Advanced Industrial Turnover Index using Statistical Learning Algorithms

The Use of R in Official Statistics
uRos 2021

S. Barragán[1], L. Barreñada[3], J.F. Calatrava[1], J.C. Gálvez Sáenz de Cueto[1], J.M. Martín del Moral[2], E. Rosa-Perez[2,3], D. Salgado[1,3]

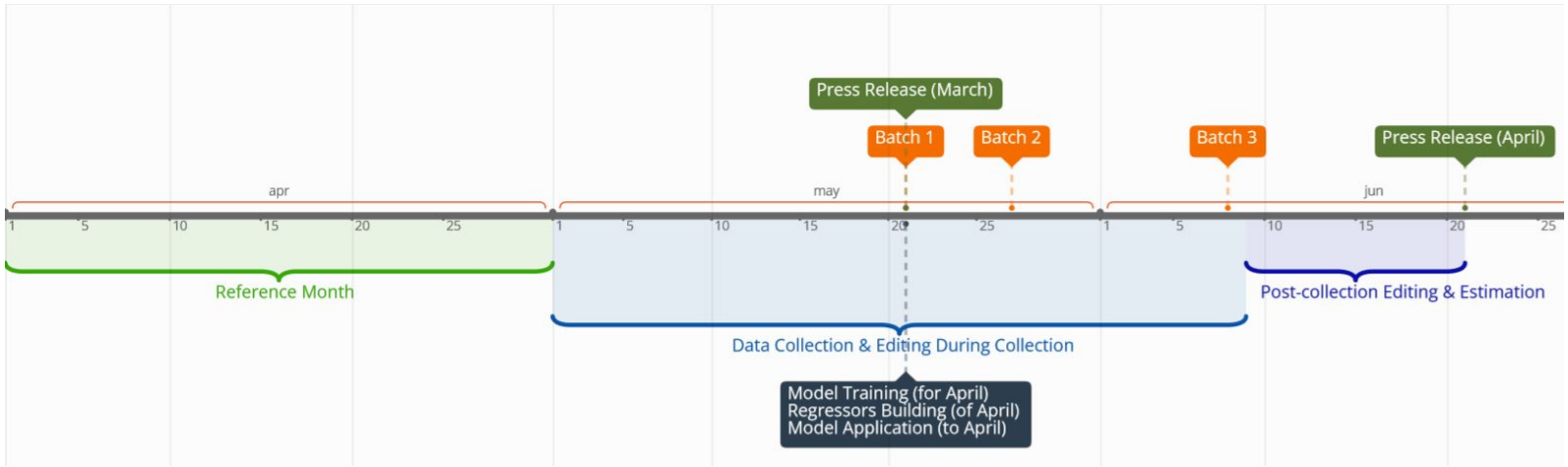
[1]Dept. Methodology and Development of Statistical Production, Statistics Spain (INE)
[2]S.G. for Industrial and Services Statistics, Statistics Spain (INE)
[3]Dept. Statistics and Operations Research, Complutense University of Madrid

Bucharest, 24-26 November 2021

Overview

(i) Short-Term Business Statistics: monthly release.

(ii) Accuracy vs. Timeliness



Methodology: Advanced Index Prediction

- Monthly fixed-base Laspeyres indices with cut-off sampling of establishments for total industrial turnover for each dissemination cell U_d

$$Y_{U_d}^{(m)} = \sum_{k \in U_d} y_k^{(m,\mathrm{val})} \rightsquigarrow I_d^{(m)}, \Delta_d^{(m,m-12)}, \Delta_d^{(m,m-1)}$$

- Tested advanced version:
  r_t subsample collected up to time $t < t_{\mathrm{release}}(m)$

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

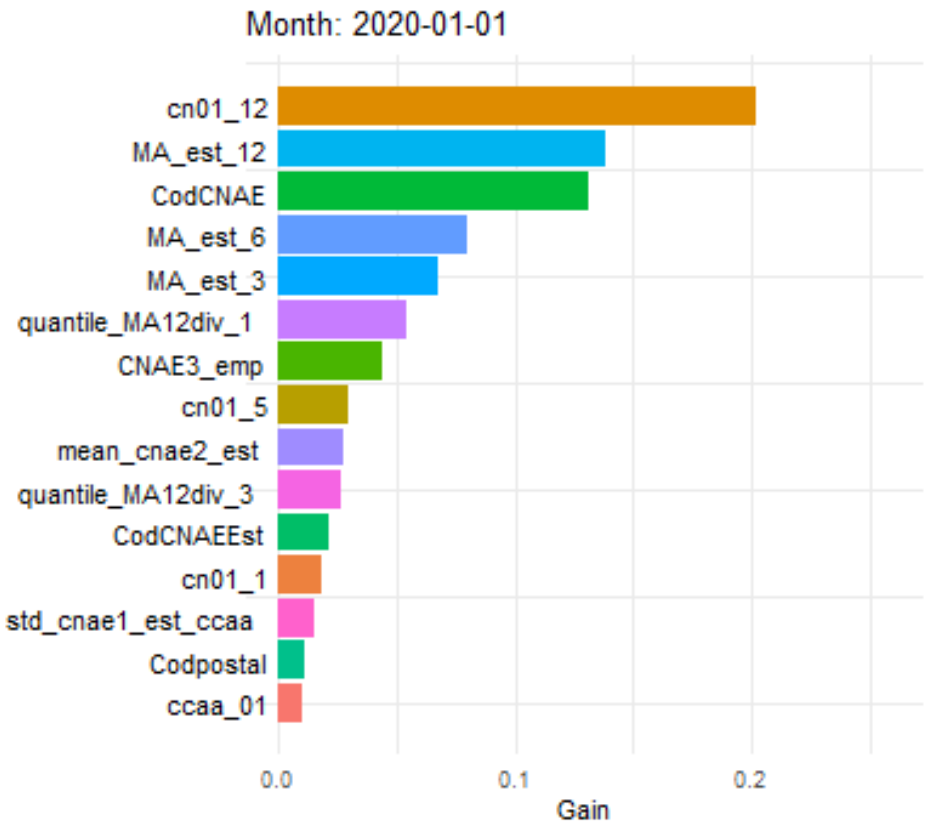
Prediction method: Statistical Learning

- Used data:
  - Survey data from ITI.
  - Survey data from IPI and IPRI.
- Target Variable: validated total turnover $y_k^{(\mathrm{m,val})}$ per unit and reference month.
- Regressors (287):

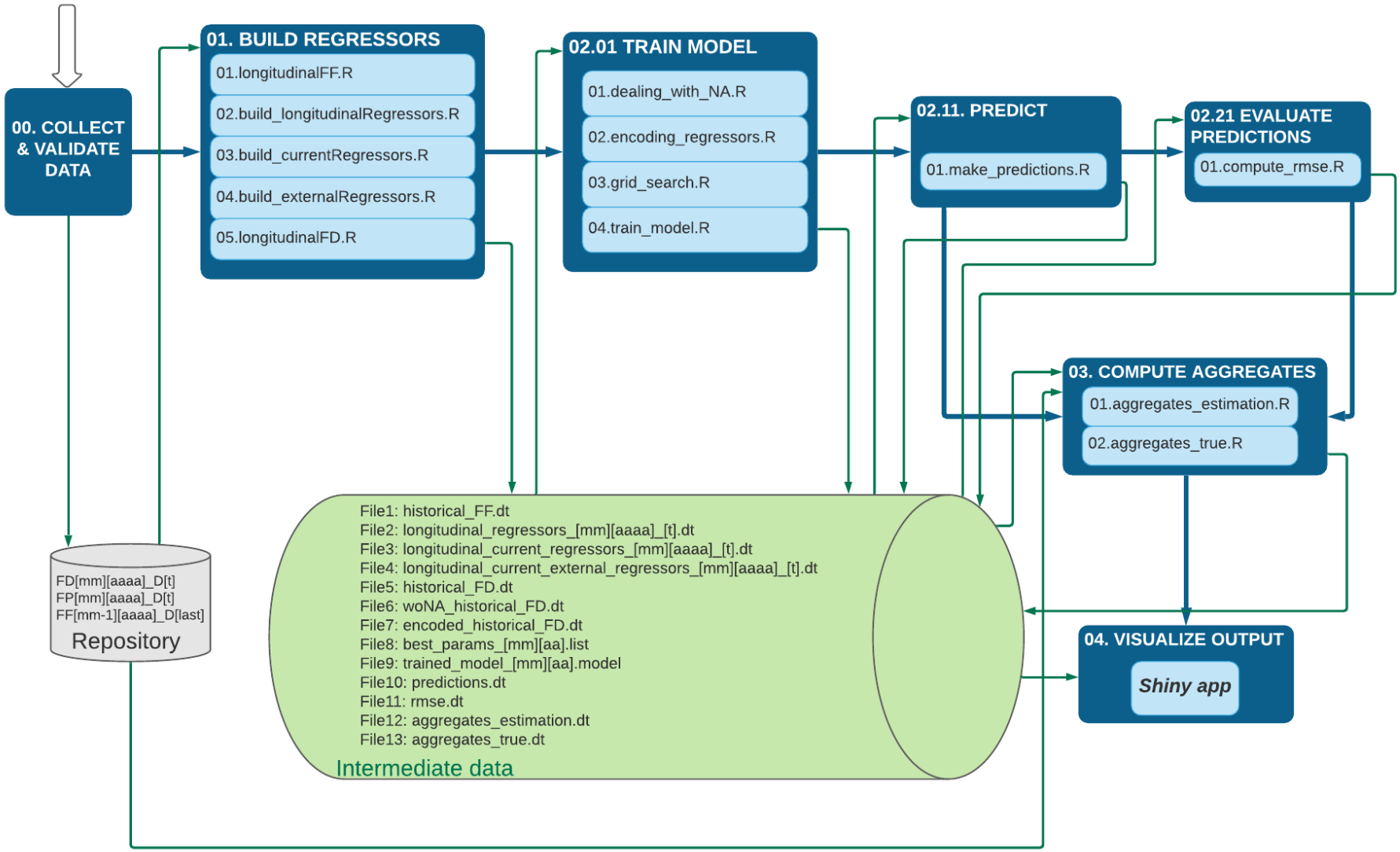
	ID	Cross	Long	Cross+ Long	External
Hist. Series	✔	✔	✔	✘	✘
Running Month	✔	✔	✘	✔	✔

- Light Gradient Boosting on Regression Trees.

Regressors Importance



Process Pipeline: Modular Design



Pilot: Software and execution

-  Packages Stack:
  - **Read and data management**

    `data.table`, `stringr`, `stringi`, `haven`, `lubridate`, `xml2`, `fastReadfwf`(1), `RepoTime`(1), `StQ`(1), `RepoReadWrite`(1), `RepoUtils`(1).
  - **Statistical Learning**

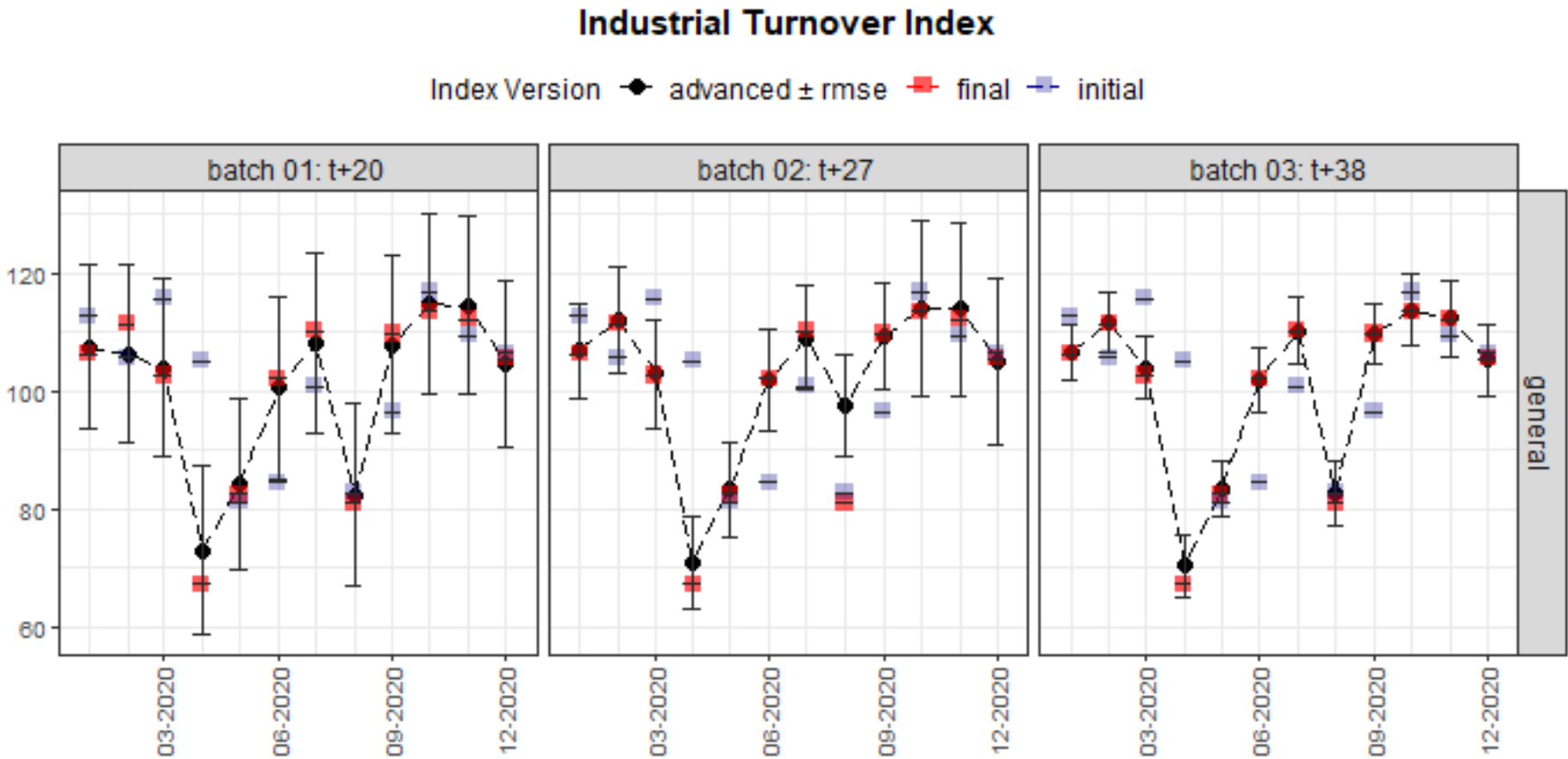
    `mltools`, `lightgbm`, `xgboost`, `psych`.
  - **Visualization**

    `shinydashboard`, `shiny`, `shinythemes`, `ggthemes`, `gridExtra`, `latex2exp`, `ggplot2`.

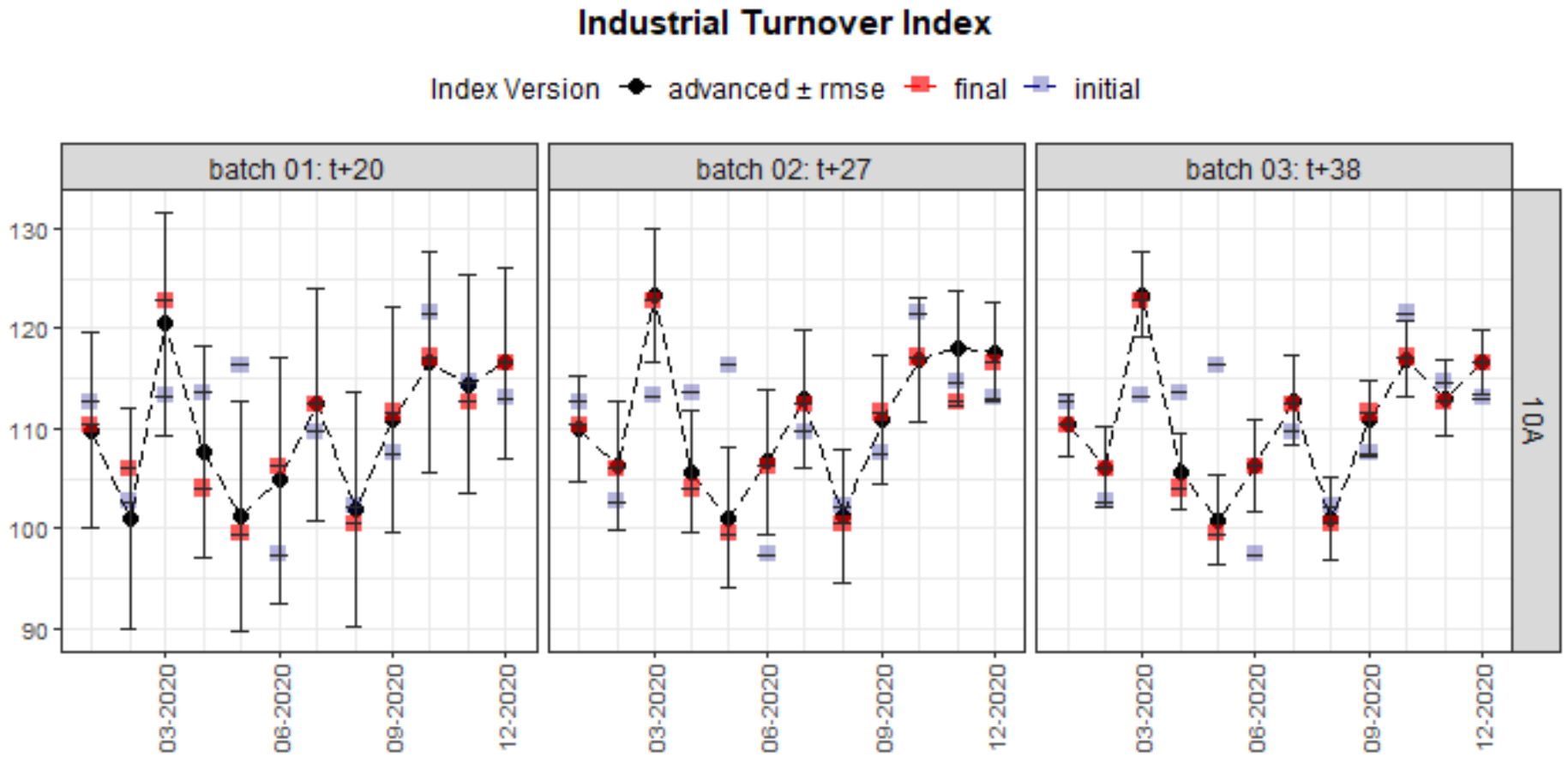
- Windows10, 32GB RAM, Intel(R) Core™ i7-4790 CPU 3.60GHz.
- Executed on more than 60 consecutive months (12000 establishments per month).

(1): In-house developed.

Results I



Results II



Ongoing and future work

- Predict semicontinuous variables $\rightsquigarrow$ seasonality
- Predict measurement errors: $y_{kt}^{(\mathrm{m,ed})} - y_{k}^{(\mathrm{m,val})}$
- Model residuals $\rightsquigarrow$ uncertainty intervals
- Implement in production $\rightsquigarrow$ interoperable software tools
- Generalize to other STS (probabilistic sampling).
- Daily Data Collection and daily turnover availability in the reference month.