UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

7-9 October 2024, Vienna

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

Statistics Spain (INE)

sandra.barragan.andres@ine.es; david.salgado.fernandez@ine.es

I. INTRODUCTION

1. Artificial Intelligence [Russell and Norvig, 2010] is considered a disruptive technology in the currently ever-increasing data ecosystem in many industries [Păvăloaia and Necula, 2023]. The production of official statistics is not an exception [see e.g. UNECE, 2021, Fraisl, 2024].

2. Statistical data editing [de Waal et al., 2011, UNECE, 2019], as an essential statistical process phase for quality assurance, is already witnessing an intense activity for the incorporation of these novel methods and technologies [see UNECE, 2018, 2020, 2022, 2023, and multiple references therein].

3. All these initiatives require an important innovation effort, including international collaboration among multiple statistical offices. In this line, the European Statistical System (ESS) has recently catalyzed these innovation efforts through the so-called ESS Innovation Agenda [ESS, 2024], endorsed by the ESS Committee and articulated through the ESS Innovation Network, cross-disciplinarily involving several ESS Directors' Groups for its implementation and coordination.

4. One of these outstanding innovation projects is the "One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS)", developing innovative solutions with respect both to statistical products and processes, allowing for more timely production of official statistics and the delivery of better responses to user needs. The project, which has recently started, covers both cross-cutting aspects and specific AI-ML use cases for the production of official statistics in multiple work packages. In this contribution, we shall focus on a description of the work package devoted to imputation.

5. This document is structured as follows. In section II we shall provide a brief description of the structure of whole project. In section III we provide a detailed description of the structure of the project work package 9 on imputation. In section IV we close with some preliminary conclusions.

II. THE EUROPEAN PROJECT "ONE-STOP-SHOP FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR OFFICIAL STATISTICS (AIML4OS)"

1. The project, which comprises 14 countries, aims to develop innovative solutions with respect to statistical products and processes using different AI-ML techniques. The project comprises 6 cross-cutting and 7 use-case-oriented work packages, among which editing (WP8) and imputation (WP9) constitute two specific focuses of development (see figure 1).

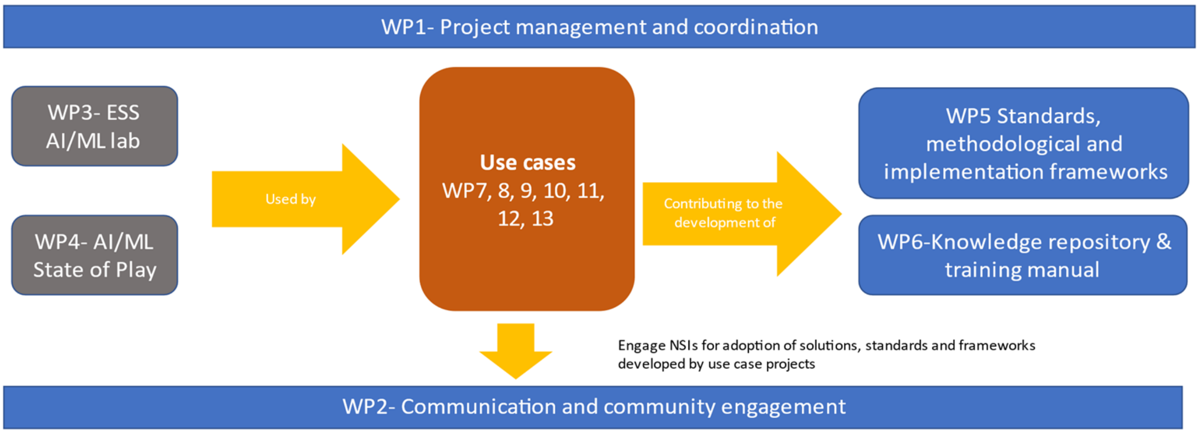
WP1- Project management and coordination
WP3- ESS AI/ML lab
WP4- AI/ML State of Play
Used by
Use cases
WP7, 8, 9, 10, 11, 12, 13
Contributing to the development of
WP5 Standards, methodological and implementation frameworks
WP6-Knowledge repository & training manual
Engage NSIs for adoption of solutions, standards and frameworks developed by use case projects
WP2- Communication and community engagement

Figure 1. Diagram of the Project Work Packages.

2. The cross-cutting work packages deal with structural aspects in the incorporation of the AI-ML technologies and underlying statistical methods in the production of official statistics. Apart from the necessary project management and coordination (WP1) and multiple communication and community engagement activities (WP2), the four main cross-cutting activities are:

- WP3: to develop a technological testing environment to be used by the different use-case-oriented work packages for experimentation, providing documentation and technical support both for their usage and for the creation of a similar infrastructure in every statistical office.
- WP4: to obtain evidence of the state-of-play in the use of AI-ML both within the ESS and beyond.
- WP5: to develop methodological and implmentation guidelines for the application of AI-ML in the production of official statistics, generalising knowledge, norms and best practices paving the way for the standardisation and industrialisation of the statistical production process.
- WP6: to share knowledge providing guidance for the integration and maintainance of AI-ML solutions in statistical offices through training and support.

3. The use-case-oriented work packages touch different aspects in the production of official statistics arising in multiple statistical domains and in different phases of the production process:

- WP7: to develop methodological and implementation guidelines to identify and apply AI-ML techniques on earth observation data and satellite imagery.
- WP8: to develop methdological and implementation guidelines for the automation, efficiency-gain, and quality improvement of statistical editing with a focus on error and outlier detection, exploring with practical examples. See [REF WP8 in this meeting].
- WP9: to develop methodological and implementation guidelines for imputation beyond the classical setting of error treatment, also exploring practical examples. This work package provides the content for this contribution.
- WP10: to develop methdological and implementation guidelines for providing automatic coding solutions for multiple international statistical classifications (NACE, COICOP, ISCO,. . .), recommending production pipelines and approaches for the multi-language challenge.
- WP11: to develop methdological and implementation guidelines for ML models estimating population-scale firm-level supply chain network datasets.
- WP12: to explore the use of generative large language models in Statistics, analysing their integration, support, fine-tuning, and identifying their enablers and constraints for the production of official statistics.
- WP13: to explore and investigate the different AI-ML algorithms for the generation of synthetic data in official statistics, delivering a proof of concept.

III. WP9: USE CASE FOCUSED ON IMPUTATION

In this contribution we summarise the structure of Work Package 9 concentrating on imputation, with a close coordination with Work Package 8, focused on editing (understood as error and outlier detection). We share the structure of the work package seeking for business functions beyond error treatment and a clear orientation towards more timely and granular production of official statistics and the delivery of better responses to user needs. The main challenge is to scale up knowledge and experience from concrete national needs to general solutions to accrete European guidelines. This project shares the vocation to establish strong international collaborations and to push forward the community-building effort in Official Statistics, now especially needed in the new data and AI-ML ecosystem.

A. Work Package motivation and orientation

1. The main objectives in this work package are to develop, test, and, if shown successful, implement AI/ML-based solutions for imputation processes. We propose three basic stages. Firstly, the goal is to identify and develop methodological proposals aligned and integrated in the statistical production process, which will be tested with specific datasets, populations, and variables. Secondly, if proved successful, these proposals will be promoted to proofs of concept (PoC), minimum viable products (MVP), or prototypes so that feedback can be easily distilled for WP5 on standards, methodological and implementation frameworks. Finally, the last goal is to design and implement a CD/CI modular process as close as possible to production conditions to identify the needs for infrastructure, thus serving as input for the ESS AI/ML Lab (WP3). The proposals are intended to two-foldly benefit from and feed both WP4 and WP6 attempting to grow the body of knowledge of AI-ML solutions for the production of official statistics.

2. In terms of the Generic Statistical Data Editing Model [UNECE, 2019], imputation can be seen as a family of business editing functions to treat detected data errors, either measurement errors, missing values, or outlier treatment. In this line of thought an obvious use case for AI-ML techniques is the improvement of these already identified business functions in the statistical production process.

We can focus on precision improvement (e.g. with more accurate models), on efficiency cost gains (e.g. with automatic pipelines), or on a combination and trade-off between both. Any intelligent agent system can be considered in this fashion to improve existing business functions in a production process.

3. Nonetheless, a complementary view is also desired where new business functions could be as well identified to improve official statistics to cover wider and novel statistics needs. In this sense two ever-pressing demands on statistical offices stand in the form of more timely and more granular information. We propose, as discussed below, to make use of the predictive power of AI-ML underlying statistical models and the existing patterns in survey and administrative microdata to explore new scenarios for imputing and providing official statistics in a more timely and granular way.

4. We shall distinguish three main research tracks pursuing accuracy, timeliness, and granularity. Notice that, although all these quality dimensions (and more) come all together when assessing a statistical product, we will make this distinction as a matter of organizing the tasks. Under these considerations, firstly we shall name **post-collection** imputation to the investigation of methodological proposals in a classic finite-population estimation scenario where some units in the sample $k \in s$ show missing or invalid target values y_k, which thus need to be imputed $y_k = \hat{y}_k$ under some prediction model. The role of AI-ML models here is to seek for a higher accuracy in the imputed values.

5. Secondly, we shall name **early** imputation to the investigation of methodological proposals to reconstruct the target values y_k for all units in the sample $k \in s$ even much before the data collection phase has finished. It can be considered a nowcasting flavour, but here we concentrate on using both survey and administrative microdata of the same statistics from its (possibly immediately) past history and ongoing data collection operations. Some very preliminary results are promising in this line [Barragán et al., 2022]. When applied to periodical short-term business statistics, notice how this proposal makes systematic use of microdata patterns increasingly accumulating with time. Furthermore, microdata are still revised and validated by subject matter experts providings thus firm ground for their use in prediction models.

6. Thirdly, under the name imputation **beyond the sample** we devise the construction of statistical models to predict target values y_k for units in the frame population but not in the sample $k \in U - s$. Indeed, a sample selection with this goal and accounting for the available auxiliary information in the population frame could in principle be thought to synthetise target values in a wide fraction of the population of analysis, increasing thus granularity. Notice, however, that accuracy is an important challenge in this proposal.

7. These research tracks will be undertaken with concrete practical examples from actual official statistics. Notice that many details will need to be considered when applied to concrete datasets (auxiliary information availability, categorical/continuous variables, model selection criteria, hyperparameter optimization strategies, ...). The three research tracks will be taken as general motivation guidelines to identify, construct, apply, and implement, if possible, different imputation models in these scenarios. In practice, a combination of factors differently impinging on accuracy, timeliness, and granularity may arise, none of them being discarded. High-quality standards will be the ultimate goal.

8. The impact of this use case on new statistical products is clear in the form of possible new experimental statistics but also in methodological terms in order to improve the quality of the process in the current statistical products in relation to timeliness, granularity and accuracy.

B. Work Package description and structure

1. We plan the execution of our research tasks in three big steps:

(1) **Methodological developments.** The first task focuses on the methodological aspects regarding the use of ML techniques for each project. In this task we include all the preliminary work such as literature review, assessment of methods and proposals to approach each problem, etc. Next, this task also entails the execution and analysis of the identified solutions for each problem. Finally, if these solutions are considered feasible and acceptable, we shall identify, distil and assess the best option for the PoC/MVP development. Otherwise, limitations for use in production will be clearly identified. The output of this task should contain our first input to WP5 on standards, methodological and implementation frameworks.

(2) **Development of PoC/MVP/prototypes and preparation for deployment in production.** This second task embraces the development of PoC/MVP/prototypes for those methodological solutions positively assessed for their use in production and the preparation for deployment in production as part of an MLOps platform or similar implementing a CD/CI modular process. This must show the empirical validity of the approach and some first results. For those methodological solutions negatively assessed for their use in production, limitations and obstacles will be conveniently explained and documented with the goal of providing orientation and recommendation to statistical producers in the use of AI/ML techniques for imputation. The impact of the imputation at microdata level with these techniques in the end-to-end process will be considered. The output of this task should contain our second input to WP5 on standards, methodological and implementation frameworks as well as our main input for the ESS AI/ML data lab (WP3) and technological needs for building a production platform implementing these methodological solutions.

(3) **Quality aspects.** This task embraces cross-cutting aspects regarding quality along different stages of the imputation process and in the multiple situations considered in this work package.

2. Accordingly, the output will be structured in the following 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.

3. The work package is integrated by experts from 11 European statistical offices, namely Spain (coord), Germany, Austria, Netherlands, Poland, Portugal, Slovenia, Italy, Luxembourg, Denmark, and Cyprus.

4. Each participant will propose a specific sub-project as part of one or more of the generic research lines to contribute in the methodology development. The datasets to be used by each participant will be from their own organization and will not be shared with the rest of participants, since they are sensitive microdata subjected to confidentiality, legal restrictions, and disclosure control. Since the data is in-house from each participant, the availability is guaranteed. Some examples of datasets that will be used are: a) post-collection imputation: accommodation establishments, employment income

data, prices of electronic products; b) early imputation: Attained Level of Education (ALE), statistics on accommodation establishments, short-term business statistics such as Industrial Turnover Index; c) imputation beyond the sample: Structural Business Statistics, Attained Level of Education.

5. No restriction is considered for the nature of statistical units (business units, establishments, households, persons, etc.) or variables (categorical, semi-continuous, continuous, etc.) and their interrelation.

6. From each sub-project multiple aspects will be evaluated and distilled to conform generalizable methodological guidelines. Some of these aspects are (i) the preprocessing work with the in-house datasets, (ii) guidelines for the exploratory data analysis, (iii) generic considerations about the ML model to be explored and used, (iv) feature engineering, (v) algorithms and their hyperparameters optimization, (vi) model evaluation (metrics, etc.), (vii) computational requirements and software, (viii) quality indicators, quality assessment and connection with the European Statistics Code of Practice and ESS Quality Assurance Framework.

7. The duration of the whole work package runs parallel to the whole project, namely 48 months from April 2024. However, the intermediate outputs will be constructed incrementally in shorter work cycles. In every work cycle different aspects of the final outputs (feature engineering, hyper-parameter optimization, model selection, model evaluation, etc.) will be approached to get closer to the final MVP/PoC oriented towards deployment in production. There are different sub-processes that can be built up incrementally over the lifetime of the project as well as the possibility to consider different data sets to replicate the methodology.

8. We view at least two kinds of stakeholders. On the one hand, the methodological and technological approach can be reused and adapted for many other statistical programmes and surveys. On the other hand, the improved/novel statistical outputs may be of interest to specialized users such as policy-makers and decision-taking institutions as well as citizens in general. In relation with the first type, we will receive their feedback through internal consultations and during conferences where there may be also external stakeholders as participants. In relation with the second type, depending on the degree of development for dissemination of the final outputs, we shall monitor the usual users feedback channels.

IV. CONCLUSIONS

1. To incorporate the disruptive innovative advancements derived from AI-ML techniques into the production of official statistics the ESS has recently launched an ambitious European project covering both cross-cutting aspects and multiple use cases.

2. Among these use cases, imputation is structured in a work package devoted both to traditional business functions dealing with detected errors, missing values, and outliers, and to novel proposals to produce early estimates and more granular statistics, thus impinging on timeliness and granularity as two highly demanded quality improvements in Official Statistics.

3. The goals have been set to go 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.

4. Both methodological and technological findins will be duly complmemented with statistical quality assessment considerations in relation with the high-quality standards in the production of official statistics.

References

S. Barragán, L. Barre nada, J.F. Calatrava, J.C. Gálvez Sáenz de Cueto, J.M. Martín del Moral, E. Rosa-Pérez, and D. Salgado. Early estimates of the industrial turnover index using statistical learning algorithms, 2022. URL https://www.ine.es/GS_FILES/DocTrabajo/art_doctr032022.pdf. Statistics Spain Working Paper 03/22.

T. de Waal, J. Pannekoek, and S. Scholtus. *Handbook of statistical data editing and imputation*. Wiley, 2011.

ESS. ESS Innovation Agenda, 2024. URL https://cros.ec.europa.eu/ess-innovation-agenda.

D. Fraisl. The potential of artificial intelligence for the SDGs and official statistics. *PARIS21 Working Paper*, April:1–21, 2024.

V.-D. Păvăloaia and S.-C. Necula. Artificial intelligence as a disruptive technology: A systematic literature review. *Electronics*, 12:1102, 2023. URL https://doi.org/10.3390/electronics12051102.

S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

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. *Machine Learning for Official Statistics*, 2021. URL https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf.

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.