Timeliness and Accuracy with Machine Learning Algorithms: Early Estimates of the Industrial Turnover Index

UNECE Machine Learning for Official Statistics Workshop 2023

David Salgado

S.G. for Methodology and Sampling Design
Statistics Spain (INE)

June 5th, 2023

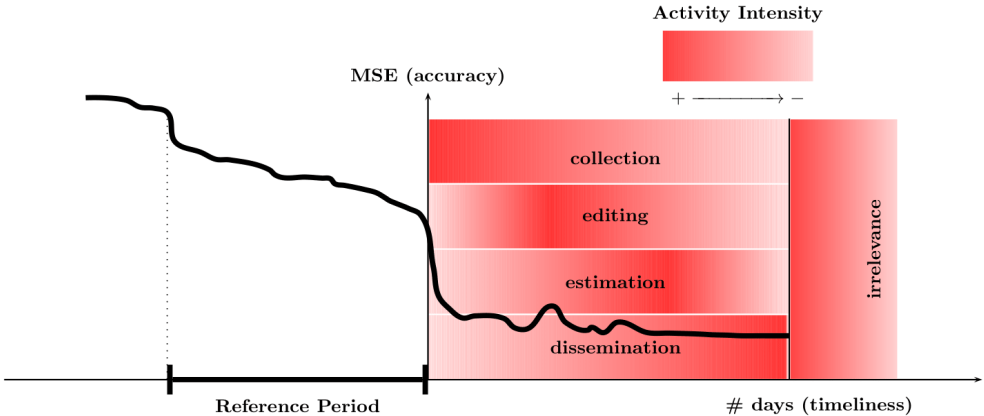
Instituto Nacional de Estadística

Overview

- General motivation
- The approach: main details
- The results
- Some lessons and conclusions

Timely or accurate? The production process

Main target: to improve **timeliness** under **accuracy** and **cost-efficiency bounds**



Timely or accurate? A critical view

- The pressing demand for timeliness has been put in the basis for the use of new data sources:
  - surveys are slow and expensive
  - digital data are fast and cheap
- The data deluge has favoured the delusion of accuracy

ServiceSectorIndex_202205_raw.txt

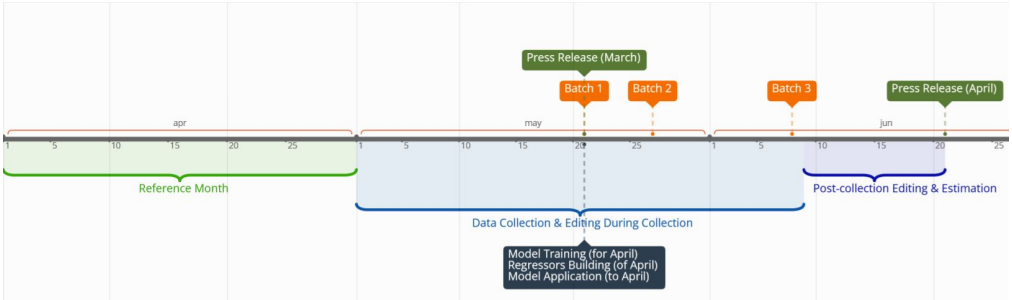
```
630044  varGestion@@147929252SS35.@@LI
630045  varGestion@@147930189SS35.@@LI
630046  varGestion@@147930266SS35.@@LI
630047  varGestion@@147930650SS35.@@LI
630048  varGestion@@147930918SS35.@@LI
630049  varGestion@@147931248SS35.@@LI
630050  varGestion@@147931987SS35.@@LI
630051  varGestion@@147932375SS35.@@LI
630052  varGestion@@147932761SS35.@@LI
630053  varGestion@@147933239SS35.@@LI
630054  varGestion@@147933406SS35.@@LI
630055  varGestion@@147933409SS35.@@LI
630056  varGestion@@147933897SS35.@@LI
630057  varGestion@@147933929SS35.@@LI
630058  This is not the economic reality
```

The key for connecting data to reality is on the **statistical methodology**

INE

The Spanish Industrial Turnover Index

- Short-term Business Statistics under European Regulation (STS)
- Monthly; around 12000 units per month
- Cut-off sampling + Fixed-base Laspeyres Index



- Subject matter experts receive data batches at $m + 20$, $m + 27$, and $m + 38$.
- First release at $m + 51$.

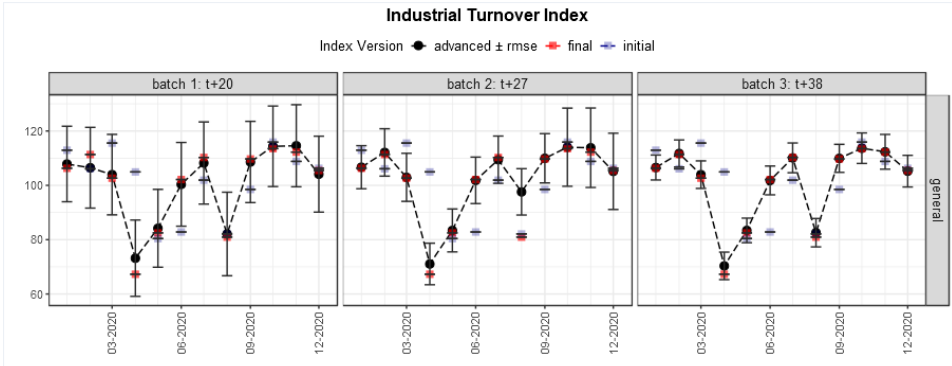
INE

The approach

Reconstruct microdata with predictions using past and on-course information:

r_t **subsample collected** up to time $t < t_{\text{release}}(m)$

$$Y_{U_d}^{(m)}(t) = \sum_{k \in r_{t,d}} y_{kt}^{(m,\text{ed})} + \sum_{k \in U_d - r_{t,d}} \hat{y}_{kt}^{(m,\text{val})}$$



Information Representation: Regressors

From **statistical variables**: Turnover + Geo Location + Economic activity

To **Regressors**:

- Geographical variables

  `code_NUTS2_ent_ed`, `code_NUTS3_ent_ed`, `code_LAU_ent_ed`...

- Time variables

  `year_ref`, `batch`, `nmonths`i`_imputd_xprt`...

- Economic activity variables

  `code_NACE2class_frame_ed`, `code_NACE2group_ed`...

- Target-Related Variables

  `trnovr_val_i`, `MA`i`_trnovr_val`, `q95_MA`i`trnovr_val_NACE2div`...

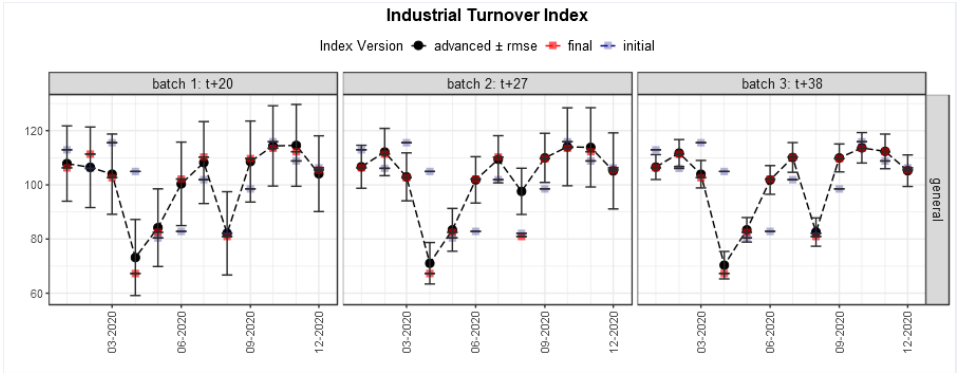
To **Encoded Regressors**: Dummy and Mean Encoding

INE

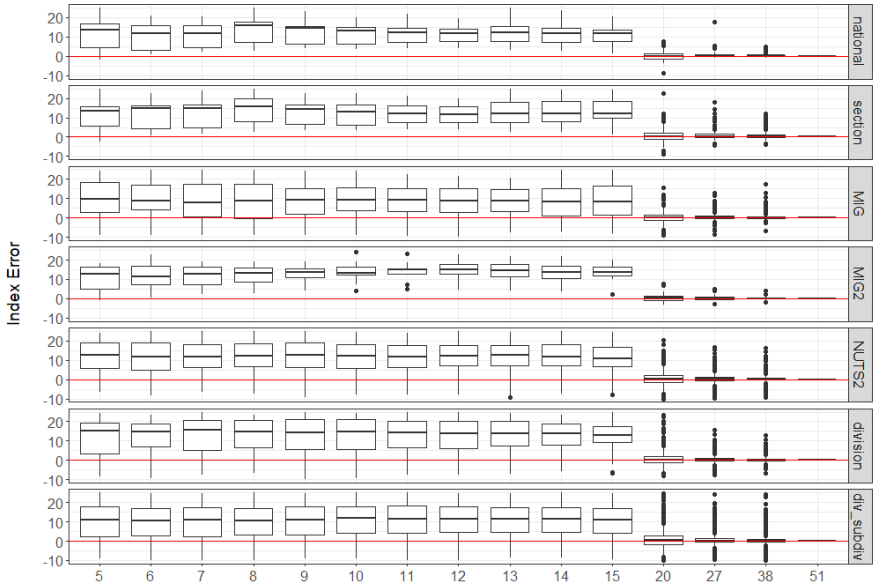
Information Representation: Regressors

q95_MA$i$trnovr_val_NACE2div	
Definition	Quantiles 0.95 of the variable MAi_trnovr_val across the population domain defined by edited values of variable code_NACE2div_ed (NACE Rev. 2 division) of the industrial establishment from the reference time period
Stat Type	Numerical
Values	$\mathbb{R}$, $i = 3, 6, 12$ (3 variables)
Example	150000
Source	Internal-Derived
Formula	$Q_{0.95}^{\text{NACE2div}}(\text{MA}i(z_k^{my,\text{val}}))$, where $\text{MA}i(z_k^{my,\text{val}}) = \frac{1}{i}\sum_{j=1}^{i} z_k^{(m-j)y,\text{val}}$
Stat Progr Ref	Spanish IOE-30052
Unit/Aggr	Aggr
Time Periods	$\{-1, \ldots, -i\}$
Long/Cross	Long + Cross
Cross-Domain Vars	code_NACE2div_ed
Encoding	-

Results: nowcasting



Results: timeliness vs. accuracy



Some lessons and conclusions

- Strategy:
  - Machine learning allows us to **improve statistical business functions**.
  - **Quality** can be improved: timeliness, accuracy, cost-efficiency, sound methodology,. . .
  - Constant **updating of synthetic microdata sets**: process refurbishing.
  - **Repository of regressors** (features).
- Methodology:
  - The core task is more about **information representation** than about statistical modelling.
  - **Combination** with existing methods is possible.
- Computational:
  - Traditional **technological infrastructure** is not enough.
  - **Data architecture** must be standard.
- Organizational:
  - A collaboration among **methodologists**, **computer scientists** and **subject matter experts** is necessary to integrate knowledge.