp. 369-384.

Daalmans J., (2017). Mass imputation for Census estimation, UNECE Group of Experts on Population and Housing Censuses, Geneva, 4-6 October 2017.

de Waal T., (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys, *Statistical Journal of the IAOS*, 32, pp. 231-243.

Di Consiglio L., Di Zio M., Filipponi D. (2019). An empirical evaluation of latent class models for multisource statistics. ITACOSM 2019 - Florence 5-7 June 2019.

Di Cecco D., Di Laurea D., Di Zio M., Filippini R., Massoli P., Rocchetti G. (2018). Mass imputation of the attained level of education in the Italian System of Registers, UNECE, Workshop on Statistical Data Editing, Neuchâtel, Switzerland, 18-20 September 2018.

Runci M.C., Di Bella G., Cuppone F. (2017). Integrated Education Microdata to Support Statistics Production. In: Lauro N., Amaturo E., Grassia M., Aragona B., Marino M. (eds) Data Science and Social Research. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham.

Scholtus S. and J. Pannekoek (2015). Mass-imputation of educational levels (In Dutch), Statistics Netherlands, Internal report, The Hague/Heerlen.

Scholtus S. (2018). Variances of Census Tables after Mass Imputation, Discussion paper CBS.

Skinner, C.J. and Vallet, L.-A. (2010). Fitting log-linear models to contingency tables from surveys with complex sampling designs: an investigation of the Clogg-Eliason approach. *Sociological methods & research*, 39 (1), pp. 83-108.

Thibaudeau, Y., Slud, E. and Gottschalck, A. (2017). Modeling log-linear conditional probabilities for estimation in surveys. *The Annals of Applied Statistics.* 11, pp. 680-697.