# The European One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS): WP9 Use Case focused on imputation

## UNECE Expert Group on Statistical Data Editing

Sandra Barragán and *David Salgado*

STATISTIK AUSTRIA
Die Informationsmanager

UNECE

INE
Instituto Nacional de Estadística

# Outline

- General considerations
- Work Package Motivation and Orientation
- Work Package Description and Structure
- Conclusions

# General Considerations

- Multiple examples of usage of ML algorithms for imputation of missing values [UNECE, 2018, 2020, 2022, 2023]
- GSDEM [UNECE, 2019] as a framework
- Close coordination with error detection (WP8)
- Not only to improve accuracy in existing business functions . . .
- but also to impact on other quality dimensions (timeliness, relevance,. . . )

# WP Motivation and Orientation

- Level of **granularity** of business functions:
  - To produce a **predicted value** according to a statistical model
  - … in **categorical/semicontinuous/continuous** variables
  - To deal with **outliers**, **erroneous** and **missing** values
  - In **household/business** statistics
  - Impinging on **quality** dimensions (only accuracy?)
- To improve:
  - Accuracy ⇝ **post-collection** imputation
  - Timeliness ⇝ **early** imputation
  - Granularity ⇝ imputation **beyond the sample**

# Post-collection imputation: the setting

- To estimate a population total $Y_U = \sum_{k\in U} y_k$ ...
- ...by a (design-based) linear estimator $\widehat{Y}_U = \sum_{k\in s} \omega_k(s, \mathbf{x}) y_k$ ...
- ...under non-response: $\widehat{Y}_U^I = \sum_{k\in r} \omega_k(s, \mathbf{x}) y_k + \sum_{k\in s-r} \omega_k(s, \mathbf{x}) \hat{y}_k$ ...
- ...testing statistical learning models $m$ under different conditions

# Post-collection imputation: applications

- IT: In LFS surveys, identify and correct automatically the systematic error in economic activity.
- PL: Imputation for non-response of Statistics on accommodation establishments.
- SI: ML imputations for employment income data applied to non-response.
- ES: Imputation with ML for non-response in labour market statistics.
- PT: ML Treatment of the Annual Survey on Construction Enterprises Using Administrative Data.
- LU: ML for non-response: a) missing prices, b) household survey maybe in LFS.
- AT: International Trade in Goods Statistics: imputing weight or code.
- DK (O): ML for non-response in education statistics.
- CY (O): Imputation with ML for non-response of education level in the earning survey.

# Early imputation: the setting

- To estimate a population total $Y_U = \sum_{k \in U} y_k \ldots$
- ...by an advanced (design-based) linear estimator $\widehat{Y}_U(t) = \sum_{k \in r(t)} \omega_k(s, \mathbf{x}) y_k + \sum_{k \in s - r(t)} \omega_k(s, \mathbf{x}) \hat{y}_k \ldots$
- ...at early times $t < t_{release} \ldots$
- ...testing statistical learning models $m$ under different conditions to predict **microdata values** $\hat{y}_k$ exploiting patterns in past and current microdata in the same statistics.

## Early imputation: applications

- IT: Attained Level of Education (ALE) for sample with longitudinal administrative data.
- PL: Imputation for flash estimates of Statistics on accommodation establishments.
- DE: Early imputation in short term business statistics (estimate totals based on early observations).
- ES: Early estimates of Industrial Turnover Index.

# Imputation beyond the sample: the setting

- To estimate a population total $Y_U = \sum_{k \in U} y_k$ . . .
- . . . by an augmented (model-based) linear estimator $\widehat{Y}_U = \sum_{k \in s} y_k + \sum_{k \in U-s} \hat{y}_k$ . . .
- . . . testing statistical learning models $m$ under different conditions to predict **microdata values** $\hat{y}_k$ exploiting patterns in past and current microdata in the same statistics. . .
- for all units $k \in U$ in the population

# Imputation beyond the sample: applications

- NL: General methods of imputation.
- IT: Estimating categorical variables by using ensemble approach (comparing with traditional methods).
- ES: Imputation with ML of SBS variables in all population units with administrative data.
- AT: Statistics on Tourism acceptance; estimating household income (EU-SILC definition); estimating poverty-rates/income distribution for children attending school.
- DK (O): Household survey.

## Cross-sectional aspects

- In-house **confidential microdata** sets
- About the model: feature engineering, algorithm and hyperparameters, model evaluation, statistical product/process evaluation
- **Computational** requirements: close to **production**
- **Quality** assessment: statistical product, production process
- ESS **guidelines from use cases**: from concrete national needs to **international guidelines**

# WP Description and Structure

- **Methodological** developments
- Development of **PoC/MVP/prototypes** and preparation for **deployment in production**
- **Quality** aspects
- **Deliverables**:

**D9.1.-** Methodological aspects from use cases in Machine Learning techniques for early imputation in the production of official statistics.

**D9.2.-** Methodological aspects from use cases in Machine Learning techniques for post-collection imputation in the production of official statistics.

**D9.3.-** Methodological aspects from use cases in Machine Learning techniques for imputation beyond the sample in the production of official statistics.

**D9.4.-** Development of prototypes and preparation for deployment of imputation use cases with Machine Learning techniques in the production of official statistics.

**D9.5.-** Quality aspects of use cases in Machine Learning techniques for imputation in the production of official statistics.

# Conclusions

- We consider both **traditional business functions** (dealing with detected errors, missing values, and outliers), and novel proposals to produce **early estimates** and more **granular statistics**.
- Goals from the identification and conformation of **generic methodological guidelines** to the development of **proofs of concepts** and **minimal viable products** as close as possible to real **production conditions**.
- Both methodological and technological findings will be duly complemented with **statistical quality assessment** considerations.

# References

UNECE. Workshop on statistical data editing, 2018. URL https://unece.org/info/events/event/18867. Neuchatel, 18-20 September.

UNECE. Generic statistical data editing model v2.0, 2019. https://statswiki.unece.org/display/sde/GSDEM.

UNECE. Workshop on statistical data editing, 2020. URL https://unece.org/info/events/event/18365. Geneva, 31 August-04 September.

UNECE. Expert meeting on statistical data editing, 2022. URL https://unece.org/statistics/events/SDE2022.

UNECE. Machine learning for official statistics workshop, 2023. URL https://unece.org/info/events/event/373380. Geneva, 05-07 June.