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

Marco Di Zio, Romina Filippini, Gaia Rocchetti [1]

Abstract

The paper describes the mass imputation procedure of the level of education in the Base Register of Individuals of the Italian National Institute of Statistics – Istat. The procedure integrates data of different nature: information deriving from administrative sources, from the 2011 population census and from the 2018 permanent census survey. The procedure is complex and is composed of different steps depending on the information of the sources. The imputation is based on log-linear models which, compared to classical methods such as the hot-deck imputation, allow greater flexibility in modelling associations. The work also illustrates the comparisons between the register estimates obtained with imputation with those of the census sample survey in order to highlight the advantages and limitations of the proposed procedure.

Keywords: statistical register, data integration, mass imputation.

[1] Marco Di Zio (dizio@istat.it); Romina Filippini (filippini@istat.it); Gaia Rocchetti (grocchetti@istat.it), Italian National Institute of Statistics – Istat.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction[2]

The Italian Base Register of Individuals (BRI) is a comprehensive statistical register storing data gathered from various data sources. In BRI, core variables as 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 resident people in BRI is proposed.

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 in the 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 sets, as for instance immigrated people entered Italy after the Census that have not attended 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 organised in registers that once combined are supposed to provide each year the main Census figures.

The focus of the work is on the prediction or mass imputation (in this application, we adopt the two terms as synonymous) of the attained level of education in the Base Register of Individuals. A mass imputation procedure 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).

[2] Although the article is the result of a joint work, the single parts are authored as follows: Sections 1 and 5, Paragraphs 3.1 and 3.3 by Marco Di Zio; Section 2 and Paragraphs 2.1, 2.2 and 2.3 by Gaia Rocchetti; Paragraph 3.2 and Section 4 by Romina Filippini.

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 to be close enough to the final ones that will be used for producing official estimates.

The paper is structured as follows. Section 2 depicts the informative context describing the data sources used for the prediction. The imputation procedure is presented in Section 3, and some results of the analysis carried out in order to assess the outcomes of the procedure are reported in Section 4. Section 5 presents some final remarks and future developments.

2. Informative context

2.1 Data sources description

In carrying out the prediction procedure, data of different nature are jointly used. In fact, the procedure combines administrative, traditional Census and sample survey data. For this exercise, the procedure is applied to a preliminary version of data for which the population in BRI considered as usually resident in Italy at 31st of December 2018 amounts to 60,433,360 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. Those core variables are obtained through an extensive utilisation of administrative data, reconciled and stored yearly in the BRI.

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 in the 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.

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 31st of October 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.

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 5% of the total population.

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[3] 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.

2.2 Reconciling classifications and computing ALE from administrative sources

Both CENS 2011 and BIT 2017 use detailed and reciprocally consistent classifications of educational level; consequently, data were univocally reclassified according to 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 2.1).

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

Table 2.1 - Correspondence table between Istat 2017 and CDIFF 2018 classifications on ALE

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 specialisation (IFTS)
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 degree) 13) *Laurea triennale* (I level, Bachelor's degree)
7 - Master's degree or equivalent level	10) Fine Arts, drama, Dance and Music Diploma (2-3 years) 14) Fine Arts, Drama, Dance and Music Second level academic diploma (Master's degree) 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

Source: Istat

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 2.2 - Consistency of data on ALE 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 aged >8 years	9,612,481	100.0

Source: Istat

Table 2.2 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.

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.

In order to reconcile the information, in the case of two different information on ALE coming from the two different sources, we replaced CENS 2011 data with BIT 2017 data. In fact, not only data that MIUR provided on yearly basis are usually reliable but also the process leading to the construction of BIT is a well-established one and is characterised by high quality standards (see Runci *et al.*, 2017).

2.3 Coverage and characteristics of subpopulation segments

An important aspect to analyse when using administrative data is the coverage of the sources, in fact they generally focus on specific populations. Table 2.3 classifies target population in main subgroups categorised 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 aged at least 9 years, the total coverage of administrative data is about 95%.

Table 2.3 - Reference population by presence in CENSUS 2011 and BIT 2017

	Total population		Aged 9 years and over	
	a.v.	%	a.v.	%
Present in BIT 2017	13,388,736	22.1	12,292,304	22.0
Present in CENS 2011 only	40,931,241	67.7	40,931,231	73.2
Records without information on ALE	6,113,383	10.1	2,685,623	4.8
Total BRI 2018 Population	60,433,360	100.0	55,909,158	100.0

Source: Istat

Despite 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 did not answered the questionnaire.

An in-depth analysis shows that these three groups of population - namely CENS 2011, BIT and people without any official administrative information on ALE – have slightly different distribution for what concerns principal core variables. Table 2.4 shows, in fact, that people without administrative and Census data on ALE are, on average, older. In more detail, individuals without information on ALE show higher percentages in the age classes 29 -39 years (29.9% as against 13.8% of total BRI population) and 40-49 years (21.6% as against 16.6% of total BRI population).

Table 2.4 - Age distribution in BRI by data sources: BIT 2017, CENSUS 2011, and records without information on ALE

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years	
Age	%	%	%	%	a.v.
9-10	9.0	0.0	1.6	2.0	1,145,028
10-11	13.6	0.0	1.7	3.1	1,720,250
14-18	23.0	0.1	1.3	5.2	2,889,161
19-22	18.1	0.1	4.6	4.3	2,385,385
23-25	11.8	0.6	4.9	3.3	1,820,519
26-28	8.8	1.7	6.6	3.5	1,943,501
29-39	12.0	13.3	29.9	13.8	7,732,149
40-49	2.4	20.5	21.6	16.6	9,264,211
50-69	1.4	39.0	22.6	29.9	16,725,020
70+	0.0	24.8	5.2	18.4	10,283,934
Total	100.0	100.0	100.0	100.0	55,909,158

Source: Istat

Moreover, in the subpopulation of individuals without any information on ALE there are more male than female (52.5% of male vs. 48.5% in BRI population) and mostly a dramatic larger percentage of Not Italian: 67.7% against 8.3% in the total BRI 2018 population (see respectively Table 2.5 and 2.6).

Table 2.5 - Gender distribution in BRI by data sources: CENSUS 2011, BIT 2017 and records without information on ALE

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years	
Gender	%	%	%	%	a.v.
Male	50.2	47.7	52.5	48.5	27,104,126
Female	49.8	52.3	47.5	51.5	28,805,032
Total	100.0	100.0	100.0	100.0	55,909,158

Source: Istat

Table 2.6 - Italian/Not Italian distribution in BRI by data sources: CENSUS 2011 and BIT 2017 and records without information on ALE

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years	
Citizenship	%	%	%	%	a.v.
Italian	94.1	94.8	32.6	91.7	51,252,687
Not Italian	5.9	5.2	67.4	8.3	4,656,471
Total	100.0	100.0	100.0	100.0	55,909,158

Source: Istat

2.4 Informative gaps and the use of auxiliary data

Data sources have some informative gaps. BIT, having MIUR data as exclusive source, reports only information for students that during the period 2012-2017 have enrolled a course that MIUR formally recognises. In particular, MIUR takes into account only courses supplied by an Italian qualified Institution on the Italian territory (*i.e.* International Institutions operating in the Country are not included). Furthermore, 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. The main consequence is an underestimation of the level of education, also for the units in the subset reporting 2011 Census ALE. In fact, the imputation procedure associates the potential lower ALE registered in CENS 2011 to those units that, during the period 2012-2017, had been enrolled in a couse not registered in BIT. As a consequence, for all population units for which schooling or training is over, it has to be experimented the use of auxiliary information.

Another critical issue concerning BIT has to do with timeliness. The lag between the moment in which BIT data are available and the BRI reference time makes it necessary to implement procedures for the prediction of the variable. BIT data are available with a delay of 12 to 24 months respect to the reference time. 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*].

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 2.7). 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. As Table 2.7 shows, information gathered by the sampling survey operation mostly overlap with ALE data coming from CENS 2011 (74.3%), though a 3.6% could help fillling the information gap for records lacking ALE.

Table 2.7 - Sample 2018 population by presence in CENSUS 2011 and BIT 2017

	Sample 2018		BRI 2018 Population aged at least 9 years	
	a. v.	%	a.v.	%
Present in BIT 2018	12,258,146	22.1	12,292,304	22.0
Present in CENS 2011 only	41,132,566	74.3	40,931,231	73.2
Records without information on ALE	1,977,346	3.6	2,685,623	4.8
Total	55,368,058	100.0	55,909,158	100.0

Source: Istat

The additional auxiliary administrative information on ALE from APR4 forms allows collecting data on about 5.2 million of units (Table 2.8). It is worth noticing that though APR4 data mostly overlap data from CENS 2011, 19.8% of observations covers the segment of population without any administrative information on ALE. This subpopulation is composed by either people more inclined to move across the Country, or entered Italy from abroad during the period 2012-2017. Thus, it is likely that the subpopulation without ALE is less “detectable” than the rest on the BRI 2018 units, and this can be the reason of the underestimation of the sample survey reported in Table 2.7.

Table 2.8 - APR4 form data by data sources in BRI: BIT 2017, CENSUS 2011 and records without information on ALE (row percentages and total absolute values)

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years
No APR4 data	22.2	74.6	3.3	50,702,985
With APR4 data	20.2	60.0	19.8	5,206,173
Total	22.0	73.2	4.8	55,909,158

To conclude, it is worth noticing that the nature of APR4 data is different both qualitatively and in substance, since it is self-declared information and never submitted to the standard editing/quality control procedures. However, a preliminary consistency analysis of information for individuals presenting data on attained ALE computed on administrative data (that is CENS 2011 updated with BIT 2017 data) and APR4 (4,175,256 observations) shows that APR4 presents a sufficient degree of consistency on the level of education 1 - Up to elementary license (83.9%) though it is decidedly lower for the other levels, with slightly higher percentages (67.7%) in 4 - Tertiary and Post Tertiary Education (see Table 2.9).

Table 2.9 - ALE in 2017 (administrative data) and in APR4 (row percentages and total absolute values)

	ALE in APR4 (2012-2017)				
	1 - Up to Primary education	2 - Lower secondary education	3 - Secondary and short cycle tertiary education	4 - Tertiary and post tertiary education	Total a.v.
1 - Up to Primary education	83.9	11.6	3.6	0.9	599,765
2 - Lower secondary education	32.8	52.8	12.5	1.8	1,281,587
3 - Secondary and short cycle tertiary education	20.2	15.3	57.2	7.3	1,536,681
4 - Tertiary and post tertiary education	16.4	3.9	12.0	67.7	757,223
Total	32.5	24.2	27.6	15.7	4,175,256

Source: Istat

3. Imputation of the attained level of education

3.1 The imputation procedure

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 2.

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

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

As aforementioned, 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^tS the ALE at time t observed in the sample survey.

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 3.1 depicts the data/information scenario that we need to take into account when making predictions for I^t. Grey cells represent missing data and the last column shows the relative frequencies of groups with respect to the population with at least 9 years.

Table 3.1 - 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%
				▨	▨	▨		▨	▨	
				▨	▨	▨	▨		▨	
				▨	▨	▨	▨	▨	▨	

Source: Istat

3.2 Mass imputation process flow

Groups A, B and C are characterised by different patterns of available information which determine different models for the estimation of ALE in 2018. In particular, ALE estimation may be either deterministic or probabilistic. The overall process of data treatment, model estimation and ALE imputation in the three groups A, B and C is summarised in Figure 3.1.

The main difference is between group A and the others. Group A is composed by “Active” people, which are attending a course in academic year t-12/t, while groups B and C are “Inactive” people, not attending any course in the same period, which is the last available from administrative sources.

In group A, administrative data provide longitudinal information on school enrollment. Thanks to the great information capacity of these administrative data, it is not necessary to resort to ALE observed in the 2018 sample. Information on ALE in the year t-12 (I^{t-12}) and information on year attendance of educational courses in academic year t-12/t (F^{t}) are available for all the individuals in group A. This allows identifying the probability of obtaining a new qualification based on schooling charachteristics of each individual.

Out of them, a subset of individuals with a zero probability of changing the educational level, from t-12 to t, is identified. Therefore, for this subset of "No-Change" people, it is not necessary to estimate a model for the imputation of ALE, since ALE in 2018 (I^{t}) is equal to ALE in 2017 (I^{t-12}).

The subset of "No-Change" is identified by one of the following conditions:

1. attending year 1, 2, 3 or 4 of primary school (Primary education is acquired at the end of year 5);
2. attending year 1 or 2 of lower secondary school (Lower secondary education is acquired at the end of year 3);
3. attending year 1, 2 or 3 of upper secondary school (Upper secondary education is acquired at the end of year 5; in high school you can attend two years in one);
4. attending upper secondary school and still having an Upper secondary education;
5. enrolled in a first level university course and still having a Bachelor's or a Master's degree;
6. enrolled in a university course and still having a Master's degree;
7. attending year 1 of a PhD course or still having a PhD.

People of group A, who do not meet any of the above conditions, have a non-zero probability of obtaining a higher qualification than that held in year t-12. For each individual of this "Change" subset the estimate of the probability distribution of achieving a new qualification in time t is based on individual characteristics and school attendance in academic year t-12/t (F^{t}). The model is estimated using only administrative sources. The underlying hypothesis is that the probability of obtaining a higher qualification between the years t-12 and t is equal to that between the years t-24 and t-12.

On the other side, group B and C are composed by “Inactive” people, this means people not enrolled in any course covered by MIUR in academic year t-12/t. Due to some informative gaps in administrative sources (see Section 2.3), there is a non-zero probability that an individual belonging to these groups is either enrolled in academic year t-12/t or has been enrolled in previous academic years in a school course not covered by MIUR.

For people in group B, information on previous educational level is available from administrative sources or from data collected in the 2011 Census.

For people interviewed in the 2011 Census who was enrolled in a school course covered by MIUR between 2011 and 2016 (but not in 2017/2018), the most updated information on ALE comes from MIUR. This subgroup is composed of individuals on average younger, who have recently dropped out of a school course covered by MIUR.

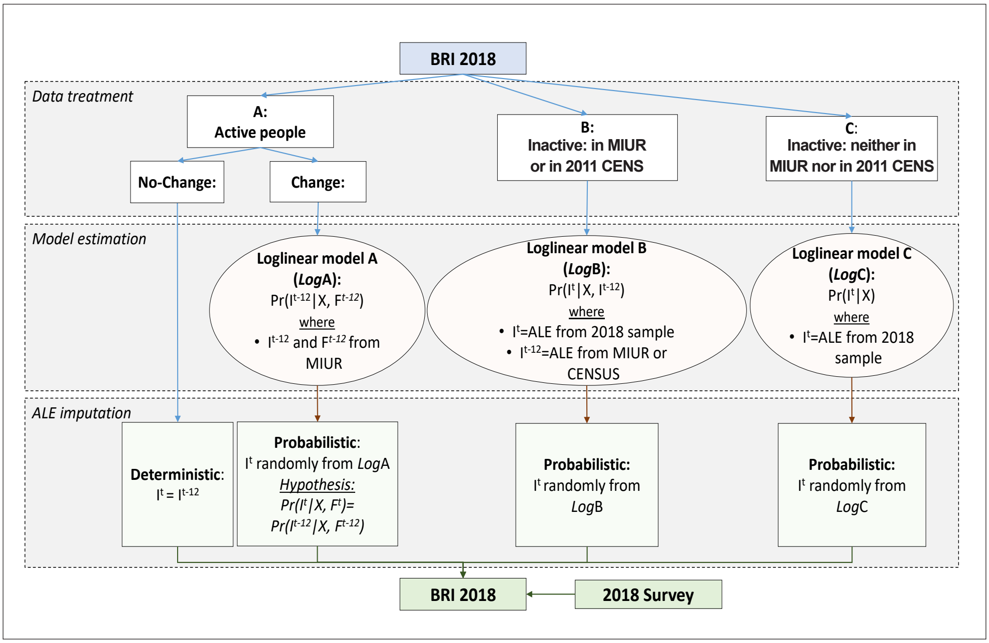
On the other hand, for people not enrolled in any school course after 2011, the only available information on ALE refers to 2011. They are mainly adults, long since out of school and probably less likely to change their educational level.

In both cases, the available information on ALE may not be error free due to coverage error (MIUR) or response error (2011 Census). For this reason, the model is estimated on units interviewed in the 2018 survey using the observed ALE as target variable. However, due to their different characteristics, individuals with information on ALE from MIUR or from 2011 Census are treated separately in the estimation process.

For people in group C, no information on ALE is available neither from MIUR nor from 2011 Census, so it is necessary to estimate a probability distribution of ALE for each pattern of available information on individual characteristics. ALE observed in 2018 survey is considered as target variable.

As a last step of imputation, for all the individuals observed in the 2018 sample, the observed ALE is directly used as prediction in BRI.

Figure 3.1 - Imputation process flowchart


BRI 2018
Data treatment
A:
Active people
B:
Inactive: in MIUR or in 2011 CENS
C:
Inactive: neither in MIUR nor in 2011 CENS
No-Change:
Change:
Model estimation
Loglinear model A (*LogA*): $Pr(I^{t-12}|X, F^{t-12})$ where • I^{t-12} and F^{t-12} from MIUR
Loglinear model B (*LogB*): $Pr(I^t|X, I^{t-12})$ where • I^t=ALE from 2018 sample • I^{t-12}=ALE from MIUR or CENSUS
Loglinear model C (*LogC*): $Pr(I^t|X)$ where • I^t=ALE from 2018 sample
ALE imputation
Deterministic: $I^t = I^{t-12}$
Probabilistic: I^t randomly from *LogA* *Hypothesis:* $Pr(I^t|X, F^t)= Pr(I^{t-12}|X, F^{t-12})$
Probabilistic: I^t randomly from *LogB*
Probabilistic: I^t randomly from *LogC*
BRI 2018
2018 Survey

Source: Istat

3.3 Model estimation and imputation

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 randomly taking a value from this distribution. The conditional probabilities $h(I^t|X)$ are estimated by means of hierarchical 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 $\hat{h}(I^t \mid X)$ is easily obtained by computing $\widehat{N}(I^t, X)/(X)$. This approach includes as a special case the random hot-deck when a saturated log-linear model is assumed, but it has the advantage of allowing the use of more parsimonious model as well. This is an important characteristic especially when the number of variables and of the contingency table cells increase.

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).

Similarly to hot-deck, it may happen that a missing observation is not imputed because its 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.

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

Different log-linear models are used within groups A, B and C, mainly because of the different available information. As already remarked, 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 are given for each group.

Imputation in Group A - Change.

This group is characterised 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].

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 instead sample survey data.

In order to ease referring to models, we adopt the classic notation for hierarchical log-linear models where only the highest-order interaction term for each variable is reported (see Agresti, 2002, pp. 320).

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

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

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).

Imputation in Group B.

People not attending any course covered by MIUR in *t-1/t* characterise this group. These are people that either have decided to stop their studies or that are 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

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

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).

Imputation in Group C.

This group is characterised 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 since 2011 (D=2).

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*.

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.

A general remark is concerned with the use of other potential covariates like income and type of occupation. Unfortunately, the type of occupation is not available in the due time to be introduced in the modelling. As far as income is concerned, we notice that it is not available for the whole population and it refers to time t-2. Nevertheless in Di Cecco *et al.*, 2018, where a model for predicions in two years (from *t*-2 to *t*) was studied, income resuled as an explicative covariate. However, with the data at hand and by considering the procedure so far illustrated, the introduction of income in the model resulted in strange results in the aggregates and, also by considering the difficulty in the construction of this information in the current Istat production system, we have opted for excluding it from the current procedure. It is indeed true that, once information on income will be more timely and stable in the Istat production system, additional analysis should be performed in order to take into account the possibility of using such information.

4. Analysis for the assessment of the predictions

In this section, we illustrate the results of some analysis carried out in order to assess the quality of the procedure. Analysis on micro-data and aggregates are performed. Results computed on BRI are analysed and compared with the ones computed on data collected in the sample of the 2018 permanent census, and with the ones computed on data from administrative sources and 2011 Census where available. As far as micro level analysis is concerned, 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.

Table 4.1 shows the number of imputed values for each imputation step. For the individuals interviewed in the 2018 sample the imputation is deterministic since the predition coincides with the observed value. Excluding the “A-No change” group, which represents the 6.7% of the population, almost all the imputations (88.4%) occur in step 1.

Table 4.1 - Distribution of imputation steps - absolute value (a.v.) **in thousands and percentage values** (%)

Imputation step	a.v.	%
Group A		
A - No change	3,762	6.7
A - Change - step 1	3,581	6.4
A - Change - step 2	4	0.0
A - Change - deterministic: estimated = admin.	9	0.0
Group B		
B - step 1	43,275	77.4
B - step 2	53	0.1
B - step 3	7	0.0
B - step 4	5	0.0
B - step 5	1	0.0
Group C		
C - step 1	2,574	4.6
C - step 2	37	0.1
C - step 3	4	0.0
2018 Sample	2,597	4.6
Total	55,909	100.0

Source: Istat

For groups A and B, transitions between observed and estimated ALE provide a first evaluation of the procedure. In group A, the estimated 2018 ALE is in most cases consistent with the 2017 information from administrative sources (Table 4.2). This happens when the estimated 2018 ALE confirms the 2017 ALE or increases it by one 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 3).

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.

Table 4.2 - 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	65.8	33.5	0.6	0.0	0.0	-	1,791
4	Lower secondary education	0.0	0.0	0.0	77.1	22.6	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

Source: Istat

In group B, the estimated 2018 ALE shows some inconsistencies with ALE in 2017 (Table 4.3). 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 4.3 - 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	8.8	2.5	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	7,298
4	Lower secondary education	0.1	0.6	5.8	78.9	13.9	0.2	0.5	0.0	12,937
5	Upper secondary education	0.1	0.2	1.0	6.5	89.3	1.2	1.6	0.0	14,05
6	Bachelor’s degree	0.0	0.2	0.6	3.1	16.8	61.9	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	0.5	1.0	2.1	0.7	27.3	68.4	136
	Total	0.7	2.6	17.7	30.0	36.4	2.1	10.1	0.4	40,931

Source: Istat

In order 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 values of the differences between percentage of each item, in absolute (*AD*) and relative (*RD*) terms. Specifically:

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

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

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

The macro level comparison between BRI and sample estimates shows that the two distributions are very similar (Table 4.4). 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 4.4 - 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 (a)	
ALE 2018		a.v.	%	a.v.	%	D_i	Dr_i
1	Illiterate	353	0.6	330	0.6	0.03	5.83
2	Literate but no ed. Attainment	2,295	4.1	2,073	3.7	0.36	9.65
3	Primary education	9,293	16.6	9,137	16.5	0.12	0.72
4	Lower secondary education	16,509	29.5	16,168	29.2	0.33	1.12
5	Upper secondary education	19,718	35.3	19,873	35.9	-0.63	-1.74
6	Bachelor's degree	1,977	3.5	1,962	3.5	-0.01	-0.16
7	Master's degree	5,531	9.9	5,598	10.1	-0.22	-2.15
8	PhD	233	0.4	227	0.4	0.01	1.67
Total		55,909	100.0	55,368	100.0	*AD*=0.21	*RD*=2.88

Source: Istat
(a) The calculations from the table may give different numbers due to the approximation.

The comparison between target and estimated distributions for groups A, B and C shows different behaviours. In particular, in group C the estimated distribution differs from that of the sample, more than it differs in the A and B groups. This is due to the lower quantity and quality of available information (Table 4.5). On the contrary, in group B the estimated and sampled ALE distributions are almost perfectly equivalent. This is mainly related to the greater number of the individuals in group B, in addition to the fact that the information on ALE from the sample is used as response variables in the model (this also applies for group C, but not for group A). It is worthwhile to notice that when class 1 and 2 are jointly considered (see Table 4.4) the difference is not high, in fact these two modalities are generally hardly to discriminate and most of the times their counts are jointly provided in the published tables. A remarkable difference is also in class 3 (Upper secondary education), we notice that this is observed both in A and C. Further analysis, jointly performed with subject matter experts, are still in progress to understand the reasons behind those differences.

Table 4.5 - Model and sample estimates of 2018 ALE (percentage values) and absolute differences (D) between model and sample percentages in the three groups A, B and C

		Model			Sample			Model – Sample		
		A	B	C	A	B	C	A	B	C
ALE 2018		%	%	%	%	%	%	$D_i^{(*)}$	$D_i^{(*)}$	$D_i^{(*)}$
1	Illiterate	0.0	0.7	1.9	0.0	0.7	1.7	-0.0	0.0	0.2
2	Literate but no ed. Attainment	8.7	2.6	6.4	7.5	2.5	5.4	1.2	0.0	1.0
3	Primary education	14.1	17.7	11.7	14.3	17.4	10.7	-0.2	0.3	1.0
4	Lower secondary education	27.3	30.0	32.0	26.8	29.7	33.2	0.4	0.3	-1.2
5	Upper secondary education	32.0	36.1	33.1	33.5	36.4	34.4	-1.5	-0.3	-1.3
6	Bachelor’s degree	8.1	2.2	3.5	8.1	2.2	3.4	0.1	-0.0	0.2
7	Master’s degree	9.2	10.4	10.6	9.2	10.7	10.6	-0.0	-0.3	0.0
8	PhD	0.6	0.4	0.6	0.5	0.4	0.5	0.0	-0.0	0.1
Total		100.0	100.0	100.0	100.0	100.0	100.0	AD=0.43	*AD*=0.16	*AD*=0.62

Source: Istat

The distribution of ALE 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 4.6) in which we observe an average difference of 0.39 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 characterised by particular features and less information available determining a different fit of the model.

Table 4.6 - 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 (a)	
	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	270	0.5	83	1.8	258	0.5	73	1.6	0.02	0.14
2 Literate but no ed. attainment	1,976	3.9	319	6.9	1,8	3.5	273	6.2	0.32	0.69
3 Primary education	8,787	17.1	506	10.9	8,658	17.0	479	10.8	0.15	0.06
4 Lower secondary education	14,968	29.2	1,541	33.1	14,686	28.8	1,482	33.4	0.37	-0.35
5 Upper secondary education	18,049	35.2	1,668	35.8	18,232	35.8	1,641	37.0	-0.58	-1.20
6 Bachelor’s degree	1,812	3.5	166	3.6	1,812	3.6	149	3.4	-0.02	0.19
7 Master’s degree	5,175	10.1	356	7.6	5,278	10.4	320	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	51,252	100.0	4,656	100.0	50,936	100.0	4,431	100.0	*AD*=0.22	*AD*=0.39

Source: Istat
(a) Warning: the calculations from the table may give different numbers due to the approximation.

Looking at the territorial level, Table 4.7 shows the differences between estimated and observed percentages of each item (D) and the mean of absolute differences of each item (AD) by region. Even if in general there is a small variability between regions, it can be seen that northern regions have lower differences between estimated and observed distributions (see Piemonte, Valle d'Aosta/Vallée d'Aoste, Lombardia, Friuli-Venezia Giulia and Emilia-Romagna), vice versa in the southern regions and islands (Puglia, Calabria, Sicilia and Sardegna). It is worth noting that item 1 (“Illiterate”) and 2 (“Literate but no educational attainment”) are over-estimated in all regions while item 5 (“Upper secondary education”) is always under-estimated. Further analyses are needed to understand the reasons.

Table 4.7 - Item absolute differences (Di) and mean of absolute differences (AD) between model and sample percentages by region

Regions	Illiterate	Literate but no att.	Primary ed.	Lower sec. ed.	Upper sec. ed.	Bachelor’s degree	Master’s degree	PhD	Mean
	(D_1)	(D_2)	(D_3)	(D_4)	(D_5)	(D_6)	(D_7)	(D_8)	(AD)
Piemonte	0.04	0.35	0.15	0.22	-0.54	-0.09	-0.13	0.01	0.19
Valle d'Aosta/*Vallée d'Aoste*	0.03	0.18	-0.08	0.43	-0.33	-0.13	-0.14	0.03	0.17
Lombardia	0.03	0.37	-0.08	0.06	-0.4	0.02	-0.01	0.01	0.12
Trentino-Alto Adige/*Südtirol*	0.02	0.22	-0.3	-0.3	-0.31	0.04	0.58	0.04	0.23
Bolzano/Bozen	*0.01*	*0.11*	*-0.58*	*0.39*	*-0.27*	*-0.06*	*0.37*	*0.01*	*0.23*
Trento	*0.04*	*0.33*	*-0.02*	*-0.98*	*-0.34*	*0.13*	*0.78*	*0.07*	*0.33*
Veneto	0.03	0.34	0.11	0.34	-0.7	0.01	-0.13	0	0.21
Friuli-Venezia Giulia	0.05	0.14	0.09	0.18	-0.6	0.08	0.07	-0.01	0.15
Liguria	0.01	0.29	0.32	0.77	-0.73	-0.07	-0.56	-0.02	0.35
Emilia-Romagna	0.02	0.33	-0.1	0.15	-0.41	-0.01	-0.03	0.03	0.14
Toscana	0.03	0.35	0.16	0.47	-0.65	0.01	-0.37	0	0.26
Umbria	0.05	0.43	0.22	0.49	-0.77	0.01	-0.41	-0.02	0.3
Marche	0.02	0.25	0.15	0.58	-0.83	0	-0.18	0.01	0.25
Lazio	0.03	0.44	0.14	0.68	-0.73	-0.07	-0.44	-0.03	0.32
Abruzzo	0.03	0.44	0.19	0.54	-0.66	-0.1	-0.45	0	0.3
Molise	0.04	0.19	-0.35	0.42	-0.13	0.08	-0.23	-0.01	0.18
Campania	0.03	0.45	-0.02	0.05	-0.54	0.07	-0.06	0.03	0.16
Puglia	0.08	0.38	0.3	0.67	-0.91	0	-0.52	0.01	0.36
Basilicata	0.1	0.4	0.05	-0.41	-0.64	0.05	0.43	0.02	0.26
Calabria	0.03	0.32	0.69	0.56	-1.37	-0.02	-0.24	0.03	0.41
Sicilia	0.05	0.37	0.43	0.46	-0.76	-0.04	-0.53	0.03	0.33
Sardegna	0.06	0.3	0.25	0.58	-0.6	0.04	-0.61	-0.03	0.31
Italy	0.04	0.37	0.12	0.33	-0.63	-0.01	-0.22	0.01	0.21

Source: Istat

Most of the analyses illustrated so far are concerned with aggregates, that are the first main goal of the procedure. In fact, the decision to impute with a random draw from the estimated conditional distribution is aimed at increasing the preservation of distributions, while unfortunately decreasing the predictive accuracy (at micro-level) of the model.

Nevertheless, it is interesting to look at the predictive accuracy of the model, since data are predicted at micro level in BRI. In Table 4.8, we report the differences at micro-level computed comparing the imputed ALE vs the ALE observed in the sample survey. Out of the whole sample, 74% of units are exactly predicted. As expected, the best predictive accuracy is in set A (88%) that is in fact the subset of data with the highest amount of administrative information. It is worthwhile to remind that in this subset, the model is estimated by using only administrative data, and this makes the result even more interesting. On the other side, we notice the poor performance in terms of micro-predictions of the model in the set C. This was expected as well, since C is characterised by a very low level of auxiliary information, but it fortunately refers to a small part of the total population (2.8% of the data used for the comparison).

Table 4.8 - Differences at micro level between estimated and observed ALE in the sample survey. DIF is equal to 1 when values are different

	Group			
	A	B	C	Total
DIF	%	%	%	%
0	87.9	71.9	34.0	74.3
1	12.1	28.1	66.0	25.7
Total	100.0	100.0	100.0	100.0

Source: Istat

5. Final remarks and future developments

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.

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".

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.

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.

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.

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

Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, NJ, U.S.: John Wiley & Sons.

Chen, S., D. Haziza, C. Léger, and Z. Mashreghi. 2019. "Pseudo-population bootstrap methods for imputed survey data". *Biometrika*, Volume 106, Issue 2: 369-384.

Daalmans, J. 2017. "Mass imputation for Census estimation". In United Nations Economic Commission for Europe – UNECE, *Conference of European Statisticians, Group of Experts on Population and Housing Censuses*. 19th Meeting, Geneva, Switzerland, 4th - 6th October 2017.

de Waal, T. 2016. "Obtaining numerically consistent estimates from a mix of administrative data and surveys". *Statistical Journal of the IAOS*, Volume 32, N. 2: 231-243.

Di Consiglio, L., M. Di Zio, and D. Filipponi. 2019. "An empirical evaluation of latent class models for multisource statistics". *Presentation at ITACOSM 2019*. Firenze, Italy, 5th -7th June 2019.

Di Cecco, D., D. Di Laurea, M. Di Zio, R. Filippini, P. Massoli, and G. Rocchetti. 2018. "Mass imputation of the attained level of education in the Italian System of Registers". In United Nations Economic Commission for Europe – UNECE, *Workshop on Statistical Data Editing*. Neuchâtel, Switzerland, 18th -20th September 2018.

Runci, M.C., G. Di Bella, and F. Cuppone. 2017. "Integrated Education Microdata to Support Statistics Production". In Lauro, N.C., E. Amaturo, M.G. Grassia, B. Aragona, and M. Marino (*eds*.). *Data Science and Social Research. Epistemology, Methods, Technology and Applications*. Heidelberg, Germany: Springer International Publishing, *Studies in Classification, Data Analysis, and Knowledge Organization*.

Scholtus, S., and J. Pannekoek. 2015. "Mass-imputation of educational levels". *Internal report* (available in Dutch). The Hague and Heerlen, The Netherlands: Statistics Netherlands – CBS.

Scholtus, S. 2018. “Variances of Census Tables after Mass Imputation”. *Discussion paper*. The Hague and Heerlen, The Netherlands: Statistics Netherlands – CBS.

Skinner, C., and L.-A. Vallet. 2010. “Fitting log-linear models to contingency tables from surveys with complex sampling designs: an investigation of the Clogg-Eliason approach”. *Sociological Methods & Research*, Volume 39, Issue 1: 83-108.

Thibaudeau, Y., E. Slud, and A. Gottschalck. 2017. “Modeling log-linear conditional probabilities for estimation in surveys”. *The Annals of Applied Statistics*. Volume 11, Issue 2: 680-697.