Istat | Istituto Nazionale di Statistica

3-7 October 2022, (virtual)

EXPERT MEETING ON STATISTICAL DATA EDITING

The imputation of the “Attained Level of Education” in the base register of individuals through Neural Networks using sampling weights

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

Istat | DIRECTORATE FOR METHODOLOGY AND STATISTICAL PROCESS DESIGN

Outline

- Context
- Data description
- Sampling weights in surveys
- Multi Layer Perceptron
- Results
- Conclusions

Context

This presentation follows a p**recedent work**, De Fausti et al (2022) about **comparison** between:

- **log-linear** models(the official imputation approach)
- machine learning approach (**Multi Layer Perceptron**) MLP

Imputation of a variable in base register of individuals

- “Attained Level of Education” ALE (the methodology can be generalized to **other variables**)

What’s new ? Introduction of the survey **sampling weights** in the precedent imputation process

Context

Quality measures of imputation

- **Micro-level** (Accuracy)
- **Macro-level** (Distribution agreement)

Advantages of ML respect log-linear approach

- **Automation** of the process
- **Efficiency** of the preprocessing phase

Data Description

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

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

Data Description

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.	
				OK		
Coverage				OK	A	Yes
					A	No
				OK	B	Yes
					B	No
					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)**.

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

Data Description

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.

Sampling weights in surveys

To create labelled dataset we use the **sample survey data**:

- Sampling design is complex
- Inclusion probabilities are unequal
- Design weights, are reciprocal of inclusion probabilities

In **standard approach** the ALE distribution estimates are calculated:

- Horvitz-Thompson estimators that take into account the weights
- Calibration estimators, which leverage available auxiliary information

Sampling weights in surveys

Machine Learning approach including sampling **weights**:

- ML approaches are mainly applied to make **micro-level predictions**
- There is **not a large literature** for sampling weights inclusion
- Inclusion of the **survey weights** during the training phase of our MLP **loss function**
- Take in account pseudo-population obtained by **cloning each training** example
- **Improve classification** results specially for groups units characterized by **higher weights**
- **Low frequency** ALE classes, which might be under-represented in the **unweighted sample**

$$loss_w = -\sum_{ic} w_i T_{ic} \log(P_{ic})$$

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

Sampling weights in surveys

Machine Learning approach including sampling **weights**:

- ML approaches are mainly applied to make **micro-level predictions**
- There is **not a large literature** for sampling weights inclusion
- Inclusion of the **survey weights** during the training phase of our MLP **loss function**
- Take in account pseudo-population obtained by **cloning each training** example
- **Improve classification** results specially for groups units characterized by **higher weights**
- **Low frequency** ALE classes, which might be under-represented in the **unweighted sample**

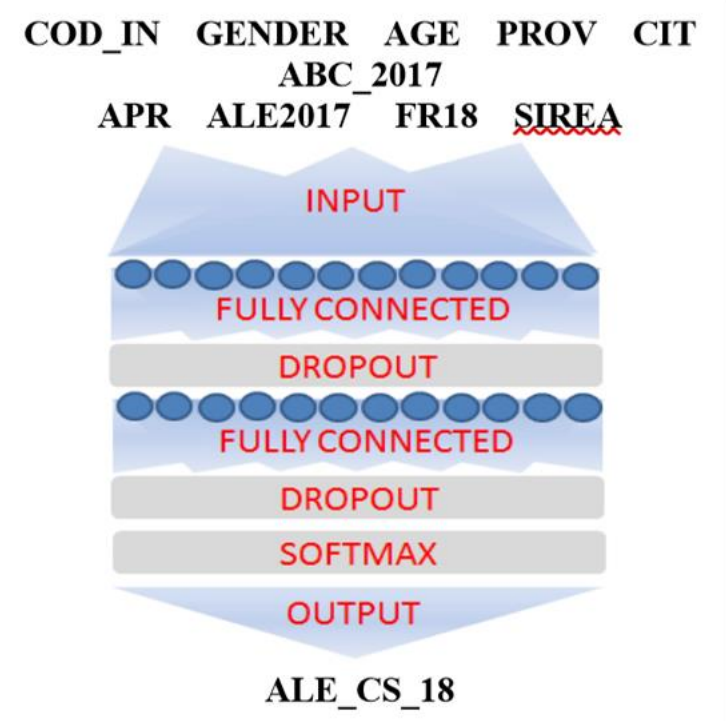
$$loss_w = -\sum_{ic} \boxed{w_i} T_{ic} \log(P_{ic})$$

sampling weights

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

Multi Layer Perceptron

- Single neural network, unlike the standard approach
- One-hot encoding
- Two hidden layers each of 128 neurons
- Dropout to prevent overfitting
- Weighted cross-entropy
- Softmax returns the probability distribution over the 8 ALE classes
- In the inference phase extraction of ALE from the softmax distribution (better macro-level estimates)



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

Experimental Study

- We **compare MLP** vs **Log-Linear** models
- Training/Test dataset corresponds approximately to **5% of total population** of interest (Lombardia Italian Region)
- **MLP model** uses the **same input variables** used in Log-Linear model
- **MLP All-in model** are not pre-processed. **All the variables** in the dataset without any selection or reclassification (**automatic approach**)
- Micro-level and Macro-level **quality measures** are calculated using a **k-fold approach** with k=5 (training/test spitting independence)
- For each model the same **imputation process is repeated 100 times** to consider the model variability and the resulting indicators are averaged over those repetitions.

Results

Micro level quality: Accuracy

K-fold	Log-linear	MLP	MLP All-in
1	71.202	71.521	73.047
2	71.254	71.648	73.059
3	71.155	71.35	73.209
4	71.183	71.405	73.279
5	71.023	71.385	73.155
Mean	**71.163**	**71.462**	**73.15**
Standard Deviation	**0.077**	**0.11**	**0.088**

Results

Macro level quality: Kullback-Leibler divergence

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	0.02	0.102
Mean	**0.019**	**0.023**	**0.068**
Standard Deviation	**0.008**	**0.011**	**0.035**

Results

Macro level quality for sub-populations: Kullback-Leibler divergence (Fold 2)

ALE in 2018	Italian			Not Italian		
	Log-linear (DKL)	MLP (DKL)	MLP All-in (DKL)	Log-linear (DKL)	MLP (DKL)	MLP All-in (DKL)
Illiterate	0,024	0,023	-0,02	0,061	0,181	-0,025
Literate but not…	-0,008	0,031	0,052	-0,832	0,16	-0,461
Primary education	-0,177	-0,086	-0,189	0,122	-0,692	-0,243
Lower secondary..	0,035	-0,071	-0,782	0,39	-0,03	2,5
Upper secondary..	0,14	0,173	1,003	1,568	0,361	0,98
Bachelor's degree	0,01	-0,057	-0,213	-1,219	-0,909	-1,817
Master's degree	0,027	0,06	0,255	0,508	1,348	-0,006
PhD	-0,04	-0,061	-0,071	-0,16	-0,08	-0,229
Mean (DKL)	**0,058**	**0,07**	**0,323**	**0,608**	**0,47**	**0,783**

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

Conclusion

.

- In MLP we modified the cross-entropy loss function using the sampling weights to create a pseudo-population
- Sampling weights improve estimates of ALE classes, which might be under-represented in the unweighted sample
- MLP and Log-Linear approaches returns similar results
- MLP encourages the possibility of using a more automated approach
- In future work we want explore other ML algorithms e.g. Random Forest
- Integration of longitudinal information in the imputation process

Bibliography

De Fausti F., Di Zio M., Filippini R., Toti S., Zardetto D. (2022). Multilayer perceptron models for the estimation of the Attained level of Education in the Italian Permanent Census. Statistical Journal of the IAOS, 38, pp. 637–646

Thanks for your attention

Fabrizio De Fausti | defausti@istat.it

Istat | Istituto Nazionale di Statistica