UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing
3-7 October 2022, (virtual)

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

De Fausti F., Di Zio M., Filippini R., Toti S., Zardetto D. (Istat, Italy)

defausti@istat.it, dizio@istat.it, filippini@istat.it, toti@istat.it, zardetto@istat.it

I. Introduction

1. The Attained Level of Education (ALE) of the Permanent Italian Census relies on a high amount of administrative information. Nevertheless, it is needed to resort to sample survey data to cope with delay of information and coverage problems. The official procedure for the estimation of the ALE (8 classes) for Italian resident population in 2018 relies on a mass imputation, see Di Zio et al. (2019).

2. Due to the complexity and heterogeneity of the available information, the solution of the problem with standard statistical methods needs the construction of different imputation models with a strong effort in terms of human intervention. Machine Learning (ML) techniques could provide a more automated data driven alternative for the imputation task.

3. In De Fausti et al. (2022), a comparison between the official imputation approach, based on log-linear models, and Multilayer perceptron models is discussed.

4. The evaluation focuses on two quality aspects: accuracy of predictions (and of estimated aggregates computed by directly using the predictions) and efficiency of the procedure. The efficiency assessment is primarily concerned with the automation of the process, which means that resources spent for data analysis and preparation can be minimized. Results are encouraging especially concerning the efficiency. In fact, they do not notice an improvement in terms of accuracy, but the same level of quality is reached by using raw data, that is without resorting to expensive data pre-treatment steps.

5. In that application, survey data are used without considering sampling weights. The role of sampling weights is to make the sample representative of the whole population, thereby leading to unbiased estimates. Although still under discussion, techniques to incorporate sampling weights in classical statistical models are developed (Pfefferman, 1993), the same cannot be said for machine learning models.

6. In this work, we extend the study in De Fausti et al. (2022) to include sampling weights in Multilayer perceptron models.

7. The paper is structured as follows. In Section II, we briefly explain how sampling weights are taken into account in a survey and in a ML approach; Section III describes the data used for our experimentation; Section IV describes the official and the machine learning imputation methods compared in this study using the sampling weights; Section V describes the experimental study; some final remarks are given in Section VI.

II. Taking into account sampling weights

A. Sampling weights in surveys

8. National Statistics Institutes (NSIs) routinely use complex sampling designs to carry out probability sample surveys. This practice results from the need of finding a tradeoff between statistical efficiency and logistic constraints. To the scope of the present paper, any sampling design resulting in unequal inclusion probabilities of the observed sample units can be considered complex. Any statistical analysis on complex survey data should be performed taking into account the selection of sample units with unequal probabilities. Failing this, inferential results would be generally invalid, even under ideal conditions (that is neglecting any form of non-sampling error, like sampling frame imperfections, non-response, measurement errors, etc.). The design-based/model-assisted approach to finite population sampling is the reference inferential framework adopted by NSIs. By properly incorporating inclusion probabilities into estimators, it leads to unbiased (or asymptotically unbiased in the large sample limit) estimation without any need of model assumptions on the target population. In this approach to inference, inclusion probabilities typically enter estimator expressions in the form of weights attached to survey units. Horvitz-Thompson estimators use so-called design weights, which are reciprocal of inclusion probabilities. Calibration estimators, which leverage available auxiliary information on the target population to improve estimation efficiency, employ so-called calibration weights, which are complex non-linear functions of design weights and embedded auxiliary information. Furthermore, despite non-sampling errors mark a departure from the ideal conditions underpinning the validity of the design-based/model-assisted inferential framework, NSIs invariably strive to mitigate estimation flaws that could arise from non-sampling errors by adjusting the weights. For instance, to adjust weights for total non-response and/or frame imperfections, propensity score modelling, and calibration are commonly applied alternatives, the choice among the two being mainly driven by the available auxiliary information.

9. Put briefly, the joint effect of (i) unequal inclusion probabilities and (ii) usage of auxiliary information for survey estimation determines unequal survey weights that should not be overlooked when fitting statistical models to data from complex surveys should be estimated.

B. Sampling Weights in ML

10. To the best of our knowledge, the question whether (and how) survey weights should be incorporated in Machine Learning models trained to survey data has received little to null attention in the literature. One possible explanation might be that research in the ML field is typically more concerned with achieving high prediction accuracy at micro-level than aimed at obtaining reliable estimates of model and/or finite population parameters. However, as explained in the introduction, the latter objective is of major relevance to the scope of our work. In fact, we need to assess the ability of MLP imputation to provide good estimates of the ALE frequency distribution, which is one of the standard statistics disseminated by the Italian Permanent Census on a yearly basis. For this reason, in order to leverage survey weights during the training phase of our MLP, we used a loss function weighted with sampling weights.

11. The intuition behind this formula is similar to the "census equations" leading to the Pseudo Maximum Likelihood (PML) approach. Basically, the loss function is computed on a pseudo-population of training observations obtained by cloning each training example i, w_i(weight) times. This way, the MLP incurs different misclassification costs for different training examples, owing to unequal survey weights. As compared to the ordinary unweighted loss, the expected effect is to improve classification results of the MLP especially for groups of survey units characterized by higher-than-average weights. In turn, this could be particularly beneficial to MLP predictions for low frequency ALE classes, which might be under-represented in the unweighted sample.

III. Data description

A. Resident population data

12. ALE for the Italian resident population in 2018 is estimated by using administrative data, traditional Census data and sample survey data.

13. 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).

14. 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.

15. 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).

16. 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 available information for mass-imputation of the attained level of education at time t

Source:	BRI	MIUR	2011 Census	CS2018		
Available inf.:	Core inf.	ALE2017	ALE2017	ALE2018	**Sub-population**	Used in the Case study
Coverage					**A**	**Yes**
					A	No
					B	**Yes**
					B	No
					C	**Yes**
					C	No

17. Core information like Age, gender, citizenship, marital status, place of birth and place of residence are available for all individuals.

18. 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;
C. Individuals neither in MIUR nor in 2011 Census belong to group C. For this group no information on ALE is available.

19. 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.

IV. Methods

A. Official procedure: Log-linear model imputation

20. The adopted official procedure is based on log-linear imputation. As stated in Singh (1988), this method generalizes hot-deck imputation by choosing suitable predictors for forming "optimal" imputation classes. In fact, the approach is based on modeling the associations between variables.

21. The general idea is to estimate a model for the prediction of ALE at time t (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 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)/ $\widehat{N}$(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 increases.

22. In order to consider sampling weights in the model, 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).

B. Machine learning procedure: Multilayer perceptron model

23. We apply the MLP for the mass imputation of ALE with the same categorical input variables of the log-linear models.

24. Our approach aims to be as general as possible, therefore:
   (a) We train a single neural network, unlike the standard approach, where different models are built, according to the variables available for each profile;
   (b) We encode the input variables of the perceptron multilayer as one-hot encoding, in this representation the missing value of a variable is encoded like any other mode of the variable;
   (c) 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.

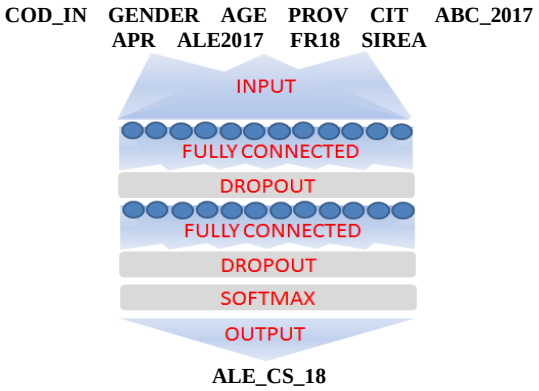
25. In order to leverage survey weights during the training phase of our MLP, we modified the cross-entropy loss function as follows:

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

where w_i is the final survey weight of the i-th training observation, c is the modality index in a one-hot representation, T_{ic} is the ground-truth value of the target variable for the i-th observation, and P_{ic} is the corresponding softmax function output probability distribution of the MLP.

26. 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



27. 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.

28. 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.

29. 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).

V. Experimental study

A. Description of the simulation

30. The comparison of MLP with the official imputation model is carried out on the 312,813 people residents in Lombardia in 2018 with no missing data on ALE 2018. The target variable is the self-declared ALE in the 2018 sample census, referring to the year 2018, which corresponds approximately to 5% of total population of interest. The subset of units is limited to the subpopulations B and C as classified in Table1.

31. A first experiment is carried out by using the MLP with the same covariates selected for log-linear models. The goal is to minimize confounding factors, therefore allowing for a neat comparison of results in terms of statistical accuracy.

32. In a second experiment (MLP all-inn), data provided to MLP are not pre-processed. All the variables in the dataset enter the MLP algorithm without any selection or reclassification. In particular, the variables age and citizenship are not aggregated into classes and the variables relating to the type of school attended are used as they are presented from administrative sources, without any type of aggregation. The variables relating to the place of residence and place of birth are also included. Moreover, the information on the data source of the three subpopulations is not considered and the flag variable (ABC_2017) which identifies the three subgroups A, B and C, is not introduced. This second experiment is clearly meant to study the possibility of using a more automated approach for the prediction of the ALE variable in large-scale production settings.

33. The variables used in the different experiments are described in Table 2.

Table 2: Variables in the dataset used in the three log-linear models and MLP approach

Id	NAME	DESCRIPTION	Log-linear			MLP	MLP without pre-processing
			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
16	SIREA	Resident in Italy in 2011 not caught by the 2011 Census			1	1	1
17	ALE_CS18	2018 ALE from 2018 Census Survey	Target variable				

34. The results of estimates obtained with MLP are compared with those of the official procedure. Quality measures are concerned with predictive accuracy of each unit and accuracy of estimated aggregates (quantities obtained by aggregating the unit predictions). The first measure is generally the one analyzed in ML approaches, while the second is usually considered in National Statistical Institutes when evaluating the quality of an estimation procedure. Since the ALE distribution will be published by gender, age classes and citizenship, it is important to evaluate the distributional accuracy in these specific subpopulations. The aggregates considered in this study refer to the main figures that are officially disseminated by Istat. In particular, we report results for the ALE distribution by citizenship.

35. Accuracy is calculated using a k-fold approach with k=5. The database 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) accuracy is calculated only on the test set as the difference between estimated ALE 2018 and the observed ALE 2018.

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

36. After the implementation of this approach, each individual (in each k-fold) has two probability distribution on the 8 ALE items, estimated using ML and log-linear models. The imputation process consists of extracting a random value from the probability distribution. The same imputation process is repeated 100 times to consider the model variability and the resulting indicators are averaged over those repetitions.

B. Results

37. Table 3 shows the micro-level predictive accuracy attained by log-linear and MLP approaches. For each method and k-fold, the proportions of units with predicted ALE equals the observed (i.e. true) value are reported as percentages.

Table 3. Micro-level accuracy in the 5 test sets averaged over 100 runs: Log-linear, MLP estimation and MLP All-in (Percentage values)

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	71.350	73.209
4	71.183	71.405	73.279
5	71.023	71.385	73.155
Mean	71.163	71.462	73.150
Standard Deviation	0.077	0.110	0.088

38. The results of the MLP are very similar to those originated from log-linear models: the average predictive accuracy, computed over the 5 folds, are respectively equal to 71.2% and 71.5%. MLP all-inn has a slightly better behaviour.

39. To evaluate the performance of the imputation procedures at macro-level, the estimated frequency distribution of ALE in 2018 ($\widehat{ALE}18$) is compared with that computed using the 2018 census sample (ALE_CS18).

40. A possible synthetic measure is given by the Kullback-Leibler (Dkl) divergence

$$D_{KL}(T|\hat{T}) = \sum_{c=1}^{K} T_c \log_2 \left(\frac{T_c}{\hat{T}_c}\right)$$

It measures the divergence of the distribution T from $\hat{T}$, or, in other words, the information lost when $\hat{T}$ is used to approximate T. If the two distributions are identical the Kullback-Leibler divergence is equal to 0.

41. Table 4 provides the DKL computed for log-linear, MLP and MLP all-in estimation methods. Also in macro-accuracy, MLP is very close to log-linear imputation, while MLP all-in shows a greater difference.

Table 4. Macro-level accuracy: Kullback-Leibler divergence (DKL) in the 5 test sets averaged over 100 runs: Log-linear, MLP estimation and MLP All-in

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

42. Table 5 reports D_{KL} for the distribution of ALE 2018 by citizenship. We notice that largest differences are related to the subpopulation of 'not Italian'. This subpopulation is much smaller than the Italian one, consisting of about 27 thousand individuals (less than 9% of total population analyzed), and less information is available for it.

Table 5. Kullback-Leibler divergence between Estimated and target ALE 2018 distribution by citizenship: Log-linear vs MLP vs MLP All-in estimation (test set 2 averaged over 100 runs)

	Italian			Not Italian		
ALE in 2018	Log-linear (D_{KL})	MLP (D_{KL})	MLP All-in (D_{KL})	Log-linear (D_{KL})	MLP (D_{KL})	MLP All-in (D_{KL})
Illiterate	0,024	0,023	-0,020	0,061	0,181	-0,025
Literate but no ed. Att.	-0,008	0,031	0,052	-0,832	0,160	-0,461
Primary education	-0,177	-0,086	-0,189	0,122	-0,692	-0,243
Lower secondary ed.	0,035	-0,071	-0,782	0,390	-0,030	2,500
Upper secondary ed.	0,140	0,173	1,003	1,568	0,361	0,980
Bachelor's degree	0,010	-0,057	-0,213	-1,219	-0,909	-1,817
Master's degree	0,027	0,060	0,255	0,508	1,348	-0,006
PhD	-0,040	-0,061	-0,071	-0,160	-0,080	-0,229
Mean (D_{KL})	0,058	0,070	0,323	0,608	0,470	0,783

43. Within the Italian subpopulation, we notice that MLP and log-linear have a similar performance to the previous tables, with a small preference for log-linear. On the other hand, it is interesting to note that MLP has a better performance in the Not Italians.

VI. Final remarks and future developments

44. This paper aims at investigating the behavior of Neural Networks with the use of sampling weights through a comparative study with the officially adopted log-linear imputation procedure. In order to leverage survey weights during the training phase of our MLP we modified the cross-entropy loss function using the sampling weights to create a pseudo-population. The effect is to improve classification results especially for groups of survey units characterized by higher-than-average weights and for low frequency ALE classes, which might be under-represented in the unweighted sample. 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 the possibility of using a more automated approach for the prediction of the ALE variable in large-scale production settings.

References

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

Di Zio M., Filippini R., Rocchetti G. (2019). An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. *Rivista di Statistica Ufficiale, n. 2-3/2019, pp. 143-174.*

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317-337.

Lumley, T., Scott, A. (2017). Fitting regression models to survey data. *Statistical Science,* 265-278

Singh A. C. Log-linear imputation. Methodology Branch Working Paper Statistics Canada. 1988; 88-29.

Scholtus S. (2018). Variances of Census Tables after Mass Imputation, Discussion paper CBS.

Skinner, C.J. and Vallet, L.-A. (2010). Fitting log-linear models to contingency tables from surveys with complex sampling designs: an investigation of the Clogg-Eliason approach. *Sociological methods & research*, 39 (1), pp. 83-108.

Thibaudeau, Y., Slud, E. and Gottschalck, A. (2017). Modeling log-linear conditional probabilities for estimation in surveys. *The Annals of Applied Statistics.* 11, pp. 680-697.