Rome. 5-6 December 2022

WORKSHOP ON METHODOLOGIES IN OFFICIAL STATISTICS

A study of MLP for the imputation of the “Attained Level of Education” in Base Register of Individuals

Fabrizio De Fausti, Marco Di Zio, **Romina Filippini,** Simona Toti, Diego Zardetto

Istat | DIRECTORATE FOR METHODOLOGY AND STATISTICAL PROCESS DESIGN

Outline

- Context
- Data description
- Sampling weights in surveys
- Methods (Log-linear - MultiLayer Perceptron)
- Experimental study
- Results
- Conclusions

Introduction

- The Attained Level of Education (ALE) of the Permanent Italian Census relies on a high amount of **administrative information**. Nevertheless, it is necessary to resort to sample survey data to cope with delay of information and coverage gaps.

- Istat adopted a **mass imputation** approach integrating administrative and survey data for the ALE estimation of the Italian resident population, based on a sequence of log-linear imputations.

Due to the complexity and heterogeneity of the available information, the solution of the problem with standard statistical methods requires an in-depth knowledge of data structure and an **expensive initial phase of data analysis and treatment.**

GOAL: experiment the use of a **MLP** with the twofold objective:

- reducing human workload
- improving estimation accuracy

Context

- The **High-Level Group** for the Modernization of Official Statistics of UNECE (HLG-MOS) launched a Machine Learning project in 2019 with the aim of investigating of the use of machine learning for official statistics (**Timely**, **Accurate** and **Reliable** estimates).
- **Istat** worked on a comparison between the official imputation approach for ALE estimation, based on log-linear models and the Multilayer perceptron model (De Fausti et al., 2022).

ALE distribution on the Italian resident population is a standard output of the new yearly Permanent Italian Census, a good estimate of the ALE **frequency distribution** is crucial

we extend the study on MLP to include **sampling weights**

Data description

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

Source:	BRI	MIUR	2011 Census	CS	
Available information:	Core information *t*	ALE*t-2* Course att.	ALE*t-2*	ALE*t*	**Sub population**
Coverage					A
					B
					C

The main difference is between group A and the others:

- Group A is composed by "**Active**" people. Only administrative information are used.
- Groups B and C are "**Inactive**" people. The aim is to reproduce the ALE distributions observed in the CS within profiles.

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto

Data description

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

Source:	BRI	MIUR	2011 Census	CS	
Available information:	Core information *t*	ALE*t-2* Course att.	ALE*t-2*	ALE*t*	**Sub population**
Coverage					A
					B
					C

The dataset for the experimentation

- Only one Italian region: Lombardia.
- Individuals with no missing data on ALE *t* (target variable).

(312.813 individuals)

Sampling weights in surveys

NSIs use complex sampling designs to carry out **probability sample surveys**.
The joint effect of (i) unequal inclusion probabilities and (ii) use of auxiliary information for survey estimation determines unequal survey weights.

In **Standard approach** survey weights are used directly in estimates or incorporated into estimators.

- The model estimated for ALE imputation includes survey weights

In **Machine Learning approach** the inclusion of survey weights has received little attention in the literature.

- In order to leverage survey weights during the training phase of our MLP, we used a loss function weighted with sampling weights

$$loss_w = -\sum_{ic} \underbrace{\boxed{w_i}}_{\text{sampling weights}} T_{ic} \log(P_{ic})$$

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto

Methods

Official procedure: Log-linear model.

For each sub-population:

- conditional probabilities $h(I_t|X)$ are estimated on weighted count data
- ALE is imputed by randomly taking a value from this distribution

ML procedure: MultiLayer Perceptron.

- Single neural network
- Weighted loss function
- ALE is imputed by randomly taking a value from the distribution
  (better macro-level estimates)

1) MLP: Same input variables used in Log-Linear model

2) MLP All-in: All the variables in the dataset without any selection or reclassification
   (to study the possibility of using a more automated approach)

Istat

Experimental Study

- Estimates are computed using a **k-fold approach** with k=5:
  The dataset is partitioned into 5 subgroups and:
  a) the model is estimated on the training set, consisting of 4 of the 5 subgroups;
  b) the results are applied on the test set, composed of the remaining subgroup;
  c) Tasks (a) (b) are repeated 5 times so to reconstruct the entire data set.

- The results of estimates are compared. Quality measures are concerned with:
  - predictive accuracy of each unit (**micro level**)
  - accuracy of estimated aggregates (**macro level**: Kullback-Leibler divergence).

  - For each model the **process is repeated 100 times** to consider the model variability and the resulting indicators are averaged over those repetitions.

Results: Micro-level accuracy

Micro-level accuracy in the 5 test sets averaged over 100 runs

K-fold	Log-linear	MLP	MLP All-in
1	71.20	71.52	73.05
2	71.25	71.65	73.06
3	71.15	71.35	73.21
4	71.18	71.41	73.28
5	71.02	71.39	73.16
Mean	**71.16**	**71.46**	**73.15**
Standard Deviation	**0.077**	**0.110**	**0.088**

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto

Results: Macro-level accuracy

Macro-level accuracy: Kullback-Leibler divergence (KLD) in the 5 test sets averaged over 100 runs

K-fold	Log-linear	MLP	MLP All-in
1	0.008	0.019	0.022
2	0.017	0.014	0.045
3	0.015	0.044	0.057
4	0.032	0.018	0.114
5	0.024	0.020	0.102
Mean	**0.019**	**0.023**	**0.068**
Standard Deviation	**0.008**	**0.011**	**0.035**

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto
Istat

Results: Macro-level accuracy by citizenship

Macro-level accuracy by citizenship: Kullback-Leibler divergence (KLD) averaged over 100 runs (Fold 2)

ALE in 2018	Italian			Not Italian		
	Log-linear	MLP	MLP All-in	Log-linear	MLP	MLP All-in
Illiterate	0.029	0.023	-0.014	0.093	0.206	-0.080
Literate but no att.	-0.014	0.025	0.047	-0.829	0.226	-0.480
Primary education	-0.176	-0.071	-0.181	0.103	-0.654	-0.262
Lower secondary	0.043	-0.075	-0.757	0.479	-0.115	2.671
Upper secondary	0.148	0.151	0.965	1.385	0.361	0.774
Bachelor’s degree	0.002	-0.021	-0.204	-1.249	-0.881	-1.726
Master’s degree	0.021	0.032	0.259	0.585	1.273	0.053
PhD	-0.043	-0.053	-0.079	-0.133	-0.090	-0.256
KLD	**0.009**	**0.011**	**0.035**	**0.433**	**0.325**	**0.694**

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto

Conclusions and future developments

This paper aims at investigating the behavior of **MLP as a tool for improving quality and efficiency** of the statistical process of ALE estimation. In order to leverage **survey weights** we modified the cross-entropy loss function using the sampling weights to create a pseudo-population.

- For the imputation of ALE the results of the MLP are very similar to those originated from log-linear models in terms of predictive accuracy and macro-level estimated frequency distribution.

- This study encourages to deepen the opportunity given by the use of ML methods of a more automated approach for the prediction of the variable ALE.

- In future work we want to explore the opportunity to manage longitudinal information in the MLP approach in order to obtain consistent estimates over time.

Thank you

FABRIZIO DE FAUSTI | defausti@istat.it
MARCO DI ZIO | dizio@istat.it
ROMINA FILIPPINI | filippini@istat.it
SIMONA TOTI | toti@istat.it
DIEGO ZARDETTO | zardetto@istat.it

Istat Istituto Nazionale di Statistica
Istat

Id	NAME	DESCRIPTION	Log-linear			MLP	MLP All-in
			A	B	C		
1	COD_IND	Record id					
2	GENDER	Gender		1	1	1	1
3	AGE_CLASS	Age classified into 14 levels	1	1	1	1	
4	AGE	Age in years					1
5	BIRTH_MU	Municipality of birth					1
6	BIRTH_CO	Country of birth					1
7	MUN	Municipality of residence					1
8	PROV	Province of residence		1		1	1
9	CIT_CLASS	Citizenship (Italian/Not Italian)	1	1	1	1	
10	CIT	Country of citizenship					1
11	ABC_2017	Subpopulation (A, B C)				1	
12	APR	ALE from APR classified into 4 levels			1	1	1
13	ALE2017	2017 ALE (combination of Administrative and 2011 Census)	1	1		1	1
14	FR18_CLASS	Aggregated type of school and year of attendance in 2017/2018	1			1	
15	FR18	Type of school and year of attendance in 2017/2018					1
		Resident in Italy in 2011 not caught by					

Results - Micro-level accuracy

WEIGHTED

K-fold	Log-linear	MLP	MLP All-in
1	71.20	71.52	73.05
2	71.25	71.65	73.06
3	71.15	71.35	73.21
4	71.18	71.41	73.28
5	71.02	71.39	73.16
Mean	**71.16**	**71.46**	**73.15**

NOT WEIGHTED

K-fold	Log-linear	MLP	MLP All-in
1	72.15	72.05	73.49
2	72.14	72.18	73.59
3	72.27	72.27	73.67
4	72.10	72.24	73.54
5	72.08	71.93	73.45
Mean	**72.15**	**72.13**	**73.55**

NOT WEIGHTED
compared with the weighted benchmark

K-fold	Log-linear	MLP	MLP All-in
1	71.31	71.22	72.51
2	71.35	71.35	72.64
3	71.29	71.28	72.57
4	71.33	71.39	72.62
5	71.18	71.06	72.45
Mean	**71.29**	**71.26**	**72.56**

Results - **Macro-level accuracy**: KLD

WEIGHTED

K-fold	Log-linear	MLP	MLP All-in
1	0.008	0.019	0.022
2	0.017	0.014	0.045
3	0.015	0.044	0.057
4	0.032	0.018	0.114
5	0.024	0.020	0.102
Mean	**0.019**	**0.023**	**0.068**

NOT WEIGHTED

K-fold	Log-linear	MLP	MLP All-in
1	0,007	0,011	0,012
2	0,008	0,019	0,014
3	0,009	0,027	0,013
4	0,026	0,014	0,024
5	0,009	0,023	0,010
Mean	**0,012**	**0,018**	**0,015**

NOT WEIGHTED
compared with the weighted benchmark

K-fold	Log-linear	MLP	MLP All-in
1	0.009	0.030	0.024
2	0.017	0.030	0.020
3	0.016	0.063	0.035
4	0.036	0.022	0.031
5	0.024	0.036	0.028
Mean	**0.020**	**0.036**	**0.028**

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto