EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

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*

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed by the European Union

Introduction and Motivation

The **availability of administrative sources** moving NSI towards a **register-based** approach:

- reduction of costs
- response burden
- micro data enhancing the production of detailed statistics

Issues using administrative sources:

- delays in data availability
- coverage problems

Introduction and Motivation

Machine learning may be useful approach

- Prediction tasks are important to produce a **complete and coherent** dataset and impute missing information
- We could leverage the **longitudinal structure** of administrative data.
- Generally a **high amount** of data

A relevant real case: The Attained Level of Education

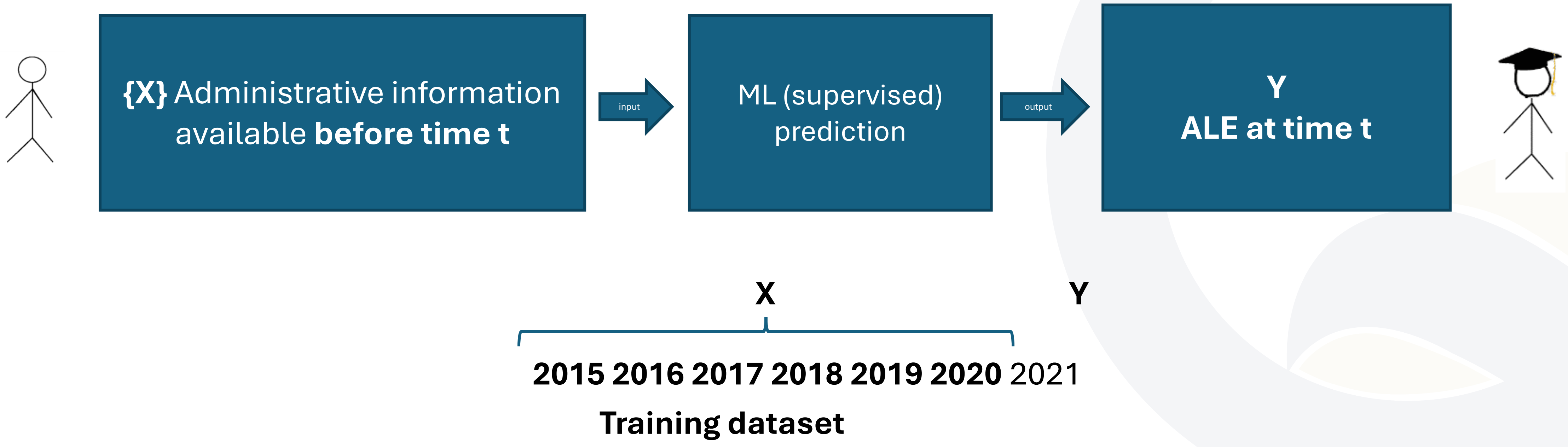
Administrative information for ***Education*, provided from Ministry of Education (MIUR)**

- Involve a **relevant subset** of Italian population
- people entering a study program **after 2011**
- Info on:
    - attained level of education (ALE)
    - course attendance
    - school characteristics
    - some demographical info (age, gender, …)

- **Not include** qualification courses like Fine Arts, Drama, Dance and Music academic diplomas
- **Time-lag.** It is available generally from 2011 to t-1, scholar year (t-2.t-1).

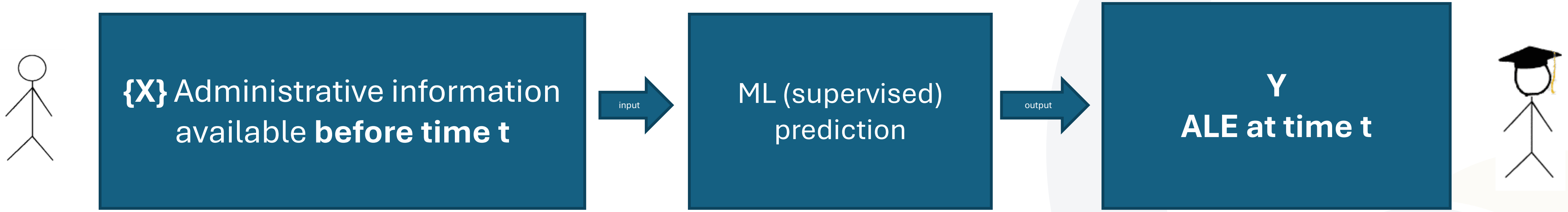
A relevant real case: The Attained Level of Education

Given the high informative power, a prediction approach is adopted by Istat to fill the time-lag and producing **estimates of ALE at time t** to enrich the RBI



A relevant real case: The Attained Level of Education

Given the high informative power, a prediction approach is adopted by Istat to fill the time-lag and producing **estimates of ALE at time t** to enrich the RBI



X **Y**

2015 **2016 2017 2018 2019 2020 2021**

Test dataset

Machine learning for exploiting longitudinal data

- **Random forest (RF).** Ensemble learning method that combines **multiple decision trees** to improve predictive accuracy and **control over-fitting**. Each tree is built on a random subset of data and features, making the ensemble **robust against noise**.

- **Recurrent Neural Network (RNN).** A class of neural networks **designed for processing sequential data**, where current inputs are dependent on previous status. Utilizes feedback loops to process sequences of data. Each neuron in an RNN maintains a hidden state, which captures information about previous inputs in the sequence and influences future predictions. Capable of learning and **remembering patterns over time**, **no long sequences.**

- **Long short-term memory (LSTM).** An advanced type of RNN, specifically designed to **avoid the long-term dependency problem,** which causes standard RNNs to forget input information over time. Highly effective for long sequence data where the relationship spans many time steps.

Experimental study

- **Data.** Emilia Romagna region (NUTS 2), aged 9 or older, observed from 2015 to 2021.Variables**.** Demographic (gender, age, citizenship), yearly educational attainment and school enrolment data up to 2020.
- **Models** RF, RNN, LSTM
- **Training data.** Data from 2015 to 2020 completely observed (no missing data).
- **Test data.** Predicting values for ALE^{2021} on data 2016-2021
- **Maintain the variability,** the prediction is obtained by a random draw from ML distribution of probability output
- **Compare the results** of estimated ALE with that at 2021 from admin data (over 100 repetitions)
- **Indicators**. Relative error of frequencies (RR) for each ALE modality (evaluation of aggregates) F1-score (evaluation of predictions)

Results

Estimated ALE distribution computed over 100 run through RF, RNN and LSTM: absolute values (a.v.), standard deviation (std) and percentage values (%). Administrative ALE distribution in 2021 (TRUE).

ALE	RF			RNN			LSTM			TRUE	
	a.v.	(std)	%	a.v.	(std)	%	a.v.	(std)	%	a.v.	%
Primary education	120,332	(202)	25.3	120,077	(46)	25.3	120,073	(52)	25.3	**120,589**	25.4
Lower secondary ed.	200,034	(326)	42.1	200,029	(171)	42.1	200,058	(173)	42.1	**200,364**	42.1
Upper secondary ed.	109,106	(260)	22.9	109,812	(714)	23.1	109,739	(629)	23.1	**109,134**	23.0
Bachelor's degree	31,149	(185)	6.6	30,978	(699)	6.5	31,107	(747)	6.5	**30,731**	6.5
Master degree	14,510	(115)	3.1	14,313	(422)	3.0	14,234	(473)	3.0	**14,410**	3.0
PhD	207	(11)	0.0	219	(25)	0.0	219	(24)	0.0	**225**	0.0
Total	475,338		100.0	475,428		100.0	475,430		100.0	**475,453**	100.0

Results

Mean percentage relative error $m(RR_i)$ and standard deviation (std) computed over 100 runs for RF, RNN, LSTM.

ALE	RF		RNN		LSTM	
	$m(RR_i)$	(std)	$m(RR_i)$	(std)	$m(RR_i)$	(std)
Primary education	0.237	(0.132)	0.425	(0.038)	0.428	(0.043)
Lower secondary ed.	0.199	(0.118)	0.170	(0.079)	0.157	(0.078)
Upper secondary ed.	0.181	(0.156)	0.721	(0.541)	0.648	(0.466)
Bachelor’s degree	1.374	(0.569)	1.884	(1.499)	2.212	(1.578)
Master degree	0.848	(0.628)	2.379	(1.822)	2.864	(1.997)
PhD	8.249	(4.495)	8.951	(6.880)	8.378	(6.997)
Mean	**1.848**		2.422		2.448	

Results

F1 score computed over 100 runs

ALE	RF	RNN	LSTM
Primary education	0.9949	0.9965	0.9966
Lower secondary ed.	0.9888	0.9918	0.9919
Upper secondary ed.	0.9149	0.9275	0.9286
Bachelor’s degree	0.6932	0.723	0.7337
Master degree	0.7085	0.738	0.761
PhD	0.6454	**0.2267**	0.8435
Global f1	0.9453	0.9527	**0.9547**

Empirical evidences

- **Distributional accuracy.** RF has a better performance
- **Prediction accuracy.** RF, RNN, LSTM similar and good. LSTM is slightly better.
- Strange behavior of RNN in the PhD class

Future works and relevant issues

- Evaluation at finer geographical detail
- Imbalanced data. Improving methods and analysis
- Dealing with pattern of missing covariates
- Comparison with the current procedure to understand if the quality of ALE estimation can improve

THANK YOU FOR YOUR ATTENTION !

Fabrizio De Fausti
DEFAUSTI@ISTAT.IT