HLG MOS – Machine Learning project

April 1st, 2020

Imputation of the variable “Attained Level of Education” in Base Register of Individuals

Fabrizio De Fausti, Romina Filippini

Marco Di Zio, Simona Toti, Diego Zardetto

Istat | Istituto Nazionale di Statistica

Background and organisational context 1/2

Why is ISTAT carrying out experiments on the use of ML?

The ongoing change in the statistical production process, based on a system of registers, with a high use of adm data, poses the need to experiment new methods, able to efficiently work a large amount of data of different nature ensuring a high level of output accuracy.

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).

ISTAT is interested in a **micro level estimation of the ALE** (8 classes) for Italian resident population in October 2018, from register and survey data.

Background and organisational context 2/2

Who initiated the pilot study?

A working group was engaged on the imputation of ALE in the Base Register of Individuals (BRI). In the specific case, Log-linear models were studied.
Due to the complexity and heterogeneity of the available information an in-depth knowledge of data structure is needed and different steps of imputation have to be performed.

The experimentation was initiated with the aim to try solutions that could be able to solve the problem in a **more automated way**.

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

Data and its context 1/2

Type and source of data:

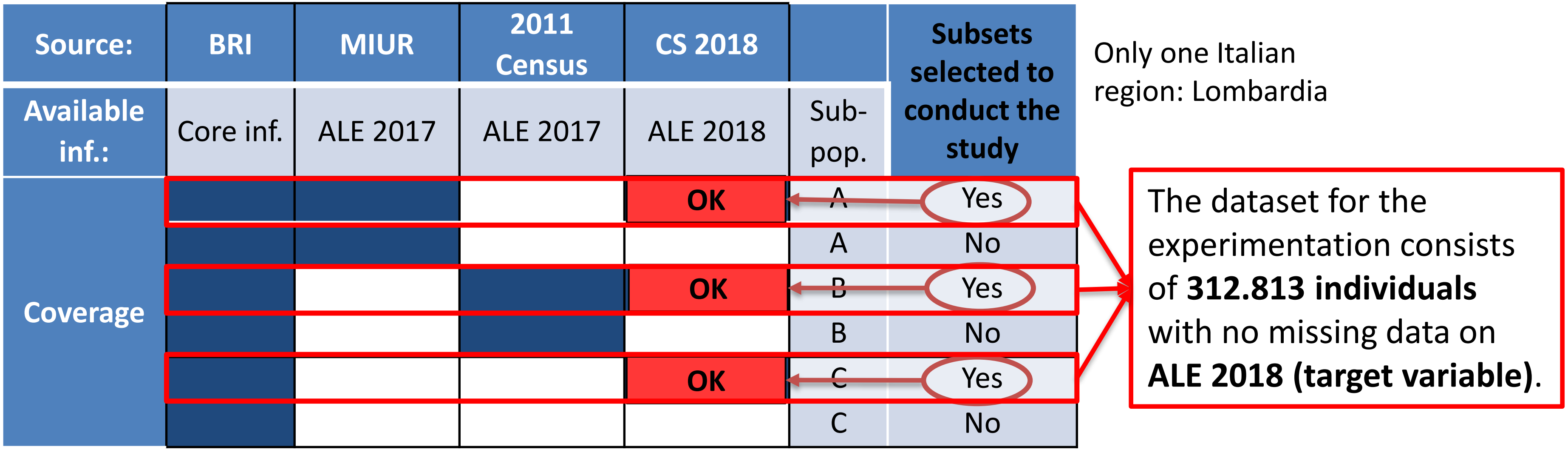
In carrying out the ALE prediction procedure, data of different nature are jointly used: administrative data, traditional Census data and sample survey data.

Source:	BRI	MIUR	2011 Census	CS 2018		Subsets selected to conduct the study
Available inf.:	Core inf.	ALE 2017	ALE 2017	ALE 2018	Sub-pop.	
Coverage	■	■		■	A	Yes
	■	■			A	No
	■		■	■	B	Yes
	■		■		B	No
	■			■	C	Yes
	■				C	No

Data and its context 1/2

Type and source of data:

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



Source:	BRI	MIUR	2011 Census	CS 2018		Subsets selected to conduct the study
Available inf.:	Core inf.	ALE 2017	ALE 2017	ALE 2018	Sub-pop.	
Coverage	■	■		OK	A	Yes
	■	■			A	No
	■		■	OK	B	Yes
	■		■		B	No
	■			OK	C	Yes
	■				C	No

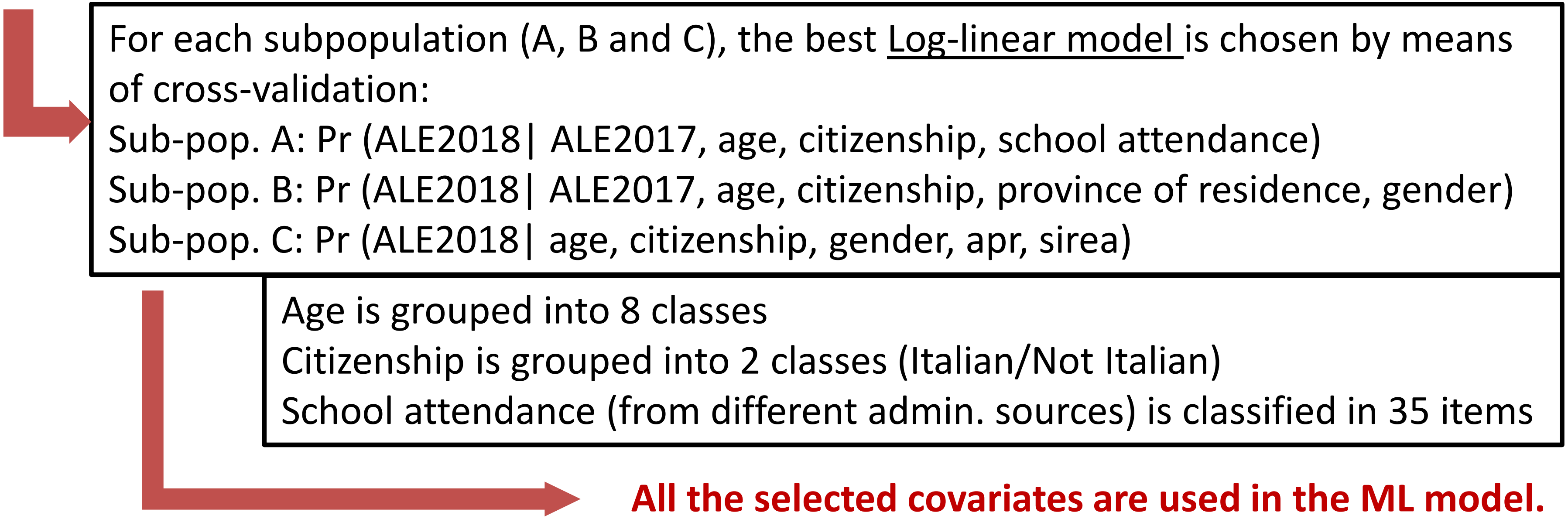
Only one Italian region: Lombardia

The dataset for the experimentation consists of **312.813 individuals** with no missing data on **ALE 2018 (target variable)**.

Data and its context 2/2

Data preparation and feature selection:

The aim of the experimentation is to compare ML and standard techniques under the same conditions.



For each subpopulation (A, B and C), the best Log-linear model is chosen by means of cross-validation:
Sub-pop. A: Pr (ALE2018| ALE2017, age, citizenship, school attendance)
Sub-pop. B: Pr (ALE2018| ALE2017, age, citizenship, province of residence, gender)
Sub-pop. C: Pr (ALE2018| age, citizenship, gender, apr, sirea)

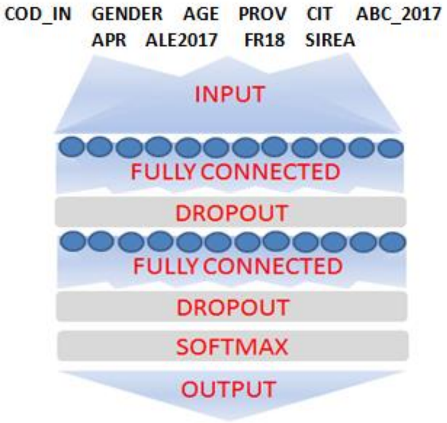
Age is grouped into 8 classes
Citizenship is grouped into 2 classes (Italian/Not Italian)
School attendance (from different admin. sources) is classified in 35 items

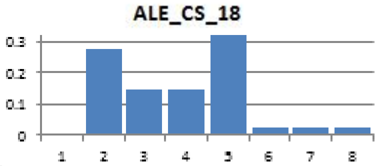
All the selected covariates are used in the ML model.

Machine Learning solutions 1/2

Our ML solution

- Experience with NN for NLP and Image Recognition.
- Simple **neural network** architecture, the Multi Layer Perceptron (MLP), to find the approximation of the **relationship** between the **input** variables and the probability distribution of the **output** variable for each pattern.
- We **impute** the ALE item **randomly extracted** from the **probability distribution** of the correspondent pattern.





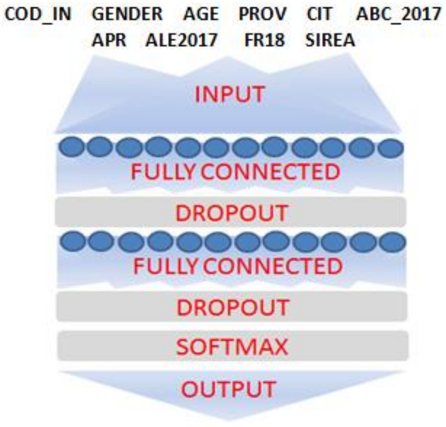
Model Training

- Dataset (312.813 samples) **splitting:** 80% Train and 20% Test
- **Input variables** are the **same** of LogLinear model
- Model selection: Best loss on Validation (20% of Train)

Machine Learning solutions 2/2

Model Training

- **1h to train** the model with Tesla V100-PCIE-16GB GPU about 27000 parameters (neural network weights)

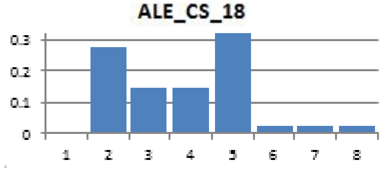


Metrics

- **Micro (Accuracy)** $$\sum_{i}^{N} \frac{(Obs\_ALE_i == Est\_ALE_i\ )}{N}$$

- **Macro (RD)** $$RD = \frac{1}{8}\Sigma_{i=1}^{8}\frac{|fr(\widehat{ALE18})_i - fr(ALE\_CS18)_i|}{fr(ALE\_CS18)_i} * 100$$

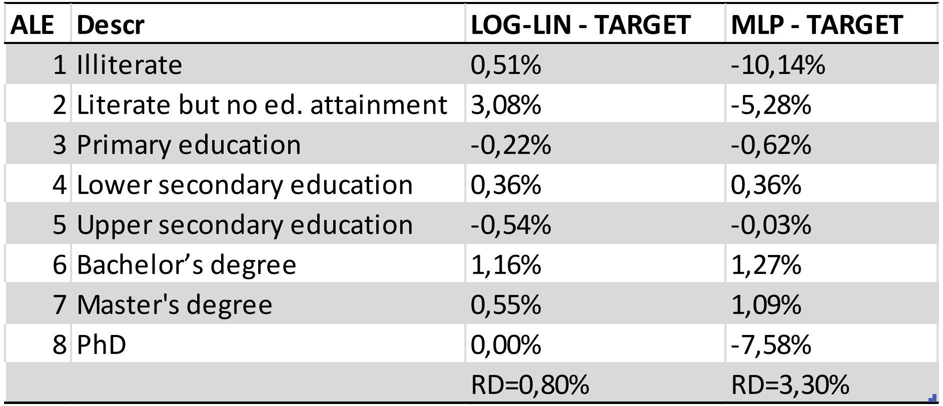
average of the absolute value of the relative differences between Ale18 distributions



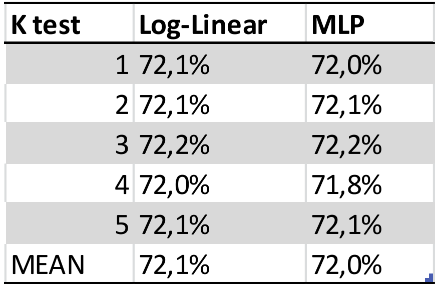
- Metrics are calculated for different splitting of dataset in train and test **(K Fold validation with k=5)**

Results 1/1

- The **accuracy** of predictions calculated on the micro-data indicate that the quality of the imputation is **comparable** in the two approaches.
- The ALE **frequency distributions** obtained by aggregating micro-data by educational level across the population and by subpopulations show that the approach with the **MLP** model makes estimates with a **greater error** in the less populated subclasses.
- 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.**



ALE	Descr	LOG-LIN - TARGET	MLP - TARGET
1	Illiterate	0,51%	-10,14%
2	Literate but no ed. attainment	3,08%	-5,28%
3	Primary education	-0,22%	-0,62%
4	Lower secondary education	0,36%	0,36%
5	Upper secondary education	-0,54%	-0,03%
6	Bachelor’s degree	1,16%	1,27%
7	Master's degree	0,55%	1,09%
8	PhD	0,00%	-7,58%
		RD=0,80%	RD=3,30%



K test	Log-Linear	MLP
1	72,1%	72,0%
2	72,1%	72,1%
3	72,2%	72,2%
4	72,0%	71,8%
5	72,1%	72,1%
MEAN	72,1%	72,0%

Outcomes 1/1

We are still in a study phase:

We have achieved similar results from ML techniques and Log-linear models using the same variables and with the same granularity. ML techniques could exploit the richer information content deriving from other variables or from the same variables with different granularity.

The studies gained interest in the organization in particular we presented this work at the internal *“Advisory Committee On Statistical Methods”.*

Excellent internal collaboration between statisticians and data-scientists.

Going beyond the demonstration or proof-of-concept

The use of ML techniques for the imputation of variables in an integrated dataset is still a **proof of concept**.

The experimentation gives encouraging results, however, some **other studies** need to be performed to better understand if and how accuracy can be improved in particular subpopulations.

There is still not a roadmap for the implementation of ML techniques.

Main **obstacle**: ML solutions are black-box algorithms.

We don’t have control on it, we can not interpret the model parameters.

Thanks to this HLG-MOS project, an **informal working group** is now active and interested to work on the topic.

<u>We hope that this trial will trigger further investigations on this topic.</u>

Conclusions

Does the application of ML add value to the production of official statistics?

Multi Layer Perceptron algorithm has the following **pros and cons**:

- More **automated** process: the estimation of ALE on the whole dataset can be performed in one step.
- **Accuracy**:
  - **Micro**: imputation results are comparable to the log-linear.
  - Aggregate estimates: frequency distributions of estimated ALE give origin to a **greater error in particular subclasses**.

Next steps

Take into account for the actual structure of the population of interest including survey **sample weights** in the estimation process.

Include **raw and new variables** in the MLP algorithm to exploit the potential of ML methods.

Analyze the **stability** of the models by generating multiple instances of the same model and repeating the random imputation process.

Explore **other standard and Machine Learning algorithms**. Preliminary studies show good performance with Random Forest and Linear Discriminant Analysis.

Explore **other architectures of neural networks** such as GAN that in the case of multivariate imputation show better performance.

HLG MOS – Machine Learnng project
April 1st, 2020

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

QUESTIONS?

Fabrizio De Fausti, Romina Filippini

Istat Istituto Nazionale di Statistica