Applying Machine Learning to Longitudinal Administrative Data: A Case Study in Education

De Fausti Fabrizio[1], Di Zio Marco[1], Filippini Romina[1], Simona Toti[1]

[1] *Italian National Institute of Statistics (Istat), Italy*

Abstract

The paper shows the results of an empirical evaluation about the use of Random Forest, Recurrent Neural Network and Long Short-Term Memory for administrative longitudinal data. Those machine learning methods are used for the prediction of the attained level of education for the Italian population with respect to the subset of units of the Italian register of individuals covered by administrative sources. The assessment is made by comparing the results of the predictions with the observed data available in the administrative sources and is carried out by looking at the distributional preservation and the prediction of each single unit, that is from a macro and micro perspective.

Keywords: mass imputation, recurrent neural network, long short-term memory, random forest.

1. Introduction

The increasing availability of administrative sources has significantly changed the official statistical production system, moving towards a register-based approach. Advantages are expected in terms of reduction of costs and of response burden, and the possibility of having micro data enhancing the production of detailed statistics. However, some issues should be dealt with when using administrative sources , typically delays in data availability and coverage problems because their target population may differ from the statistical population of interest. In this context, the production of a complete and coherent dataset becomes a crucial activity, making necessary the application of various procedures for estimating delayed and missing data. An important peculiarity is that, once the data time-lag is overcome, updated administrative information becomes available, providing the opportunity for an evaluation and a refinement of the statistical procedures used to transform administrative data to meet our statistical interests. An important subject for which administrative information is available is the *Education* of people. In the Italian National Institute of Statistics (Istat), information on students, school attendance and educational level (ALE) are available from the Ministry of Education since 2011. To obtain a yearly estimate of the ALE for the Italian resident population, Istat has adopted a mass imputation approach that integrates administrative, survey and the 2011 census data (Di Zio et al., 2019). To fully leverage the opportunities presented by longitudinal administrative sources and the potentiality of machine learning methods (De Fausti et al., 2022), we study Random Forest (RF), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) to predict the ALE. Section 2 and 3 report details on the data and the methods used. Section 4 describes the study and the results obtained. Some conclusions are discussed in Section 5.

2. The use case: administrative data for the ALE

The strategy behind the new population census relies on the integration of the information stored into registers with data specifically collected through a sample survey. In this context, the Base Register of Individuals (BRI) is the basis of the new population census, referred to as the Permanent Census. The register is constructed at individual record level, primarily through extensive integration of administrative data. The set of variables derived from administrative sources, named register variables, includes demographic information, such as gender, age, marital status, place and date of birth and citizenship that are mainly derived from demographic sources. The BRI, however, does not contain all data concerning information traditionally collected by the census, as for instance the ALE. Hence, to generate thematic census outputs, a survey has been conducted to complement and enhance the coverage and quality of the existing register data. Specifically, respondents are asked about their educational level. The high amount of available administrative information on this topic, in particular longitudinal information, may allow the production of ALE figures from registers as well as from the census survey.

The primary sources of administrative information on ALE originate from the Ministry of Education, Universities, and Research (MIUR). MIUR's administrative data pertain to individuals enrolled in a school course from 2011 onward. For this subset of the population, MIUR provides information regarding the ALE and their course attendance (e.g., attending the first year of primary education). However, MIUR data have some informational gaps: they only cover individuals entering a study program after the 2011, they trace only students enrolled in an educational course held in Italy, and do not include qualification courses like Fine Arts, Drama, Dance and Music academic diplomas, as well as other training and vocational careers managed by Italian Regions that are not required to provide data to MIUR. The main consequence is a potential underestimation of ALE in the administrative source. Another critical issue is the timeliness. MIUR data are typically available with a delay of 1 or 2 years compared to the BRI reference time.

Due to the complexity and heterogeneity of the available information, the official procedure for estimating ALE relies on different imputation procedures, which are combined to address sub-populations characterized by varying amounts of information (Di Zio et al., 2019). Specifically, the presence or absence of information on ALE in administrative sources determines the partitioning of the target population into two main subgroups: for the subset of individuals with available information from MIUR, ALE at time *t* is predicted using time-lagged data; for the remaining individuals, ALE from the survey is used as response variable. The general idea is to estimate a model for the prediction of ALE given the values of known covariates *X*. Considering $I^{(t)}$ as the target variable (i.e., ALE at time *t*), the conditional probabilities $h(I^{(t)} |X)$ is estimated and then I_t is imputed by randomly selecting a value from this distribution.

In this study, the focus is on the sub-population where longitudinal administrative information is available, comprising approximately 22% of the overall population. The conditional probabilities $h(I^{(t)} |X)$ are estimated by means of different machine learning models that leverage the longitudinal information available from administrative sources.

Once the data time-lag has elapsed, updated administrative data become available. In 2023, administrative information pertaining to 2021 are available and can be regarded as the gold standard for evaluating results generated by the various machine learning models. This enables a more accurate evaluation and potentially a fine-tuning of the procedure.

3. Machine learning with longitudinal data

In recent years, machine learning methods for classification and prediction have emerged as an alternative to the classical statistical approach. In addition to classification problems that statically associate input with output, within ML it is possible to model time using so-called dynamic classifiers.

Recurrent Neural Network (RNN), where a recurrent connection is introduced in the network, is a possible approach to dynamic ML (De Mulder, 2015, Sherstinsky, 2020). Each neuron in an RNN maintains a hidden state, which captures information about previous inputs in the sequence and influences future predictions. This capability makes RNNs particularly effective for tasks involving sequential data processing, such as language modelling, speech recognition, and time series prediction. If the series are more than 10 time points, RNN can suffer from difficulty in capturing long-term dependencies. In addition, RNN is vulnerable to vanishing or exploding gradient problems.

The Long Short-Term Memory (LSTM) is a specialized type of recurrent neural network architecture designed to address the limitations of RNNs in capturing long-range dependencies and mitigating the vanishing gradient problem. LSTM units contain a set of gates (input, forget, output) that regulate the flow of information through the network, enabling better preservation of relevant information over long sequences (Hochreiter *et al.*, 1997).

Finally, we consider the Random Forest (RF) (Breiman, 2001) which is an ensemble learning method used for classification, regression, and other tasks. RF do not explicitly account for time, it operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. RF are robust against overfitting, exhibit high accuracy, and are relatively resistant to noise and outliers in the data.

4. Real data application

The dataset for the application consists of approximately 480 thousand individuals, aged 9 or older, residing in the Emilia Romagna region (NUTS 2) in 2021. It includes complete longitudinal administrative information on ALE from 2015 to 2021, classified into 7 modalities.

The training set for the three ML methods is composed by the subset of individuals with complete longitudinal information from 2015 to 2020 (*t*-1). The covariates considered in the model are the school enrolment up to 2020, as well as demographic information such as gender, age and province of residence. The estimated conditional probabilities for the ALE at 2020 are then applied to one-year-forward shifted data (test set), and ALE in 2021 (*t*) for each unit in the dataset is obtained by randomly selecting a value from the estimated ALE probability distribution.

The estimated ALE ($\hat{I}$) is compared with the ALE from MIUR (*I*) (considered as the target ALE value). Table 1 shows the estimates obtained with RNN, LSTM and RF compared with the ALE *I* observed in MIUR data. Since the predictions are obtained through a random draw from the estimated distributions, the procedures are repeated 10 times to take into account the variability, and the results are computed averaging over those repetitions.

Table 1: Estimated absolute (a.v.) and percentage (%) values of ALE distribution through RNN, LSTM, RF and administrative ALE distribution in 2021.

ALE	RNN		LSTM		RF		TRUE	
	a.v.	%	a.v.	%	a.v.	%	a.v.	%
Primary education	120,087	25.3	120,065	25.2	120,335	25.3	120,589	25.4
Lower secondary ed.	199,984	42.1	200,068	42.0	200,012	42.1	200,364	42.1
Upper secondary ed.	109,082	22.9	109,829	23.1	109,164	23.0	109,134	23.0
Bachelor’s degree	31,470	6.6	31,251	6.6	31,144	6.6	30,731	6.5
Master degree	14,594	3.1	14,001	3.0	14,514	3.1	14,410	3.0

PhD	214	0.0	218	0.0	210	0.0	225	0.0
Total	**475,431**	**100.0**	**475,432**	**100.0**	**475,378**	**100.0**	**475,453**	**100.0**

We remark that only modalities referring to an acquired educational level are reported, the class “no educational attainment” represents a residual class and is not included in the table, and this accounts for the discrepancy in the total amounts.

All the methods provide an estimated distribution that is close to the target ALE distribution. Relative errors *RR_i* for each ALE modality *i* is computed:

$$RR_i = \frac{\left(\hat{I}_i - I_i\right)}{I_i}$$

where I_i is the true absolute frequency of modality i and $\hat{I}_i$ is the corresponding estimated value.

Table 2 shows the mean of the relative errors *$m(RR_i)$* computed over the 10 repetitions.

RF yields better results: the mean of the relative errors *$m(RR_i)$* is the lowest among the three methods. Moreover, the standard deviation is lower for almost all the ALE modalities. We notice high values of RR_i corresponding to the PhD class, this is due to the very low frequency of this modality.

Table 2: Mean relative error *$m(RR_i)$* and standard deviation (std) computed over 10 runs for RF, RNN, LSTM.

	RF		**RNN**		**LSTM**	
ALE	**$m(RR_i)$**	**(std)**	**$m(RR_i)$**	**(std)**	**$m(RR_i)$**	**(std)**
Primary education	0.235	(0.132)	0.416	(0.039)	0.435	(0.032)
Lower secondary ed.	0.176	(0.110)	0.190	(0.080)	0.151	(0.071)
Upper secondary ed.	0.097	(0.059)	0.332	(0.299)	0.719	(0.561)
Bachelor’s degree	1.343	(0.417)	2.404	(1.636)	1.996	(1.689)
Master degree	0.874	(0.429)	3.695	(1.944)	3.278	(1.380)
PhD	6.844	(4.529)	6.000	(7.219)	7.244	(4.936)
Mean	**1.595**		**2.173**		**2.304**	

Table 3 shows the f1 score for each ALE class and the global f1 score. The f1 score is a measure of predictive performance of a classifier and it is the harmonic mean of the precision and recall. It thus symmetrically represents both precision and recall in one metric. The highest possible value is 1.0, indicating perfect precision and recall, and the lowest possible value is

0, if either precision or recall are zero. The calculation of the f1 score requires that a positive class of interest and negative complementary classes be defined. The f1 score is given by the formula f1=TP/(TP+0.5(TP+TN)) where TP is the number of correctly predicted instances as belonging to the positive class, FP is the number of instances erroneously predicted as positive and FN which is the number of instances erroneously predicted as negative. The calculation of the global f1 score (generally named micro f1) requires calculating aggregate TP, FP, FN , adding the partial TP, FP, FN obtained by defining each time one of the 7 classes as positive. Those measures are computed to assess the methods with respect to the prediction for each units, that is a micro-level evaluation. We notice a behaviour opposite to the evaluation of distribution preservation (Table 2), that is the macro-level assessment. LSTM outperforms the others, with a preference increasing with the highest ALE classes, in fact in the modalities until "upper secondary education" the f1 scores are close each other, for the classes "bachelor and master degree" there is a more sensible difference, and finally the f1 score for 'PhD' is very different from the others. There is a strange low value of the indicator for RNN, some more analysis should be performed in order to understand the reason. On the other hand, this class is characterised by a low frequency of observations, and probably this is the reason because the global f1 score of RNN is not affected by this value.

Table 3: F1 score computed over 10 runs.

ALE	RF	RNN	LSTM
Primary education	0.9950	0.9965	0.9966
Lower secondary ed.	0.9889	0.9918	0.9920
Upper secondary ed.	0.9138	0.9265	0.9286
Bachelor's degree	0.6910	0.7216	0.7340
Master degree	0.7063	0.7353	0.7613
PhD	0.6305	0.1625	0.8484
Global f1	**0.9450**	**0.9521**	**0.9548**

The code is developed in the Python language using the machine learning library scikit-learn (Pedregosa, 2011) to implement Random Forests, while we used the library keras (Chollet, F. 2015) for the deep learning algorithms RNN and LSTM. For further analysis, confusion matrices are reported in Appendix.

5. Conclusions

The paper shows the results of an empirical evaluation about the use of random forest, RNN and LSTM for longitudinal data. Those methods are applied to the prediction of the attained level of education for the Italian population with respect to the subset of units of the Italian register of individuals covered by administrative sources. The results show that random forest is preferable to the other methods when distributional accuracy is our main interest. On the other hand, when the interest is in the micro-prediction, the preference is not so evident even if LSTM has a slight better performance. It should be noticed that, the data used are not characterised by a long series, and this is certainly an element to take into account in the interpretation and generalisation of the results. The good behaviour of random forests, that are not introduced in literature mainly for dealing with the time dimension of the data, could be explained by the fact they – at least as used in this paper – estimate the conditional distribution without any particular constraint used for taking into account the time. This is certainly a problem when a long time series is the input of our case and in this situation models as RNN and LSTM, specifically introduced for these problems, can be more efficient. Further studies will be devoted to the cases presenting those characteristics.

References

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Chollet, F. (2015) keras, GitHub. https://github.com/fchollet/keras

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.

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(2), 637-646.

De Mulder, Wim, Steven Bethard, and Marie-Francine Moens. "A survey on the application of recurrent neural networks to statistical language modeling." Computer Speech & Language 30.1 (2015): 61-98.

Hochreiter, Sepp, and Jürgen Schmidhuber (1997). Long short-term memory. Neural computation 9.8: 1735-1780.

Sherstinsky, Alex (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena 404: 132306.

Pedregosa et al.,(2011) Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830.

Appendix

Table A1: Confusion matrix (average over 10 runs): RF estimated ALE vs true ALE.

	True ALE (observed in MIUR)						
RF estimated ALE	**0 – no ed. attainment**	**1 - Primary education**	**2- Lower secondary ed.**	**3 - Upper secondary ed.**	**4 - Bachelor's degree**	**5 - Master degree**	**6 - PhD**
0 – no ed. attainment	0	0	0	0	0	0	0
1 - Primary ed.	61.4	119,865.1	660.7	1.8	0	0	0
2- Lower sec. ed.	13.7	460.9	197,963.7	1,891.1	18	16.6	0
3 - Upper sec. ed.	0	7.5	1,344.6	99,743.2	6,813.0	1,223.5	2.2
4 - Bachelor's degree	0	0.9	26.7	6,336.9	21,378.6	2,982.1	5.8
5 - Master degree	0	0.2	16.3	1,187.9	2,926.8	10,214.2	64.6
6 - PhD	0	0	0	3.4	7.4	77.2	137.0

Table A2: Confusion matrix (average over 10 runs): RNN estimated ALE vs true ALE.

	True ALE (observed in MIUR)						
RNN estimated ALE	**0 – no ed. attainment**	**1 - Primary education**	**2- Lower secondary ed.**	**3 - Upper secondary ed.**	**4 - Bachelor's degree**	**5 - Master degree**	**6 - PhD**
0 – no ed. attainment	0	0	0	0	0	0	0
1 - Primary ed.	20.0	119,912.3	649.7	3.1	1.2	1.1	1.6
2- Lower sec. ed.	0.8	166.4	198,540.5	1,647.2	2.4	3	3.7
3 - Upper sec. ed.	0.9	6.8	788.9	101,088.3	6,285.1	962.2	1.8
4 - Bachelor's degree	0.6	0.5	2	5,493.6	22,441.9	2,778.5	13.9
5 - Master degree	0.1	0.5	2	849.4	2,736.6	10,663.8	157.6
6 - PhD	0	0.3	0.7	0	2.5	185.8	35.7

Table A3: Confusion matrix (average over 10 runs): LSTM estimated ALE vs true ALE.

	True ALE (observed in MIUR)						
LSTM estimated ALE	**0 – no ed. attainment**	**1 - Primary education**	**2- Lower secondary ed.**	**3 - Upper secondary ed.**	**4 - Bachelor's degree**	**5 - Master degree**	**6 - PhD**
0 – no ed. attainment	0	0	0	0	0	0	0
1 - Primary ed.	21.3	119,923	644.1	0.5	0	0.1	0
2- Lower sec. ed.	0.1	140.9	198,609.6	1,612.4	0.8	0.2	0
3 - Upper sec. ed.	0.1	0.8	811.8	101,664.4	5,813.2	843.7	0
4 - Bachelor's degree	0	0.1	1.6	5,677.2	22,747.3	2,304.8	0
5 - Master degree	0	0.2	0.7	874.7	2,689.3	10,815.3	29.8
6 - PhD	0	0	0	0	0	37.3	187.7