Imputation of the variable "Attained Level of Education" in Base Register of Individuals

Organisation: ISTAT

Author(s): Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto

Date: 27/02/2020

Version: 1

1. Background and why and how this study was initiated

The Italian production system of statistics is deeply changing, moving towards a register based statistics production. To this extent, the new Italian Census (Permanent Census) will be as much as possible register-based.

Among others, Census gathers information on the Attained Level of Education (ALE). The high amount of available information on this topic, in particular administrative longitudinal information, may allow the production of statistics from register rather than from survey. The aim is to insert the variable ALE in the set of core information in the Base register of individuals (BRI), which represents the widest possible set of individuals on which to make a selection for the identification of resident population.

In particular the Italian National Institute of Statistics is interested in a micro level estimation of the ALE (8 classes) for Italian resident population in October 2018. To this aim, a working group has been working on the prediction/mass-imputation of ALE in BRI [1]. In the specific case, Log-linear models are studied [2]. Due to the complexity and heterogeneity of the available information, in order to carry out an accurate ALE prediction, an in-depth knowledge of data structure is needed and different steps of imputation have to be performed.

The experimentation carried out among the HLG-MOS group is initiated with the aim to try a different solution that could be able to solve the problem in a more automated way and to improve the results. ML techniques are the methods studied to this aim.

To perform the experimentation of interest a collaboration within ISTAT has born. In particular, ML experts and statisticians involved in the estimation of ALE have started a fruitful collaboration.

2. Data

2.1 Input Data

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.

In carrying out the ALE prediction procedure, data of different nature are jointly used: administrative data, traditional Census data and sample survey data.

Administrative data: administrative information on ALE is gathered making use of the information collected by the Ministry of Education, University and Research (MIUR). MIUR provides information about ALE and course attendance for people entering a study program after 2011 and covers the period from 2011 to 2017 (scholar year 2017/2018).

Traditional Census data (2011 Census): for people that have not attended any course since 2011 we turn to data from 2011 Census to fill the gap.

Sample survey data: direct measurement for ALE in 2018 is available only for a subset of population (about 5%), coming from the first Permanent Census Survey that took place in Italy in October 2018 (CS2018).

The reference population is characterised by different patterns of variables.

The structure of available information is summarized in table 1. Blue cells indicate that the information is available for the specific subpopulation.

Table 1: Structure of the dataset

Source:	BRI	MIUR	2011 Census	CS 2018		
Available inf.:	Core inf.	ALE 2017	ALE 2017	ALE 2018	**Subpopulation**	Used in the Case study
Coverage	■	■		■	**A**	**Yes**
	■	■			**A**	No
	■		■	■	**B**	**Yes**
	■		■		**B**	No
	■			■	**C**	**Yes**
	■				**C**	No

Core information from BRI are available for all individuals. This information are age, gender, citizenship, marital status, place of birth and place of residence.

The different availability of information on ALE from 2011 to 2017, determines the partition of our population of interest into three subgroups:

A. All persons for whom information on ALE is available from MIUR belong to subgroup A;
B. Persons not in MIUR who were interviewed in the 2011 Census belong to subgroup B. This means that subgroup B is made up of individuals for whom the only information on ALE comes from the 2011 Census[1];
C. Individuals neither in MIUR nor in 2011 Census belong to group C. For this group no information on ALE is available.

[1] If the MIIUR was not affected by under-coverage, the fact that an individual was not present in the MIUR would mean that he never attended a school course from 2011 to 2017 and that the ALE in 2011 did not change in the following years.

The classification adopted for ALE is composed by 8 items: 1 – Illiterate, 2 - Literate but no formal educational attainment, 3 - Primary education, 4 - Lower secondary education, 5 - Upper secondary education, 6 - Bachelor's degree or equivalent level, 7 - Master's degree or equivalent level, 8 - PhD level.

2.2 Data Preparation

To compare ML and standard model imputation approach we use a reduced dataset, considering only one Italian region (Lombardia) and the subset of population for which the target variable is available (see last column of table 1). The target variable is the self-declared ALE in the 2018 sample census, referring to the year 2018 and which covers about the 5% of total population of interest.

People with complete longitudinal information on course attendance from administrative sources (part of subpopulation A) are excluded from this experimentation since the knowledge of the schooling history, until scholar year 2017/2018, allows to forecast ALE 2018 with great accuracy. For this part of population, we resort to a different approach, not using ALE 2018 as target variable[2].

The dataset for the experimentation consists of 312.813 individuals resident in the Lombardia region in 2018 with no missing data on ALE 2018 (target variable). This is the sum of sub-populations A-Yes, B-Yes and C-Yes in table 1.

2.3 Feature Selection

Different patterns of available variables may induce the selection of different models. For each subpopulation (A, B and C), the best Log-linear model is chosen by means of cross-validation. In particular, Log-linear models for each sub-population are:

Subpopulation A: Pr (ALE2018| ALE2017, age, citizenship, school attendance)

Subpopulation B: Pr (ALE2018| ALE2017, age, citizenship, province of residence, gender)

Subpopulation C: Pr (ALE2018| age, citizenship, gender, apr[3], sirea[4]).

[2] For this part of population we estimate the probability of changing ALE from 2016 (t-2) to 2017 (t-1) given school attendance in 2016/2017, and we use the same probability to forecast ALE 2018, given school attendance in 2017/2018.

[3] Apr is an auxiliary information on ALE coming from an administrative source, which covers a particular subpopulation of individuals: those who changed their place of residence after 2014. Moreover, this information is more aggregate (4 items) and not so accurate. We decide to use ALE from apr source only in subpopulation C, where we do not have any other information on ALE.

[4] People not caught by the 2011 Census. During post-Census operations, the collaboration with municipalities (named SIREA operation) allowed to identify individuals not found in the 2011 Census but that were resident:

All the selected covariates enter the dataset used for the experimentation.

2.4 Output data

We perform the final ML procedure on a dataset containing 312.813 records (individuals) and 11 variables (Table 2).

Table 2: Variables in the dataset

id	NAME	DESCRIPTION
1	COD_IND	Record id
2	GENDER	Gender
3	AGE	Age classified into 14 levels[5]
4	PROV	Province of residence
5	CIT	Citizenship (Italian/Not Italian)
6	ABC_2017	Subpopulation (A, B C)
7	APR	ALE from APR classified into 4 levels[6]
8	ALE2017	2017 ALE (combination of Administrative and 2011 Census)
9	FR18	School attendance in 2017/2018
10	SIREA	Resident in Italy in 2011 not caught by the 2011 Census
11	**ALE_CS18**	2018 ALE from 2018 Census Survey (**Target variable**)

In addition, the dataset contains the results of the imputation using the Log-linear models.

The Log-linear solution uses different models for the three subpopulations combining differently the variables in the dataset; the ML solution uses together all the covariates in the dataset.

3. Machine Learning Solution

3.1 Models tried

In recent years in ISTAT we have gained a lot of experience on the use of neural networks for the treatment of big data, in particular the deep-learning techniques with the convolutional networks (CNN) [3] [4] for the treatment of images and natural language. With this background, we initially used the Multi Layer Perceptron (MLP) as the machine learning algorithm for this experimentation.

they were “detected” with the purpose of counting resident population but they did not answered the questionnaire.

[5] Age levels are identified taking into account the structure of the Italian school system (0-8; 9-10; 11-11; 12-13; 14-17; 18; 19; 20-22; 23-25; 26-28; 29-39; 40-49; 50-69; 70-max)

[6] APR (registration and cancellation forms for transfer of residence) is an auxiliary administrative source containing information on ALE. It is a self-declared information with a low level of quality and too much aggregated classification (ALE comes with 4 levels of classification: 1- Up to primary education; 2 - Lower secondary education; 3 - Secondary and short cycle tertiary education; 4 - Tertiary and post tertiary education).

Some first experimentation using Random Forest (RF) has been done, we do not report these results in this document because they are in a very preliminary stage, however, the model RF seems to be very promising for the imputation of qualitative variable ALE.

3.2 Model finally selected and the criterion

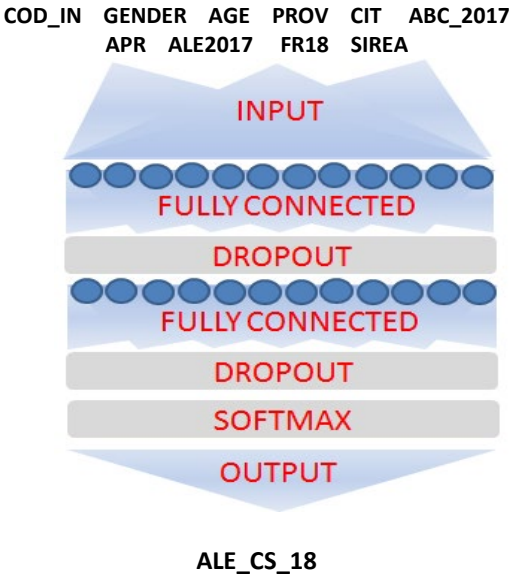
In particular, we use the MLP as neural network architecture, this choice is due to the ability of the MLP to find, after a training phase, a good approximation of the relationship between the input variables and the distribution of the output variable [5].

We use in our comparative study the same categorical input variables used in the Log-linear models. Our approach aims to be as general as possible, therefore:

- We train a single neural network, unlike the standard approach, where different models are built, according to the variables available for each profile.
- We encode the input variables of the perceptron multilayer as dummy/one-hot encoding, in this representation the missing value of a variable is encoded like any other mode of the variable.

We encode the input variables of the perceptron multilayer with the aim of minimizing the cross-entropy loss function. The cross-entropy is a measure of the distance between the distribution of the output variable and the distribution of the target variable. The architecture of the network is shown in figure 1 and has two hidden layers each of 128 neurons, an output layer with 8 neurons (one per modality of the target variable). To limit the over-fitting in the learning phase, two layers of dropout have been interposed. The best configuration of some hyper-parameters (number of hidden neurons, dropout probability, learning-rate) was explored through a suitable grid-search.

Figure 1. Architecture of the model implemented



For each record of the dataset, the model generates a probability distribution on the 8 ALE items. In a conventional ML approach, the imputed value is the modal value of the distribution. However, in our case study, an important goal is to reproduce the distribution of ALE in the population of interest. To increase the distributional accuracy, for each record we impute the ALE item randomly extracted from the probability distribution of the correspondent pattern as in the Log-linear models.

3.3 Hardware used

For our case study, we use a Linux server, Ubuntu 16.04.5 LTS distribution on the Azure cloud platform with Tesla V100-PCIE-16GB GPU. The GPU is not strictly necessary but reduces the runtime to train the model.

3.4 Runtime to train the model

We spend about an hour to train our MLP model.

The runtime depends from several aspects:

- The model complexity, in particular our model has about 27000 parameters (the neural network weights)
- The training set dimension (250250 samples)
- The number of the iterations of the optimization algorithm (500 epochs).

4. Results

4.1 Micro level accuracy

Model accuracy is calculated using the k-fold approach, k=5. The database is partitioned into 5 subgroups,

(1) the model is estimated on the training set, consisting of 4 of the 5 subgroups,

(2) the results are applied on the test set, composed of the remaining subgroup,

(3) accuracy is calculated only on the test set as the difference between estimated ALE 2018 and the observed ALE 2018.

Operations 1-3 are repeated 5 times so to reconstruct the entire data set. The same approach is used for both ML and standard Log-linear models, so results can be compared.

Micro level accuracy of imputed ALE 2018 using MLP[7] approach is very similar to those originated from Log-Linear models: 72,1% vs 72,0% - variance of results is in both cases negligible (table 3).

Table 3. Micro-level accuracy in the 5 test sets: Log-linear vs MLP estimation

test	Log-Linear	MLP
1	72.14%	72.05%
2	72.12%	71.99%
3	72.18%	72.24%
4	72.05%	72.08%
5	72.07%	71.90%
MEAN	**72.11%**	**72.05%**

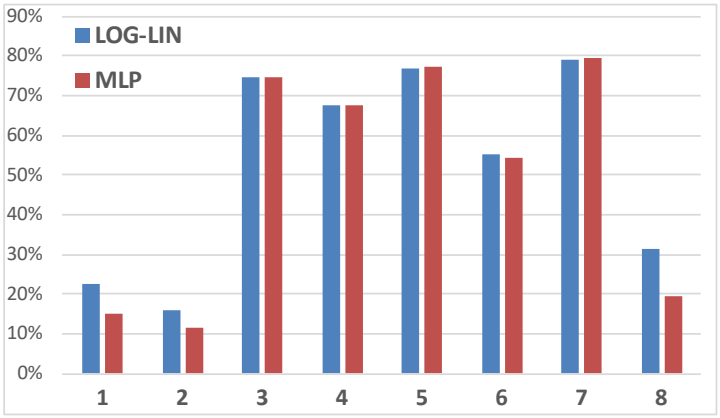
Micro-level accuracy can be calculated for each item; specifically, item accuracy is:

$$\frac{\#(\text{observed ALE}=i \text{ and estimated ALE}=i)}{\#(\text{observed ALE}=i)}$$

It is interesting to note that both, Log-linear and MLP, give origin to high accuracy in central items (3, 4, 5 and 7) which are also the most frequent, and lower accuracy in extreme items (1, 2 and 8) (figure 2). In particular, even if with low accuracy, the extreme items are better estimated by the Log-linear model than MLP.

[7] Note that instead of using the modal value as in the standard MLP approach, we made a random extraction of ALE value from the estimated distribution. This reduces micro accuracy (that would otherwise be around 80%) but improves distributional accuracy, which is our main goal.

Figure 2. Item accuracy: Log-linear vs MLP estimation



1 Illiterate
2 Literate but no ed. attainment
3 Primary education
4 Lower secondary education
5 Upper secondary education
6 Bachelor's degree
7 Master's degree
8 PhD

4.3 Macro level accuracy

To evaluate the imputation procedure in a macro level approach, the estimated ALE in 2018 ($\widehat{ALE18}$), is compared with the data collected in the 2018 census sample (ALE_CS18). In particular, we focus on the differences between the frequency distributions of estimated 2018 ALE and the distribution observed on the sample. 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:

$$AD = \frac{1}{K}\sum_{c=1}^{K}|D_c| = \frac{1}{K}\sum_{c=1}^{K}\frac{\left|\sum_i \hat{T}_{ic} - \sum_i T_{ic}\right|}{N}$$

$$RD = \frac{1}{K}\sum_{c=1}^{K}|Drel_c| = \frac{1}{K}\sum_{c=1}^{K}\frac{\left|\sum_i \hat{T}_{ic} - \sum_i T_{ic}\right|}{\sum_i T_{ic}}\ 100$$

where T_{ic} is the target variable (ALE_CS18) for the observation i, c is the item index in a dummy representation, K is the number of the items and N is the total number of observations. Similarly $\hat{T}_{ic}$ is the predicted target variable.

Macro level accuracy of imputed ALE 2018 using the MLP approach is slightly lower to those originated from the Log-Linear models (table 4). The estimations differ from the observed data by 0.08% and 0.09% points on average on each item respectively for the Log-linear and MLP approach. In relative terms the MLP approach performs a little worse: the average relative differences are 2.28% and 3.22% respectively for Log-linear and MLP. This means that inaccurate estimates are more concentrated in less

numerous classes such as extreme ALE items. This is valid in general but particularly for the MLP.

Table 4. Macro-level accuracy, in absolute (AD) and relative (RD) terms, in the 5 test sets: Log-linear vs MLP estimation

test	LOG-LIN - TARGET		MLP – TARGET	
	AD	RD	AD	RD
0	0.066%	3.096%	0.092%	2.499%
1	0.068%	1.154%	0.081%	4.118%
2	0.064%	1.665%	0.086%	3.299%
3	0.165%	3.989%	0.094%	3.754%
4	0.032%	1.519%	0.088%	2.416%
MEAN	0.079%	2.285%	0.088%	3.217%

This is particularly evident in some specific subpopulation. Since the ALE 2018 distribution will be published yearly by ISTAT taking into account for some other variables such as gender, age classes, citizenship, etc. it is important to take into account the distributional accuracy in these specific subpopulations. Looking at ALE 2018 distribution by citizenship and comparing the two estimation approaches with the target variable distribution some differences are evident especially on subpopulation of foreigner. This subpopulation is smaller than the Italian one, consisting of about 27 thousand individuals (less than the 9% of total population analysed), and less information are available.

Table 5. Relative differences between Estimated and target ALE 2018 distribution by citizenship

ALE in 2018*	Italian		Non italian	
	Log-Lin. (*Dreli*)	MLP (*Dreli*)	Log-Lin. (*Dreli*)	MLP (*Dreli*)
1 Illiterate	2.920	-12.409	-5.556	-16.667
2 Literate but no ed. att.	-1.325	-0.295	-16.270	-20.238
3 Primary education	-0.156	-0.362	7.212	6.971
4 Lower secondary ed.	-0.128	-0.261	-1.873	-1.873
5 Upper secondary ed.	-0.356	0.041	1.268	3.006
6 Bachelor's degree	1.707	-0.569	10.048	-12.440
7 Master's degree	1.534	2.202	0.457	4.566
8 PhD	1.093	-11.475	17.647	105.882
Mean	RD=1.152	RD=3.452	RD=7.541	RD=21.455

* the results presented in this table come from the test set 2

The relative differences between estimated and target distributions are greater for Non Italian people (Table 5) with respect to Italian and are concentrated in the extreme and less frequent values. As it can be noticed, the greater differences are evident for the

MLP: the differences between the estimated and target frequencies calculated on the MLP estimates are almost always greater than those estimated with the Log-linear models.

5. Code/programming language

We make available a beta version of the python code for imputation with MLP on github temporary link: https://github.com/defausti/MLP_Imputation.git

6. Evolution of this study inside the organisation

In ISTAT the high use of administrative data poses the need to experiment with new methods able to work efficiently with a large amount of data of different nature and to ensure a high level of output accuracy.

ISTAT participation in the ML project was born with the aim of trying new methodological solutions in this new framework of statistical production and was the kick-off for the collaboration between the IT and the statistical units.

For the first time we tried the application of ML techniques to solve an imputation problem. The study results are considered of interest to the organization but further analysis is needed to improve accuracy in some particular subpopulations and to better understand how the ML works.

7. Is it a proof of concept or is it already used in production?

Actually, the use of ML techniques for the imputation/estimation of variables in an integrated dataset is still a proof of concept. The experimentation gives encouraging results demonstrating a gain in efficiency using the ML techniques and a high level of accuracy.

In particular, micro accuracy is very similar in the two approaches while the Log-linear models perform slightly better at the macro level, in which we are most interested. Some other studies need to be performed to better understand if and how accuracy can be improved particularly in some subpopulations. We hope that this trial will trigger further investigations on this topic.

As far as efficiency is concerned, we will experiment the use of raw (not pre-treated) variables in the ML approach; in Log-linear model variable pre-treatment cannot be skipped.

7.1 What is now doable which was not doable before?

ML techniques allow for a more efficient imputation process, solving the problem in one step and providing results similar to the Log-linear method.

This ML project was an opportunity to start working on this topic and start a collaboration between ISTAT offices, but there is still a lot of work to be done and the production needs, often do not allow you to dedicate the time necessary to the experimentation.

Other techniques, such as Random Forest and Linear Discriminant Analysis, should be experimented. Some preliminary analyses have been performed using these techniques and the results seem to be very promising, so we will have the opportunity to continue with the experimentation.

7.2 Is there already a roadmap/service journey available how to implement this?

There is still not a roadmap for the implementation of ML techniques but an informal working group is now active and interested to work on the topic, in the hope of being able to dedicate the right time to it.

7.3 Who are the stakeholders?

The stakeholders are managers for Census data validation and outputs. The main interest is on the results.

7.4 Fall Back

For the publication of ALE, ISTAT is evaluating the estimate coming from the application of log-linear models. The ML approach is not yet considered in the official statistical production process.

8. Conclusions and lessons learned

ML techniques can be used for the imputation/estimation of variables. In particular, Multi Layer Perceptron algorithm has the following pros and cons:

Gain in efficiency: the estimation of ALE on the whole dataset can be performed in one step (one MLP model for all subpopulations A, B C), while the Log-linear approach involves the construction of different models.

Micro accuracy: The accuracy of predictions calculated on the micro-data indicates that the quality of the imputation is comparable in the two approaches.

Aggregate estimates: The ALE frequency distributions obtained by aggregating the microdata by educational level in the population and by subpopulations (Italian / non-Italian) show that the approach with the MLP makes estimates with a greater error in the less populated subclasses[8].

9. Potential organisation risk if ML solution not implemented

Finding an ML solution for our case study is not strictly necessary since standard methods for imputation already exist.

10. Has there been collaboration with other NSIs, universities, etc?

No collaborations outside the HLG-MOS ML project.

11. Next Steps

We are going to explore other standard and Machine Learning algorithms. Preliminary studies show good performance with Random Forest and Linear Discriminant Analysis.

To take into account for the actual structure of the population of interest, survey sample weights must be taken into account. Both, Log-linear and ML approach, will be re-estimated introducing sampling weights. In particular, as regards MLP, the sample weight w_i is applied to the contribution of each observation to the cross entropy loss function as follows:

$$loss = -\sum_{ic} w_i T_{ic} \log\left(P_{ic}\right)$$

where T_{ic} is target variable for the observation i; c is the modality index in a dummy representation. P_{ic} is the probability distribution of the output of the MLP for the observation *i*.

Finally, we intend to explore other architectures of neural networks such as GAN [6] that in the case of multivariate imputation show better performance.

[8] Considering the whole population the 2 methods are very similar. Log-linear models give origin to a little better results but differences may be negligible.

12. Bibliography

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

[2] Di Zio M., Filippini R., Rocchetti G. “An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data”, Workshop on Statistical Data Editing, Neuchâtel, Switzerland, 31 August - 2 September 2020

[3] Bernasconi, Eleonora, et al. "Satellite-Net: Automatic Extraction of Land Cover Indicators from Satellite Imagery by Deep Learning." arXiv preprint arXiv:1907.09423 (2019).

[4] De Fausti Fabrizio, Pugliese Francesco and Diego Zardetto. "Toward Automated Website Classification by Deep Learning." arXiv preprint arXiv:1910.09991 (2019).

[5] Cybenko, George. "Approximation by superpositions of a sigmoidal function." Mathematics of control, signals and systems 2.4 (1989): 303-314.

[6] Yoon, Jinsung, James Jordon, and Mihaela Van Der Schaar. "Gain: Missing data imputation using generative adversarial nets." arXiv preprint arXiv:1806.02920 (2018).