Selective Editing of Categorical Variables with Random Forests: Results of the Implementation in the Statistical Production

Use of R in Official Statistics 2020

S. Barragán, M.R. González-García, **D. Salgado**, T. Vázquez

Statistics Spain (INE)

Statistics Austria, Vienna, 2-4 December, 2020

Overview

1. The production process
2. Methodology: selective editing of the occupation variable with random forests
3. Implementation
4. Results
5. Some conclusions

The production process (1/2)

Spanish branch of the European **Health Interview Survey**:

- We shall focus on **editing** the **occupation** variable.
- **CNO11** classification:
  - **Spanish** adaptation of **ISCO-08** classification.
  - **Social Class**: subject-matter aggregation of CNO11 codes.
- **Estimators** (target variable, domain U_d):

$$\widehat{Y}_d^{\text{target}} = \sum_{k \in s} \omega_{ks}(\mathbf{x}) \cdot \delta_{kU_d}^{\text{SC}} \cdot \delta_k^{\text{Target}}$$

- Sample size around 20000 statistical units with questionnaires for the **household** and a (randomly) selected **adult**.

The production process (2/2)

INSTITUTO NACIONAL DE ESTADISTICA

Spanish branch of the European **Health Interview Survey**:

- **Traditional** editing work:

  systematic interactive editing on each questionnaire **taking into account other nuclear sociodemographic variables** such as age, gender, professional situation, economical activity, income interval...

- Data collection into **weekly batches** loaded into the system for manual editing by **subject matter clerks** following **strict** guidelines.

- New production step based on random forests **not fully integrated**

- A **concurrent asynchronous** pipeline to **prioritize units** for the manual editing.

Methodology

$$s_k = \omega_k \cdot \left| y_k - \hat{y}_k \right|$$

$$\uparrow \frac{|y_k - \hat{y}_k|}{\nu_k} \rightarrow 0$$

$$s_k = \sqrt{\frac{2}{\pi}} \cdot \omega_k \cdot \frac{M\left(-\frac{1}{2}, \frac{1}{2}, -\left(\frac{\sigma_k^2}{\sigma_k^2 + \nu_k^2} \right)^2 \frac{(y_l - \hat{y}_k)^2}{2\nu_k^2} \right)}{1 + \frac{1-p}{p} \sqrt{\frac{\nu_k^2 + \sigma_k^2}{\nu_k^2}}}$$

$$s_k = \omega_k \cdot \mathbb{P}\left(\delta_{kU_d}^{\mathrm{sc}} \neq \delta_{kU_d}^{\mathrm{sc}(0)} \middle| \mathbf{Z}_k^{\mathrm{cross}} \right)$$

$y_k = \hat{y}_k + \nu_k$
$y_k = y_k^{(0)} + \sigma_k$

$y_k = \delta_{kU_d}^{\mathrm{sc}}, y_k^{(0)} = \delta_{kU_d}^{\mathrm{sc}(0)}$
$\mathbb{P}\left(\delta_{kU_d}^{\mathrm{sc}} \neq \delta_{kU_d}^{\mathrm{sc}(0)} \middle| \mathbf{Z}_k^{\mathrm{cross}} \right) = \sum_{m=1}^{M} w_m \cdot \mathbb{I}_{R_m}(\mathbf{x}_k; \mathbf{Z}_k^{\mathrm{cross}})$

$$s_k = \mathbb{E}_m \left[\omega_k \left| y_k - y_k^{(0)} \right| \middle| \mathbf{Z}_k^{\mathrm{cross}} \right]$$

Implementation (1/2)

- **Target** variable: **social class measurement error** $\epsilon_k^{sc} = I_k(sc_k^{raw} = sc_k^{(0)})$.
- **Predictors** (raw values):

  `Age`, `Sex`, `Stratum`, `Proxy`, `EconActivity` (NACE section + 1st, 2nd, 3rd code digit), `Occupation`, (ISCO Main Group + 1st, 2nd, 3rd code digit), `SocClass`, `NUTS2`, `EduDegree`, `ProfSituation`, `JobSituation`, `samplingWeight`

- ```
  model1 <- randomForest::tuneRF(
  train[, ..vars], train$target1, ntreeTry=1000, stepFactor=2,
  improve=0.05, trace=TRUE, plot=TRUE, doBest = TRUE)
  ```

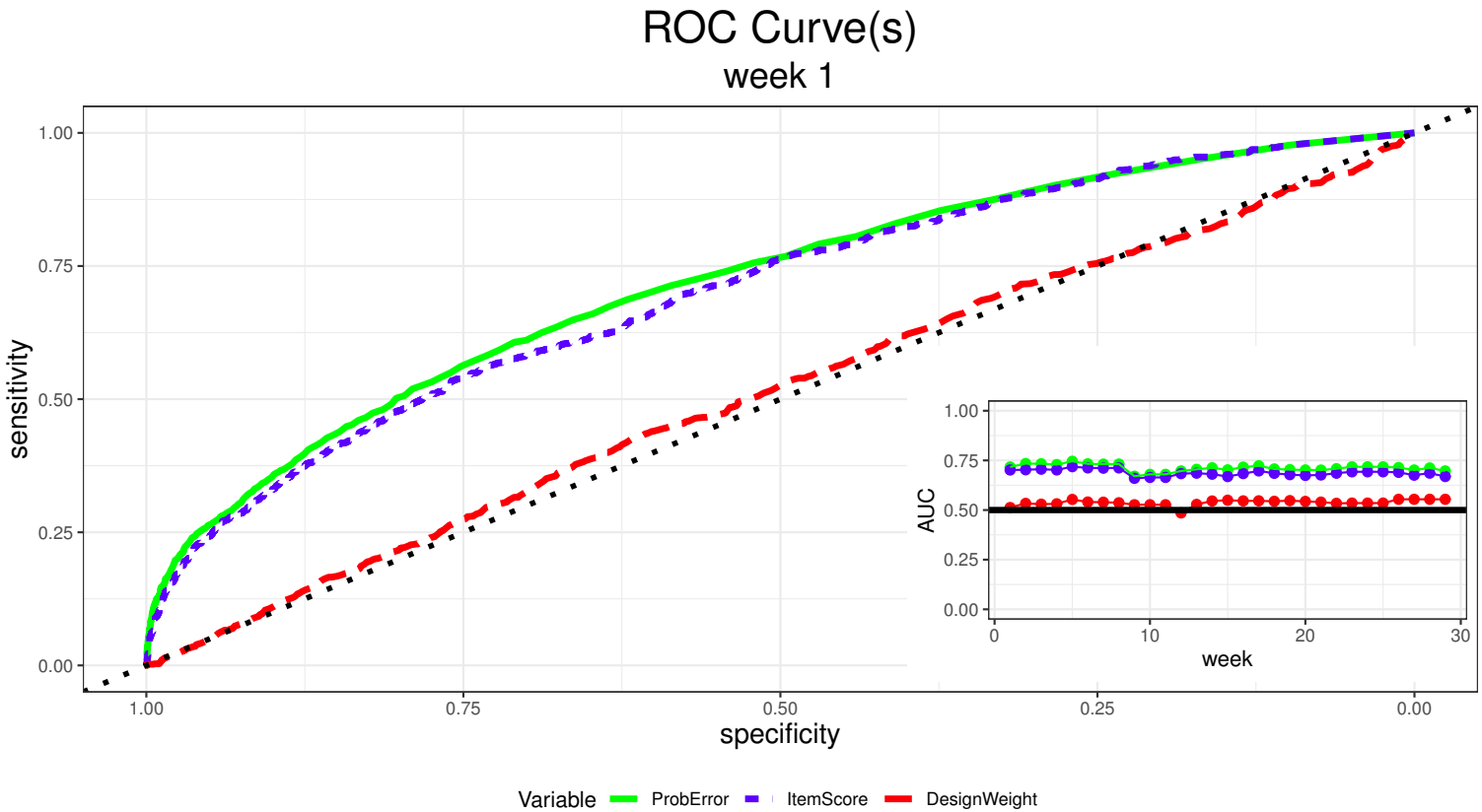
- For **each fieldwork batch**:
  - **Train** set: **historic** data sets + **on-course edited** data sets (raw + validated).
  - **Application** set: collection of **questionnaires not yet edited**.

- **First 2** (accumulated) batches (5200 units): **all** questionnaires edited **manually**
  ```

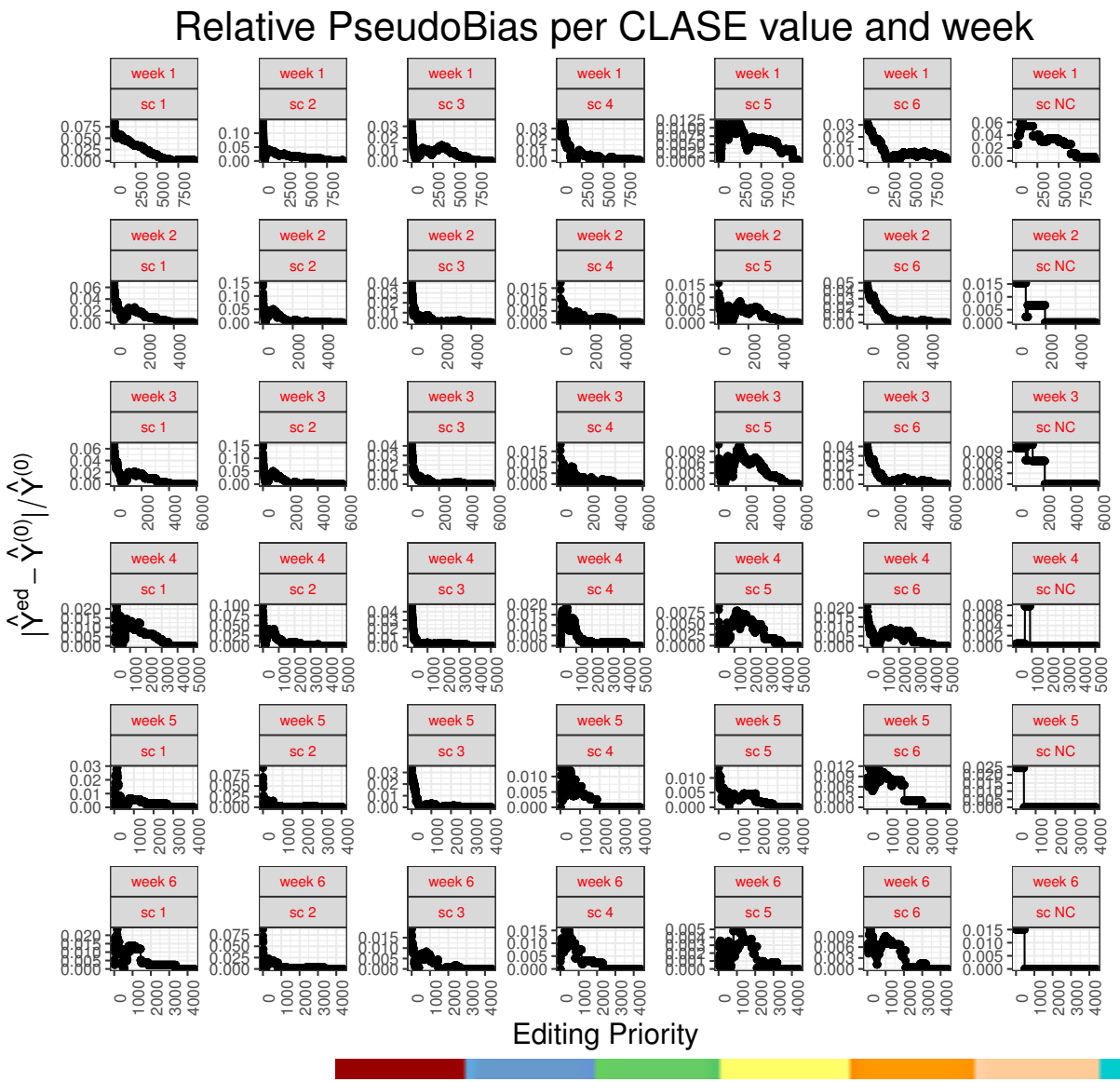
Implementation (2/2)

- Model **fitted** and non-edited units **(re)prioritised** every **week**
- **Collection** Batches $\neq$ **Editing** Batches
- **Traditional** editing computer system **complemented** with:
  - Editing **Priority**
  - **Revision Flag**
  - Measurement Error **Probability**
  - Item **Score**
- **First 2** (accumulated) batches (5200 units): **all** questionnaires edited **manually**
- **Intermediate** batches: First half model-selected + second half expert-selected
- **Final** batches: First half model-selected

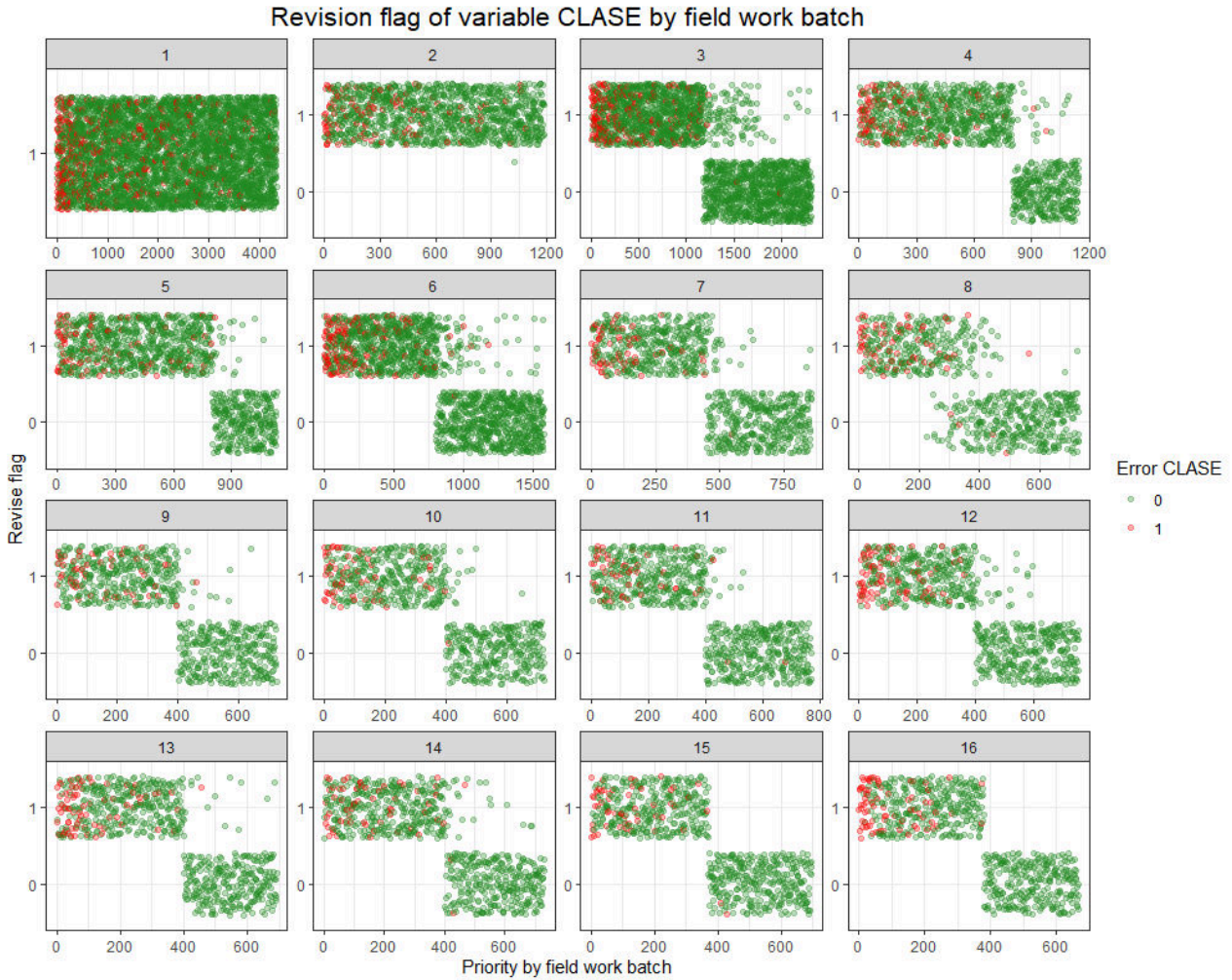
Results: rank efficiency?



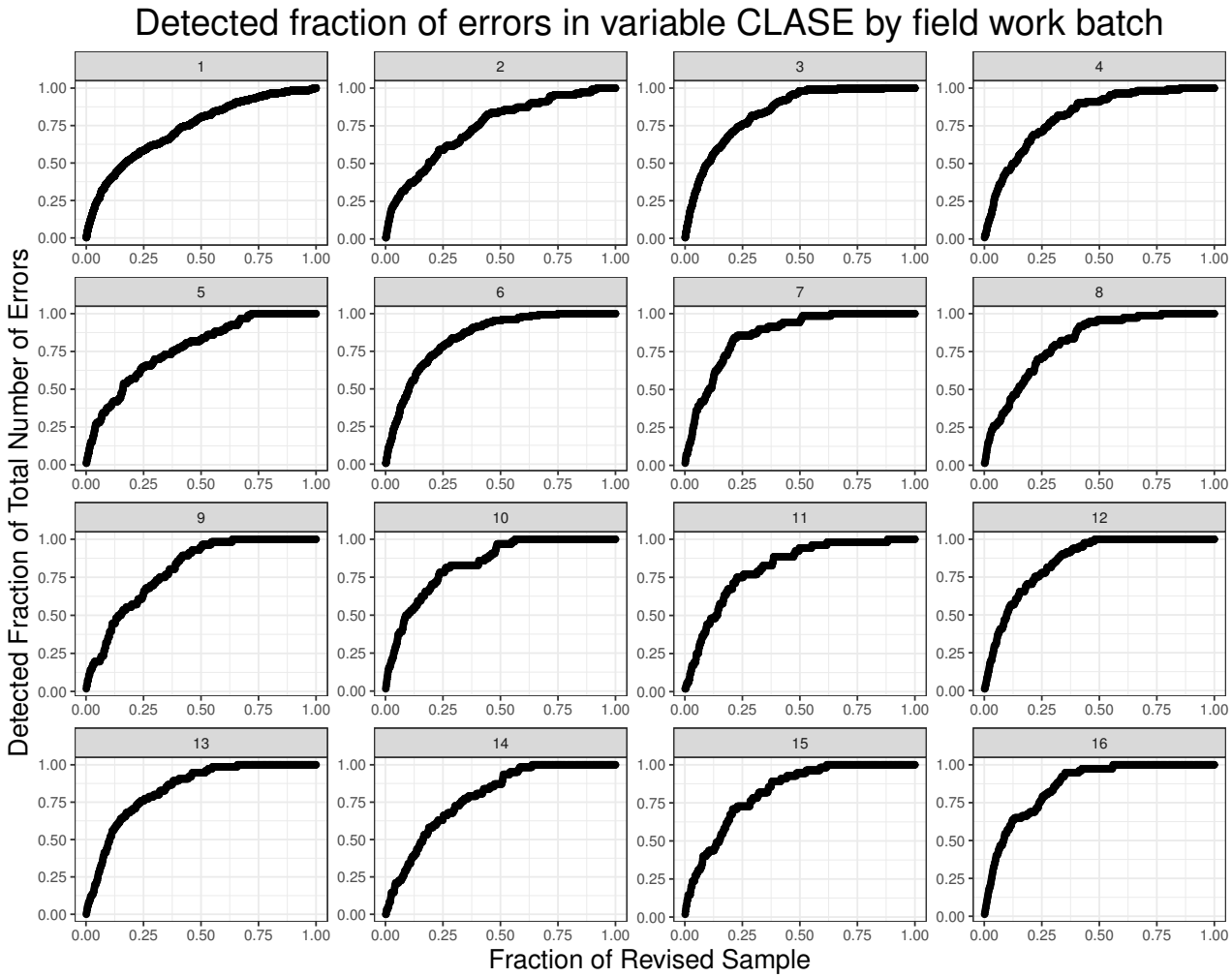
Results: pseudobias?



Results: errors ranked first?



Results: Errors vs. Revised Units



Results: Variable Importance

Main Conclusions

- **Effective prioritization** of units with influential errors.
- Collaboration of **subject matter** experts, **methodologists** and **IT** experts is fundamental.
- First: modular **deployment**; then: model **optimization**
  - **Regressors** $\rightsquigarrow$ collection and editing paradata.
  - **Hyperparameters** $\rightsquigarrow$ `mtry`, `stepFactor`, `improve`,...
  - **Unbalanced** learning $\rightsquigarrow$ sub/over-**sampling**, **cost-sensitive** learning, **ensemble** techniques...
  - **Multivariate** editing.
- **Software** development
  - Statistics Spain **internal** ecosystem with **key-value pair data representation**
    - `https://github.com/david-salgado/categObsPredModelParam`
    - Based on `randomForest` package
  - Future: mimic `validate` ecosystem