ON THE USE OF RANDOM FOREST TO IMPUTE CATEGORICAL VARIABLES BEYOND THE SAMPLE

Ilaria Bombelli, Romina Filippini, Simona Toti

Abstract. The new paradigm of the Official Statistical production is based on a system of registers, resulting from the integration of administrative and survey data. Administrative data provide a complete enumeration of the units they cover: however, this population usually represents only a specific subset of the statistical target population, and such a subset is typically not obtained through a probabilistic sampling design. Similarly, survey data also cover only a subset of the population of interest, but they are made representative through the use of weights. Another common limitation of administrative sources is the delay in data availability.

In this context, generating a complete and consistent dataset is a critical task, which requires the implementation of specific procedures to account for delayed data and to impute missing values. One possible strategy is to use survey data as the target variable. However, this raises the issue of how to properly incorporate weights into the estimation model.

An example of a mass imputation approach is provided by the official estimates of the Attained Level of Education (ALE) adopted by the Italian National Institute of Statistics (Istat) for all the resident population in Italy. The official procedure is based on the estimation of different log-linear models.

In this application, we focus on the use of Random Forest (RF) to leverage the opportunities presented by Machine Learning (ML) technique on using all available information, including longitudinal data and variables with many categories. This approach allows for capturing complex relationships between variables, which is often challenging to incorporate comprehensively using standard methods.

The results are evaluated by focusing on different scenarios, each related to a level of available information, and focusing on three population subsets, each characterized by a different pattern of available information.

1. Introduction

Nowadays, Official Statistics is required to produce reliable and timely data to analyse and describe phenomena related to demography, economy, environment, and other aspects of society. Such data can be useful for decision makers and stakeholders to clearly understand problems and plan necessary actions. At the same time, it is required to satisfy several constraints. On the one hand, more and new statistics are demanded, requiring considerable resources, both monetary and in terms of effort. On the other hand, survey units are less likely to spend their time in filling in questionnaires, leading to high non-response rates. Therefore, Official Statistics must also address the issue of respondent burden.

In this framework, Official Statistics can exploit new data sources, such as administrative data, and new data collection methods. These alternatives can be used for statistical purposes. Moving from considering only traditional data sources and collection methods to being able to combine both traditional and new ones implies an important change in the paradigm. The new paradigm of the Official Statistical production is based on a system of registers, resulting from the integration of administrative and survey data.

The integration of administrative and survey data addresses several challenges. First, both administrative and survey data typically cover only a portion of the target population. Indeed, administrative data provide a complete enumeration of the units within the population they capture, but this population usually represents only a specific subset of the statistical target population. Similarly, survey data also cover only a subset of the population of interest, but representativeness is ensured through the application of sampling weights, derived from the survey design, and through calibration techniques. Another typical issue concerns the timing of data availability, as administrative sources are often provided with a lag time. Moreover, while survey data are specifically designed for statistical purposes, administrative data are not; therefore their integration requires additional effort, particularly in the harmonization of target parameter and the definition of target populations.

Consequently, within this framework, generating a complete and consistent dataset becomes a critical task. This requires the implementation of various procedures to predict delayed data and impute missing values. One possible strategy is to use survey data as the target variable.

An example of applying a mass imputation approach is provided by the official estimates of the Attained Level of Education (ALE) adopted by the Italian National Institute of Statistics (Istat) to cover the whole resident population in Italy.

The official procedure consists in estimating various log-linear models. In this paper, we propose an alternative approach, implementing a ML technique, specifically Random Forest (Breiman, 2001), to impute ALE. Indeed, ML approach allows to leverage all available information, including longitudinal data and variables with many categories. Moreover, these methods are able to capture complex relationships between variables that are often difficult to model comprehensively using traditional methods. However, the use of ML methods brings up the issue of how to appropriately incorporate sampling weights into the estimation process.

The remainder of the paper is organized as follows. Section 2 provides a general framework for ALE imputation, describing the data availability and their structure and detailing the official procedure based on log-linear models. Section 3 presents an overview of the Random Forest (RF) model. In Section 4, we illustrate the application of RF for ALE imputation, including a discussion on the incorporation of sampling weights and variable selection. Finally, Section 5 summarizes the main findings and discusses relevant issues and further developments.

2. Framework on Attained Level of Education

Many statistical outputs on the Italian resident population produced by Istat originate from the Base Register of Individuals (BRI), a statistical register that provides the most comprehensive coverage of individuals and contains core demographic information for each unit, gathered from various sources. Since a substantial amount of information related to education is available at the micro level from administrative sources, it is of interest to produce register-based statistics on ALE rather than survey-based. However, a key limitation of administrative data is that they are not fully up to date: for reference year *t*, administrative information on ALE is only available up to year *t-1.* Moreover, administrative sources cover only a specific subset of the target population.

To this end, a strategy has been implemented to estimate ALE for all individuals included in the BRI. This approach allows to produce official ALE statistics at different levels of detail and aggregation.

As ALE for the reference year *t* is observed only for a sample of individuals, interviewed for the Italian Permanent Census, imputations are required in order to enrich the BRI with the information on ALE, for units beyond the sample. This can be achieved by exploiting the high amount of administrative information, available at a micro level.

2.1. Data Description

For each unit of the reference population, a set of core variables gathered from different sources is available. This information is age, gender, citizenship, place of birth, and place of residence.

As regards the topic of education, longitudinal information is available from administrative sources, in particular, the Ministry of Education, University and Research (MIUR) provides information about ALE and course attendance for people entering a study program after 2011 and covers the period from 2011 to *t-1* (scholar year *t-1/t)*.

For individuals who have not attended any school course since 2011 and are therefore not included in the MIUR dataset, data from the 2011 Census are used to fill the gap. However, a residual subset of individuals—accounting for about 5% of the reference population—is not covered by any administrative sources on ALE. This subgroup consists of individuals who either entered Italy after 2011 or were present in Italy in 2011 but not captured by the Census, and who have not enrolled in any study program recorded by MIUR.

Another source containing information on ALE is the so-called APR4-module. This is the registration and cancellation form that must be filled in when applying for a change of residence, either for a new registration in Italy from abroad and/or when changing one's usual residence within the country. ALE information on APR4 is self-declared and classified into 4 aggregated categories.

Finally, an important source of information on ALE comes from the survey. Since 2018, Istat has collected ALE data through the Italian Permanent Census sample survey. As a result, directly observed ALE data for the reference year *t* are available for approximately 5% of the population, known as the Master Sample (MS).

Since Istat classification for ALE estimates is based on 8 categories, all ALE variables (with the exception of APR4) are harmonized and reclassified accordingly: 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.

The different availability of information on ALE determines the partition of the target population into three subgroups, each characterized by a different pattern of available information.

A. Individuals who enrolled in a school course covered by MIUR in at least one school year after 2011, are characterized by a high amount of information available, including demographic information, longitudinal information on ALE, and school enrollment characteristics.
B. Individuals not in MIUR, interviewed in the 2011 Census, are characterized by the availability of demographic information and information on ALE referring to the year 2011.
C. Finally, individuals not in MIUR nor in the 2011 Census are associated only with demographic information, while ALE data are available only for a subset of them through the APR4-modules.

For each subgroup of individuals, ALE at year *t* is available for a representative sample interviewed in the MS.

2.2. Official procedure

The official procedure to impute ALE relies on the use of different log-linear models. Specifically, for individuals attending a school course in year *t-1*, the model predicts ALE given the year of the course attended the previous year. On the other hand, for individuals not enrolled in any school course covered by MIUR, due to the MIUR under-coverage, it is necessary to resort to sample survey data.

The conditional probabilities of each ALE category, are estimated for each profile, and the predicted class of each individual is then obtained by randomly drawing from the estimated probability distribution (Di Zio *et al.*, 2019).

3. Random Forest Algorithm

The Random Forest algorithm, introduced by Breiman (2001), is a general-purpose ML method suitable for both classification and regression tasks. It is an ensemble learning technique that generates multiple independent decision trees and combines their predictions using bagging (bootstrap aggregating). Specifically, each tree is constructed using a different bootstrap sample of the data and, at each node split, the best predictor is selected from a randomly chosen subset of features.

The final output of the RF model is obtained by aggregating the predictions of the individual trees. In regression tasks, this involves averaging the outputs of all trees. In classification tasks, the final prediction is typically based on a majority vote, although it is also possible to obtain the class probabilities instead of just the predicted class labels.

In our case, the classification forests were constructed using the Gini index as the splitting criterion. Class predictions were derived from the estimated class probabilities. Particularly, the predicted class of each individual is obtained by randomly drawing the class from the probability distribution of the classes related to each individual.

To implement RF in R, we used the ranger package (Wright & Ziegler, 2017). To achieve optimal predictive performance with RF, it is important to tune the hyperparameters — parameters that must be set before training and are not learned from the data (Probst *et al.*, 2019). Among the various hyperparameters, we focused on tuning: *mtry*, which controls the number of variables randomly selected as candidate features at each split; *min.node.sez,* which determines the minimum number of observations required in a terminal node (i.e., the minimum size of the leaves).

The results presented in the following sections were obtained using the tuneRanger R package (Probst *et al.*, 2018), which provides automated tuning of the hyperparameters for the ranger implementation of RF.

4. Application

The experiment in the present paper considers data from the Emilia Romagna Region, with the aim of producing a complete and consistent estimate of ALE for the reference year t=2022. To this purpose, a RF model is trained on the MS survey to predict the variable ALE for all residents in the Region.

The prediction of ALE for the units beyond the sample can be obtained from RF mainly in two ways: using the majority voting approach, i.e., predicting the class with the highest probability; predicting the class randomly drawn from the RF estimate probabilities.

Since our goal is to generate predictions that closely matches the distribution of ALE in the MS, the results are evaluated by assessing the degree of alignment between the predicted distribution and the benchmark distribution provided by the MS. As a consequence, we adopt the second approach as it provides a higher coherence at the macro-level (De Fausti *et al.*, 2022).

The majority voting approach, indeed, allows for good performance at the micro-level. However, it completely reduces all probabilistic information to a single category, thus losing the shape of the predicted distribution. In contrast, the random-draw approach allows the shape of the predicted distribution to be preserved, avoiding the distortion induced by the majority voting and preserving the variability (see, for example, Little and Rubin, 2002). In this sense, the second approach allows for better performance at the macro-level.

4.1. Sampling weights usage

The probabilistic sampling design of the MS assigns to each unit a sample weight that quantifies the number of population units that the unit represents.

Standard statistical model requires to include sampling weights at the likelihood level. In the ML framework, different possibilities are explored to consider the sample weights; we focus on two options: a) Ignoring them; b) Considering them by building the augmented version of the sample, by repeating each record a number of times given by the sampling weight of the record.

Option a) ignores the statistical probabilistic nature of the sample; in option b), the input dataset is augmented by replicating each row according to its sampling weight. In this way, the augmented MS mimics the target population structure.

In the initial experiments with RF, the same set of input variables used in the log-linear models is considered. Specifically, these variables are described in Section 4.2 and summarized in Table 2, Scenario 1.

Results are evaluated at the macro-level by comparing the distribution of ALE across the whole population with the benchmark, represented by the ALE distribution in the augmented MS, corresponding to the weighted ALE distribution. Performance is measured by computing the difference between the percentage relative frequencies of the predicted distributions and those of benchmark distribution.

Table 1 – *Absolute (a.v.) and percentage (%) distribution of ALE in the Augmented MS (benchmark). Differences between predicted and benchmark distribution.*

ALE	Augmented MS		Estimated – Augmented MS		
	a.v.	%	Log-Linear	RF Ignoring weights	RF Augmented data
1	16717	0.4%	0.05%	-0.01%	-0.02%
2	130504	3.2%	-0.21%	-0.25%	-0.25%
3	585713	14.2%	0.25%	0.09%	0.05%
4	1135472	27.4%	0.02%	-0.13%	-0.11%
5	1571471	38.0%	0.27%	0.32%	0.33%
6	187616	4.5%	-0.14%	0.00%	0.00%
7	486049	11.7%	-0.25%	-0.05%	-0.03%
8	23777	0.6%	0.00%	0.03%	0.02%
	Sum of absolute values		1.20%	0.88%	0.81%
	Mean of absolute values		0.15%	0.11%	0.10%

The results in Table 1 show that applying RF models leads to better performance, measured in terms of the sum of absolute differences, compared to log-linear models. The improvement is particularly evident for ALE classes 3 (Primary education), 6 (Bachelor’s degree), and 7 (Master’s degree). Since both models rely on the same set of input variables, the performance gain can be mainly attributed to the ability of RF to capture non-linear relationships.

Moreover, the inclusion of sampling weights in the construction of the augmented MS leads to slightly improved results.

Overall, the predictions obtained with the RF that incorporates sampling weights (applied to the augmented MS) are more consistent with the benchmark distribution than those obtained without using weights. Specifically, the sum of the differences between the estimated and benchmark distributions is 0.88% when weights are ignored, and 0.81% when weights are included.

Therefore, in the following analyses, the reported results refer to RF models trained on the augmented MS data.

4.2. Variables selection and encoding

After establishing the approach for accounting for sample weights and using the augmented MS as input data, we explored the full space of potential predictors to identify those to include in our model. We implemented three distinct scenarios to evaluate the effect of adding predictors and different levels of aggregation for predictor categories. These scenarios are designed to assess the contribution of additional data and the potential of ML techniques to capture complex relationships.

The first scenario (Scenario 1) includes the same variables used in the log-linear regression model. This allows for a direct comparison of performance between the methods, highlighting the ability of ML techniques to capture non-linear relationships. This feature set includes demographic characteristics, *i.e.*, age, gender, citizenship, and province of residence; educational data for the school year *t-2/t-1*, *i.e.*, ALE, school enrollment status, year of attendance, type of upper secondary school.

The second scenario (Scenario 2) includes the same variables used in Scenario 1, with more granular levels of detail for the variable Type of upper secondary school: from 3 to 28 categories.

Table 2 – *Description of input variables: type, presence of missing values and, encoding used in the RF. An: "X" indicates the inclusion of each variable in the scenarios.*

Variable	Type	anyNA	Encoding	Scen. 1	Scen. 2	Scen. 3
Age	Numerical		Ordinal	X	X	X
Gender	Binary		Dummy	X	X	X
Italian or not	Cat. (2 classes)		Dummy	X		X
Province of residence	Categorical		Dummy	X	X	X
Subgroup	Cat. (3 classes)		Dummy	X	X	X
ALE from Administrative Sources (MIUR or Cens11)	Cat. (8 classes+1)	X	Ordinal	X	X	X
ALE from APR4	Cat. (4 classes+1)	X	Ordinal	X	X	X
Year of attendance	Cat. (32 classes+1)	X	Ordinal	X (*t-1*)	X	X (*t-1 ; t-2*)
Evening school	Cat. (2 classes+1)	X	Dummy			X *(t-1 ; t-2*)
Type of attendance	Categorical	X	Dummy			X *t-1 ; t-2*)
Type of primary/lower secondary school	Cat. (2 classes+1)	X				X *(t-1 ; t-2*)
Type of upper secondary school	Cat. (2 classes+1)	X	Dummy	X (*t-1*)		
Type of upper secondary school	Cat. (28 classes+1)	X	Dummy		X	X *(t-1 ; t-2*)

Finally, Scenario 3 is implemented to fully exploit the potential of ML methods. This subset incorporates additional information on school attendance (type of primary and lower secondary school, evening school, and type of attendance) as well as data from previous years.

Table 2 provides a synthetic overview of the three scenarios.

In the three scenarios, nominal categorical predictors are converted into dummy variables (in R factors), and missing data are replaced with an “out-of-range” value, namely 99 or “99”.

A comparison of the results across the different scenarios (Table 3) shows that, as the amount of available information increases (from Scenario 1 to Scenario 3), overall prediction performance improves accordingly, with Scenario 3 yielding the best results.

Table 3 – *Summary statistics of differences between % relative frequency distribution obtained by RF model and the % relative frequency distribution of benchmark distribution, separately by subpopulations.*

	Overall	Pop A	Pop B	Pop C
Augmented data – scenario 1				
Sum of absolute values	0.81%	1.92%	0.79%	1.74%
Mean of absolute values	0.10%	0.24%	0.10%	0.22%
Augmented data – scenario 2				
Sum of absolute values	0.76%	1.90%	0.77%	1.97%
Mean of absolute values	0.09%	0.24%	0.10%	0.25%
Augmented data – scenario 3				
Sum of absolute values	0.69%	1.79%	0.78%	2.80%
Mean of absolute values	0.09%	0.22%	0.10%	0.35%
Augmented data – scenario 3 **Fixed variables**				
Sum of absolute values	0.82%	1.79%	0.84%	1.55%
Mean of absolute values	0.10%	0.22%	0.10%	0.19%

This trend is evident when focusing on population A, characterized by more informative profiles. For population B, the predictive performance of the RF remains stable across the three scenarios. In contrast, for population C, the trend is the opposite of that observed for population A—that is, prediction performance worsens as the amount of information increases. This occurs because the third scenario introduces additional variables that are informative only for population A and B. In RF, the trees are built by randomly selecting a subset of predictors from the full set (which is larger in Scenario 3) to evaluate at each split. As a result, the probability of selecting the informative predictors—those useful for estimating population C based only on demographic information—decreases from Scenario 1 to Scenario 3.

Indeed, the tuning across the three scenarios selects a high number of predictors: in all cases, the *mtry* value is very close to the total number of covariates, and the leaf size is very small. This indicates that the tuned RF trees are very deep and make use of nearly all the available covariates.

In the latest experiment, demographic information was constrained to be included at every split. The results of the RF model built in this way are comparable to those obtained in the three scenarios, with the best performance in predicting population C. However, in this case, the *mtry* value is much lower than the total number of predictors—approximately equal to the square root of the number of predictors.

5. Discussion

In this work, we showed how the use of ML technique can help for imputation purposes. We proposed to impute the ALE for units beyond MS using RF model.

The evaluation of the results was carried out at a macro level, using as a benchmark the distribution derived from the augmented MS with sampling weights.

When comparing ALE predictions from the official LL procedure with those generated by the RF model—based on their divergence from the benchmark distribution—the RF model shows superior performance, particularly when trained on the augmented sample.

In the second part of the analysis, we explored the possibility of incorporating all available information into the model. ML models generally offer the advantage of handling a high number of input features, including fully disaggregated categorical variables. This flexibility represents a key strength of ML approaches compared to traditional statistical models.

We designed three scenarios to evaluate the contribution of additional information and to assess the ability of the RF model to leverage this information across three subpopulations, each characterized by distinct information profiles. Overall, the results show a clear trend: the difference from the benchmark distribution decreases with a progressively closer alignment as we move from the least informative to the most informative scenario.

However, the improvement manifests differently across subpopulations: there is a clear gain for the population with more populated regressors (subpopulation A), relative stability for the group with moderately informative profiles (subpopulation B), and a marked deterioration for the subpopulation with only demographic information (subpopulation C). In fact, the increase in the number of regressors makes the selection of the informative predictors for ALE in subpopulation C less likely during the tree-splitting process. Even for subpopulation C, the results are comparable to the overall ones, because the increase in the number of regressors is balanced by the hyperparameter *mtry*, which is appropriately tuned.

Usually, the suggested default value for *mtry* is close to the square root of the number of regressors (see, for example, Breiman, 2002). However, as the author suggested, when very noise regressors are present, the *mtry* parameter should be higher. In our application, the patterns of available information differ substantially across subpopulations. For sub-population C only three covariates are informative while the remaining regressors are essentially noise. For this reason, the tuning process indicates that the optimal value of *mtry* is close to the total number of regressors. However, if the trees are forced to consider, at each split, the non-missing regressors of subpopulation C among the randomly selected candidates, then the overall prediction performance becomes comparable to that of the other RF. In this case, prediction for subpopulation C improves significantly, with the optimal *mtry* value close to the square root of the number of regressors.

In conclusion, this research shows that the suggested value of *mtry* cannot be applied without taking into account the structural characteristics of the data. While it is a good choice when information is uniformly distributed across the population and no noisy regressors are present, it becomes inadequate in scenarios characterized by heterogeneous information patterns and the presence of noisy regressors, as in the case of our study.

The flexibility of the RF structure, which can adapt to different prediction scenarios through the specification of different hyperparameters, makes the tuning phase essential and allows for the semi-automatic handling of prediction problems in presence of highly heterogeneous populations.

In general, when the population of interest is highly heterogeneous in terms of available information, ML provides a valuable tool, i.e., the ensemble model. The idea of an ensemble model is to combine different ML techniques, each one specifically suited for a subpopulation, to obtain a final prediction for each unit. This approach can be useful in our application. As a further development, we plan to implement additional ML models, better suited for subpopulations B and C, such as XGBoost (Chen and Guestrin, 2016), and Neural Network (Rumelhart *et al*., 1986). Finally, we aim to combine these different ML models into a single ensemble model (see, for example, Dietterich, 2000).

References

BREIMAN L. 2001. Random forests, *Machine learning*, Vol. 45, pp. 5–32.

BREIMAN L. 2002. *Manual On Setting Up, Using, And Understanding Random Forests V3.1.*

CHEN T., GUESTRIN, C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp. 785–794.

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*, Vol. 38, No. 2, pp. 637-646.

DIETTERICH T.G. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, Vol. 1857. Springer, Berlin, Heidelberg.

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*, Vol. 2, No. 3, pp. 143-174.

LITTLE R.J.A, RUBIN D.B. 2002. *Statistical Analysis with Missing Data*. Wiley, New York.

PROBST P., WRIGHT M., BOULESTEIX A. 2018. Hyperparameters and Tuning Strategies for Random Forest, *Data Mining and Knowledge Discovery*.

PROBST P., BOULESTEIX A., BISCHL B. 2019 Tunability: Importance of Hyperparameters of Machine Learning Algorithms, *Journal of Machine Learning Research*, Vol. 20, pp. 1-32.

RUMELHART D. E., HINTON G. E., WILLIAMS R. J. 1986. Learning representations by back-propagating errors. *Nature*, Vol. 323, No. 6088, pp. 533–536.

WRIGHT M.N., ZIEGLER A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, *Journal of Statistical Software*, Vol. 77, No. 1, pp. 1–17.

Ilaria BOMBELLI, Istat, ilaria.bombelli@istat.it
Romina FILIPPINI, Istat, filippini@istat.it
Simona TOTI, Istat, toti@istat.it