Improving statistical data editing with Machine Learning: first use cases in Statistics Spain (INE)

Sandra Barragán, David Salgado

Dept. Methodology and Development of Statistical Production

Statistics Spain (INE)

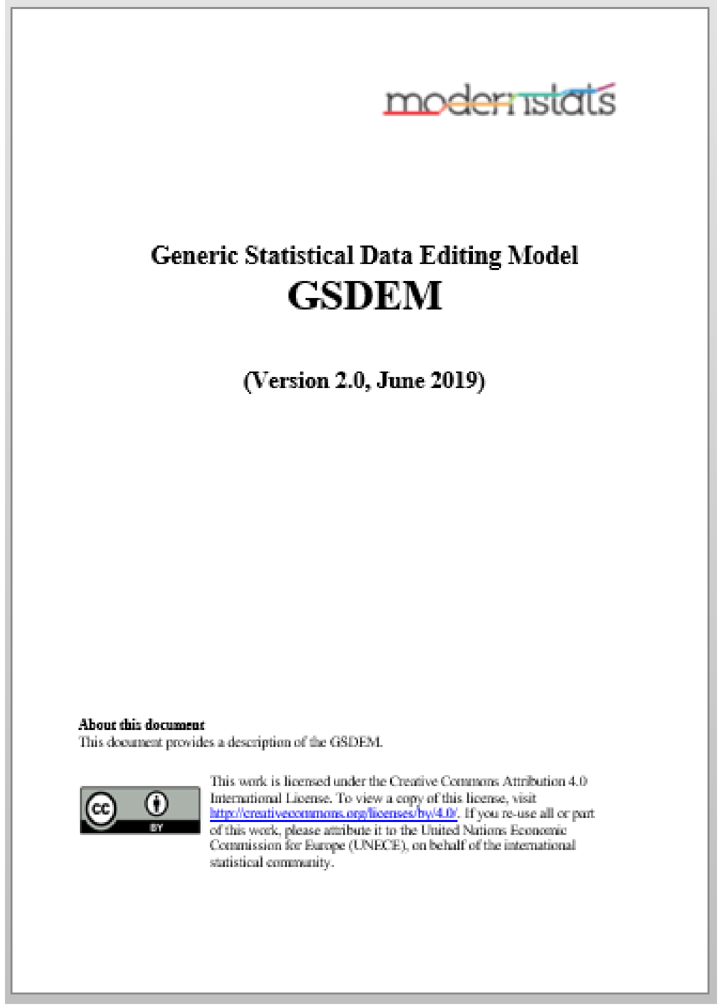
Statistical Data Editing 2022

October 5th, 2022

Outline

- **Context**
- Case 1: Local **scores**
- Case 2: **Semicontinuous** variables
- Case 3: **Imputation - Nowcasting**
- Case 4: **Imbalanced** data
- Case 5: Measure **Quality**
- Case 6: **NLP** questionaire comments
- **Conclusions**

Context


modernstats

Generic Statistical Data Editing Model

GSDEM

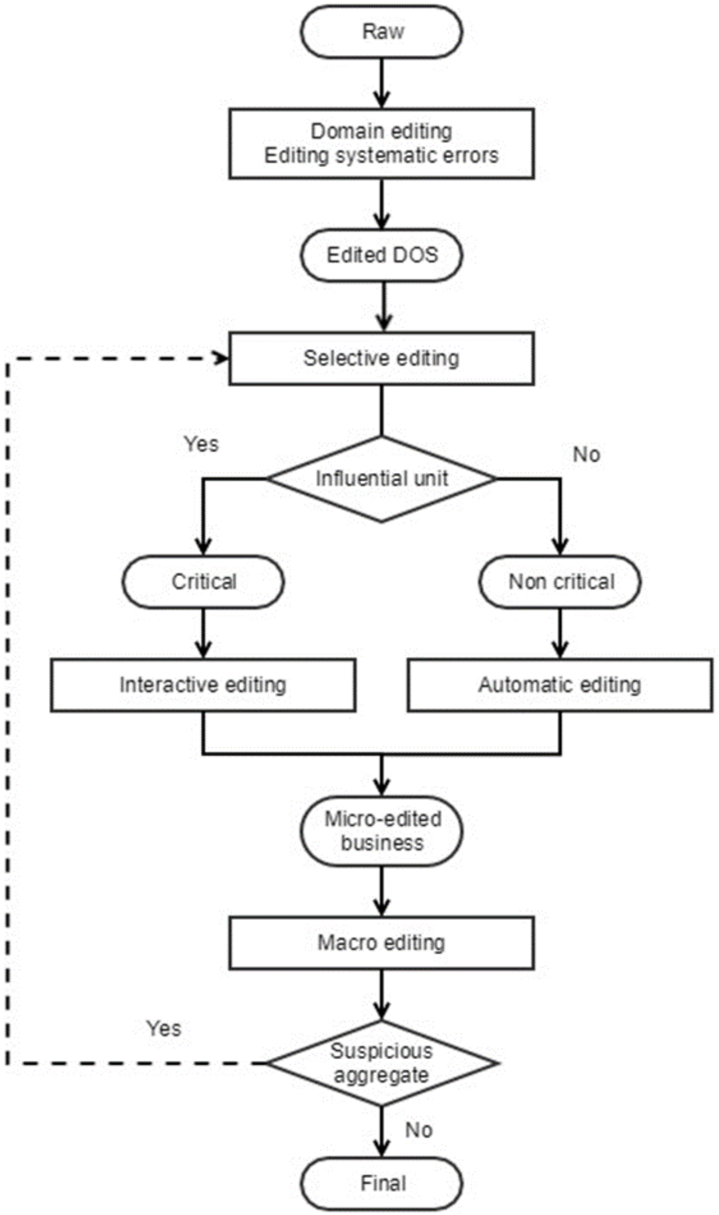
(Version 2.0, June 2019)

About this document

This document provides a description of the GSDEM.

Editing business functions

- **Review**
  - measuring the plausibility of values
  - assessing data for logical consistency
  - Units review (scores)
- **Selection**
  - Selection of **units**
  - Selection of **variables**
- **Treatment**
  - **Imputation** of variables
  - Treatment of **units**



UNECE modernstats by HLG - MOS

Case 1: Local Scores

- Traditionally $s_k = s(\hat{y}_k, y_k^{raw}) = d_k \cdot |y_k^{raw}\text{-}\hat{y}_k|$
- **Optimization** approach:

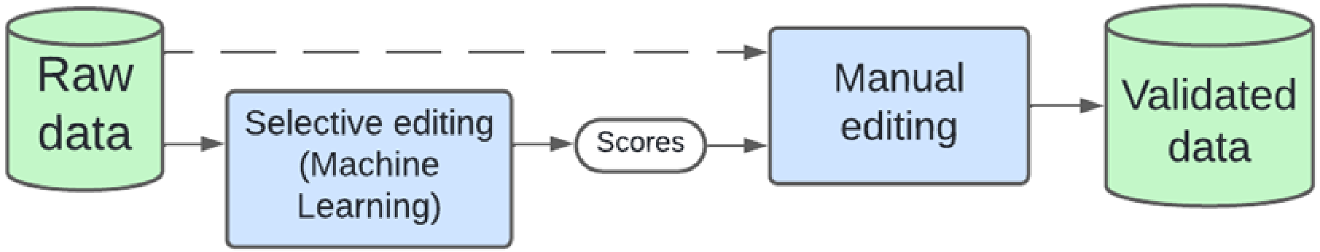
$$s_k = \mathbb{E}[d_k \cdot \underbrace{|Y_k^{raw}\text{-}Y_k^{true}|}_{\boxed{\textbf{target}}} \,|\underbrace{X_k}_{\boxed{\textbf{auxiliary information}}}] \longrightarrow$$ **Machine Learning** models

- Application to **categorical variables**:

$$s_k = d_k \cdot \mathbb{P}(\epsilon_k = 1|X_k)$$

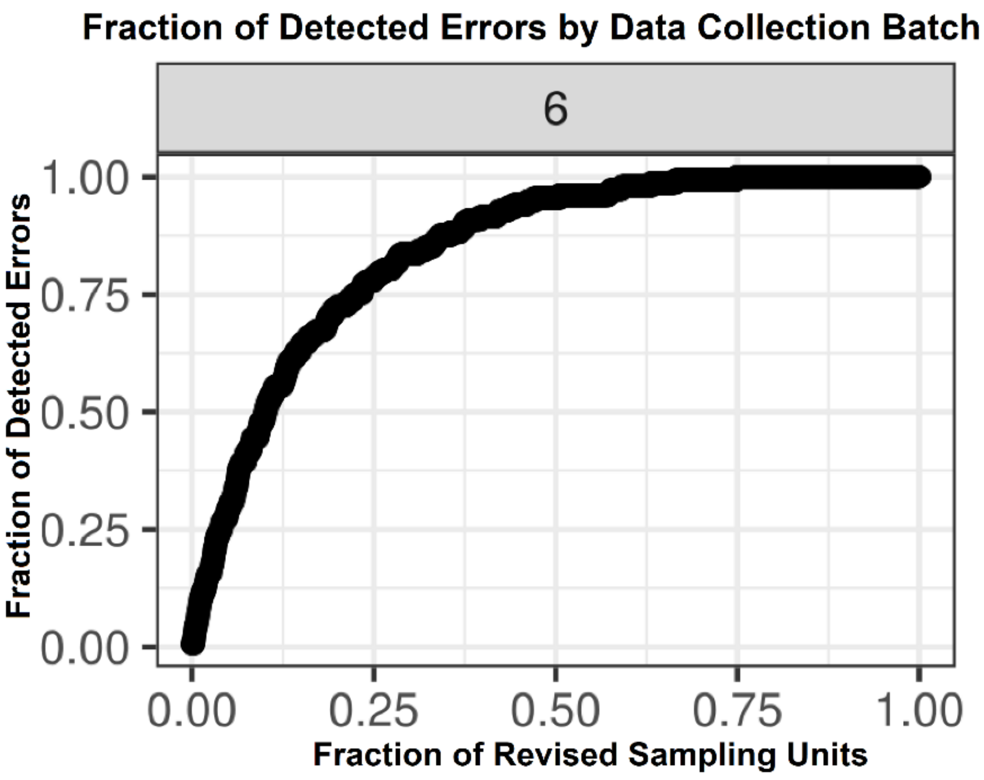
where ϵ_k=$|y_k^{raw}\text{-}y_k^{val}|$ is the **measurement error** (binary in categ. vars)

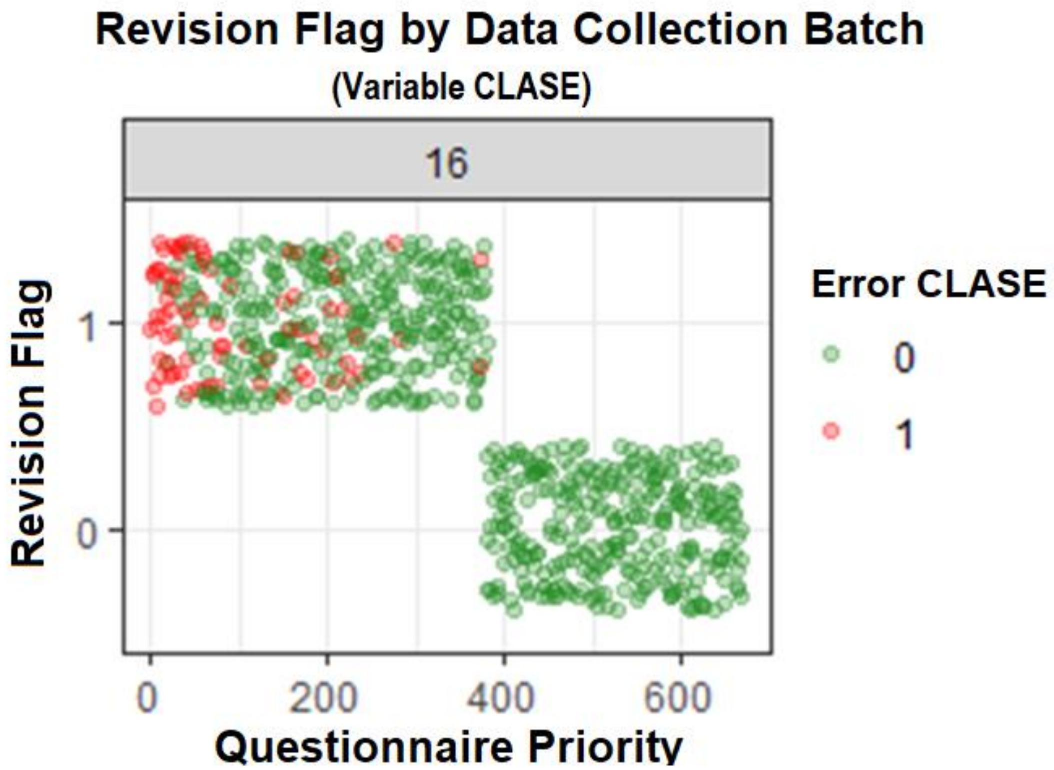
- **European Health Interview Survey** in Spain: occupation (CLASE).
  - Random Forest for Classification. Target: error indicator.



Case 1: Local Scores

- **First half** of the sorted sample already contained 75% of all measurement errors





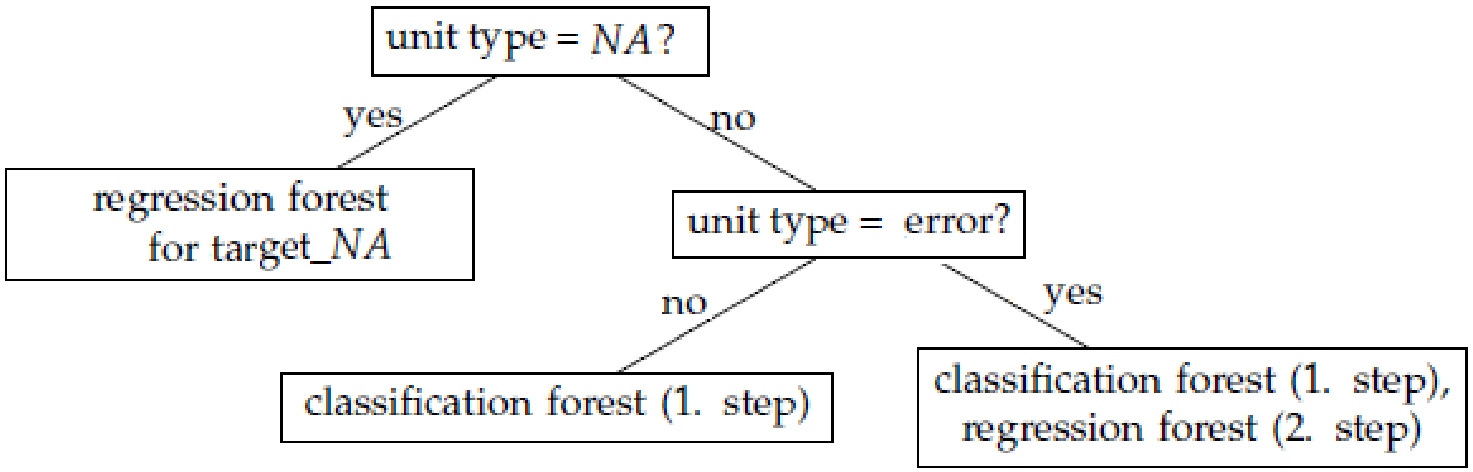
Case 2: Semicontinuous variables

- **Continuous** variable: y_k *(Services Sector Activity Indicators: turnover)*
- Target in the models:

  STEP 1 (RF classification): $I(\epsilon_k > 0)$ *(binary indicator of error)*

  STEP 2 (RF regression): $\epsilon_k = \left|Y_k^{raw} - Y_k^{val}\right|$ *(absolute error)*

- **Two-stage approach** to model semicontinuous variables including missing values:



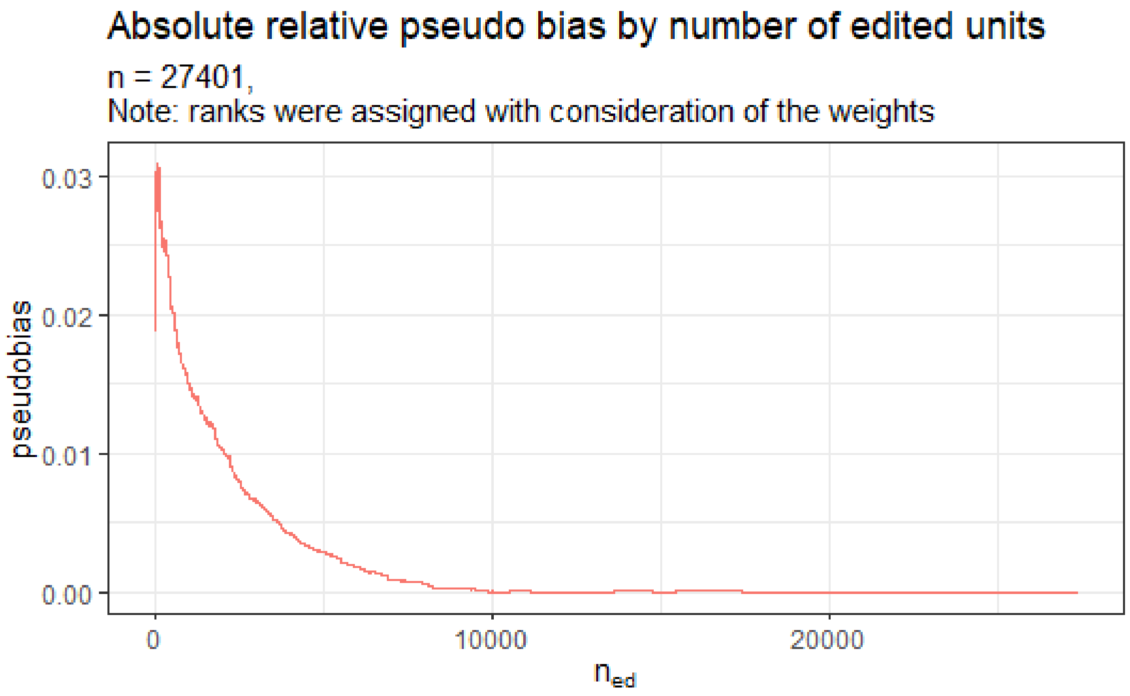
- ***Score* local**:

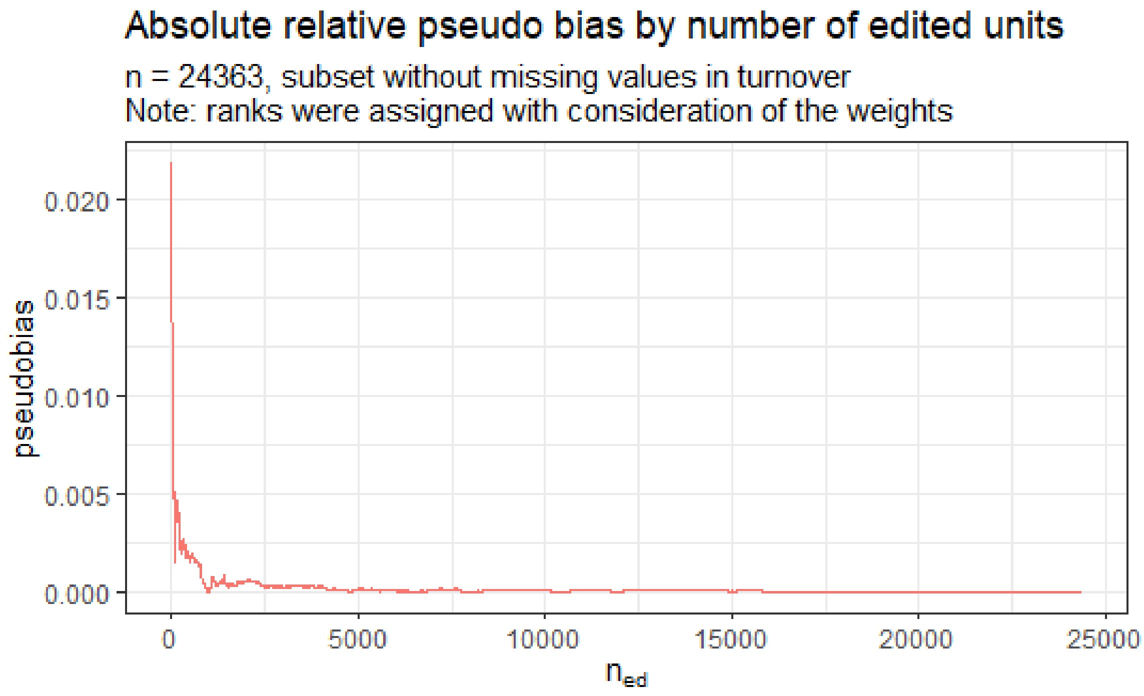
$$s_k = d_k \cdot \mathbb{P}(\epsilon_k > 0|X_k) \cdot \mathbb{E}[\epsilon_k|\epsilon_k > 0, X_k]$$

Case 2: Semicontinuous variables

- Relative pseudobias in absolute value:

$$ARB\left(\hat{Y}(n_{ed})\right) = \frac{|\hat{Y}(n_{ed}) - \hat{Y}^0|}{\hat{Y}^0}$$





Case 3: Imputation - Nowcasting

- **Early estimates of Spanish Industrial Turnover Index Survey**
- **Mass imputation** exercise over units not yet collected during the data collection
- **Gradient boosting** algorithm (lightgbm).

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

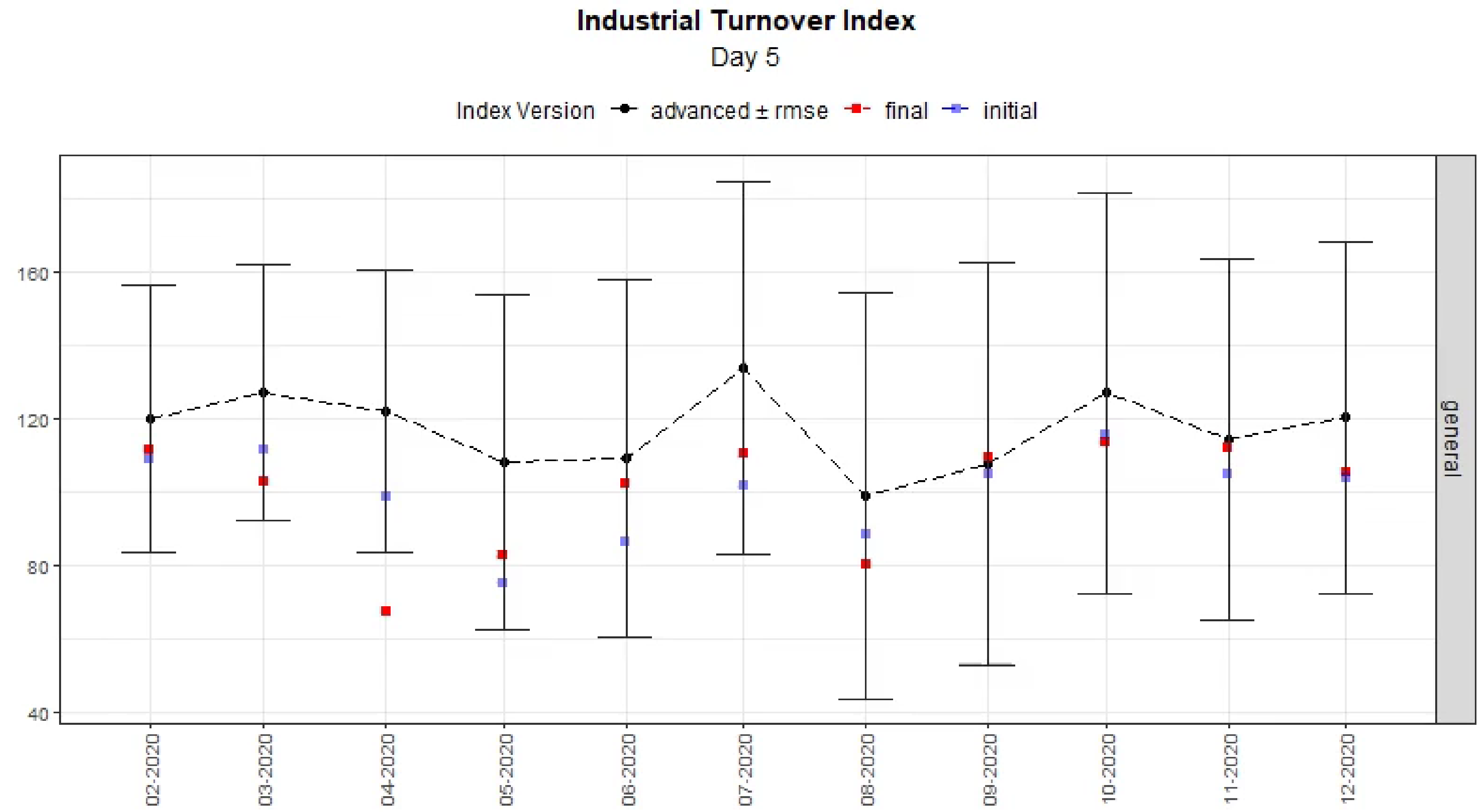
$t < t_{release}$
r_t: collected sample in t
$\hat{y}_{kt}$: estimation with Gradient boosting algorithm (light gbm).

- **Regressors** (287):

	ID	Cross	Long	Cross+Long	External
Hist. Series	✔	✔	✔	✘	✘
Running Month	✔	✔	✘	✔	✔

- **Process Pipeline: Modular design**

Case 3: Imputation - Nowcasting

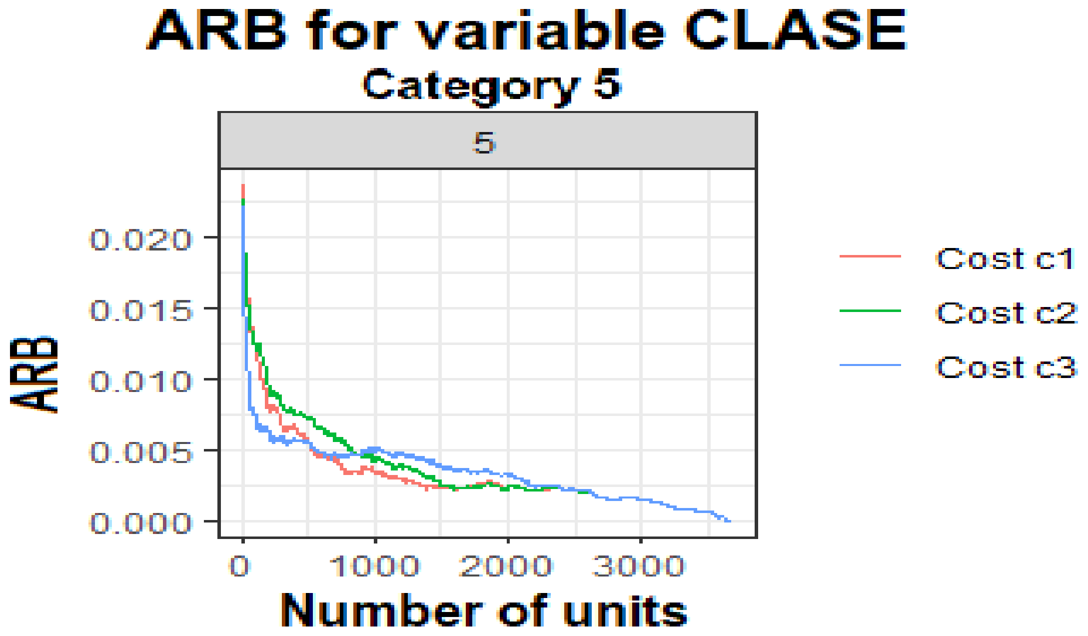


$t_{release}$= t+51

Case 4: Imbalaced data

- Three approaches:
  - Undersampling, oversampling, **cost-sensitive learning.**
- $s_k = \begin{cases} d_k \cdot \mathbb{P}(\epsilon_k = 1|X_k) \cdot c & if\ \mathbb{P}(\epsilon_k = 1|X_k) \leq \frac{c}{1+c}, \\ d_k \cdot \mathbb{P}(\epsilon_k = 0|X_k) & if\ \mathbb{P}(\epsilon_k = 1|X_k) > \frac{c}{1+c}. \end{cases}$
- European Health Interview Survey in Spain: occupation (CLASE).

		predicted	
		1	0
true	1	0	c
	0	1	0

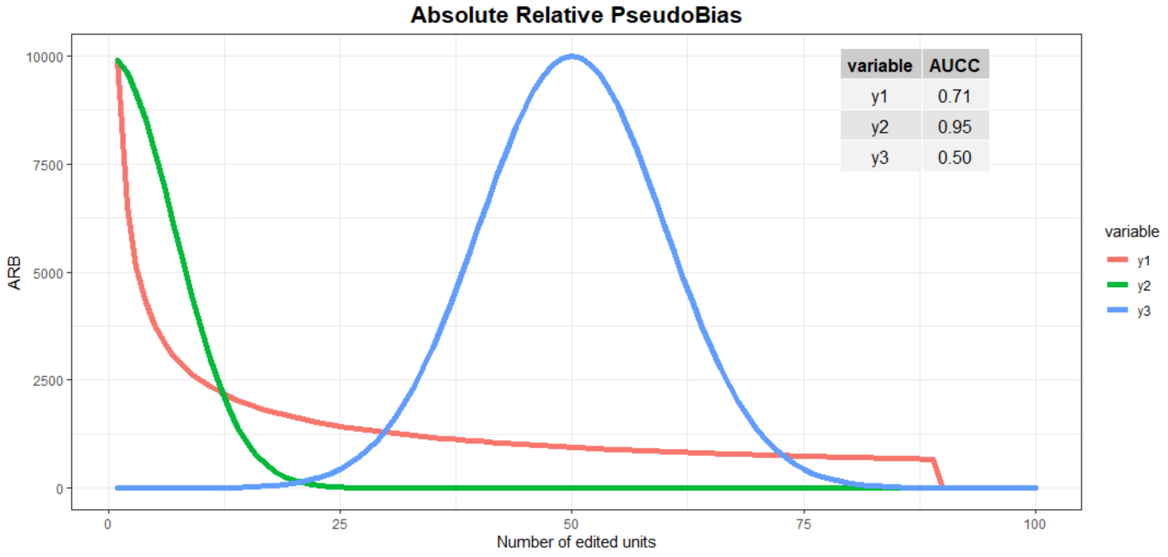


Case 5: Quality Measure

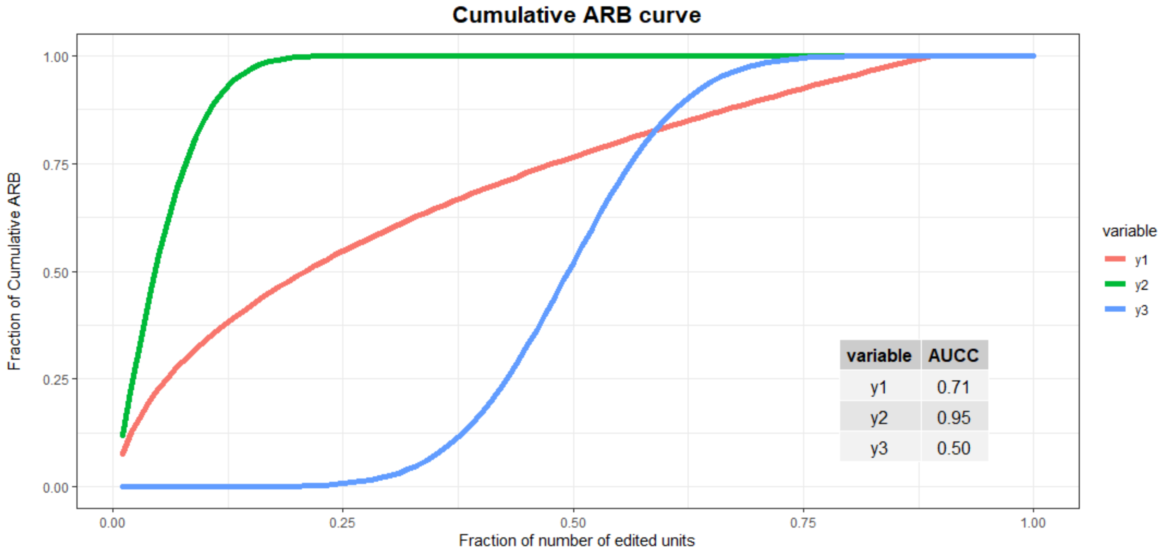
- Area under the ARB cumulative curve (AUCC) with coordinates (x = n_{ed}, y = y_k):

$$y_k = \frac{\sum_{i=0}^{k} ARB(n_{ed}=i)}{\sum_{i=0}^{n} ARB(n_{ed}=i)} \text{ with } k = 0, \dots, n.$$

- $0 \leq AUCC \leq 1$



variable	AUCC
y1	0.71
y2	0.95
y3	0.50



variable	AUCC
y1	0.71
y2	0.95
y3	0.50

Case 6: NLP of questionaire comments

- **Microdata** and **paradata** from data collection:
  - Remarks and comments (read during editing)
- Spanish Industrial Turnover Index Survey (monthly; 12000 units).
- NLP steps:
  - **Preprocess** (lowercase, remove stopwords, substitute into generic expressions…)
  - **Tokenize**
  - Apply **hash trick** to code all tokens
  - **Random forest** of classification. Target: $\epsilon_k \in \{0,1\}$ (revised/not revised)
- Results:

Token	AUC
1-grams	0,5923
2-grams	0,5732

More research is needed

Machine Learning in Official Statistics

Task	ML technique	GSBPM SubPhase
Record linkage	Clustering	2.4, 5.1
Coding	Classification	2.4, 4.3, 5.2
Outlier detection	Clustering	2.4, 4.3, 5.1, 6.2
Stratification	Classification	4.1, 4.3, 5.4, 5.6
Estimation	Regression/classification	4.3
Imputation	Regression/classification	5.4
Calibration	Regression/classification	5.6
SDC	Regression/classification	6.4
Error detection	Regression/classification	5.3
Imputation	Regression	5.4
Estimation w/ admin data	Regression/classification	5.1, 5.5, 5.7

Yung et al (2017) – Uses for Primary Data

Conclusions

- Machine learning algorithms are proving to be extremely useful to modernise and streamline many statistical production tasks even with traditional (survey) data.

- Detection of erroneous values of continuous, categorical, and semicontinuous variables.

- Improvement of accuracy, timeliness, and cost-efficiency in the editing phase.