

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics

by Francesca Kay

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics

Francesca Kay ¹

Abstract

Several challenges encountered when constructing U.S. administrative record-based (AR-based) population estimates for 2020 are identified. They include locational accuracy, person coverage and its consistency over time, filtering out non-residents and people not alive on the reference date, uncovering missing links across person and address records, and predicting demographic characteristics. Several ways to address these issues are discussed. Regression results illustrate how the challenges and solutions affect the AR-based county population estimates.

Keywords: Official statistics; Quality; Artificial intelligence; Machine learning.

1. Introduction

The use of Artificial Intelligence and Machine Learning (AI/ML) for the production of official statistics is one of the strategic domains that need to be developed further and where coordinated, systemic action is beneficial. The One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS) will play an important role in developing innovative solutions with respect to statistical products and processes, allowing for more timely production of official statistics and the delivery of better responses to user needs. The AIML4OS is bringing together a consortium of 14 countries to develop knowledge and use cases supporting the use of AI/ML-based solutions for the production of official statistics.

Objectives of AIML4OS Project

The overarching objective of the project is dual:

- To realise added value by aiming at providing tailored guidance and assistance to the European Statistical System (ESS) and beyond, and
- To continue developing knowledge and use cases supporting the use of AI/ML-based solutions for the production of official statistics.

1.1 Specific objectives

The specific objectives of the AIML4OS are to provide a single entry point for ESS staff involved in innovation to identify concrete opportunities. The project will provide guidance and support in deploying AI/ML solutions within quality, methodological and implementation frameworks to facilitate new statistical products and processes taking advantage of AI/ML state of the art research and developments.

The key underlying activities for the AIML4OS are:

- Develop, maintain and evolve a coherent set of relevant capabilities (including methodologies, guidelines, sandboxes, labelled data, processes, methodological, implementation and quality frameworks) for implementing AI/ML-based solutions in official statistics across the ESS.
- Set up a platform/hub providing a single entry point for ESS staff to access relevant capabilities.
- Provide support and guidance for the integration and maintenance of relevant AI/ML-based solutions in ESS organisations through training and active and efficient support.
- Build communities around open source solutions developed and maintained by ESS members.

¹ Francesca Kay, Central Statistical Office, Ireland, Francesca.kay@cso.ie. The author acknowledges Eurostat staff devoted to the project – Albrecht Wirthmann, Konstantinos Giannakouris, Jean-Marc Museux

- Share ideas, experiences, success stories and lessons learned to stimulate innovation based on the use of AI/ML.
- Enable and facilitate the transition from development and experimentation of AI/ML-based solutions to industrialisation and production.

1.2 Outcomes of AIML4OS Project

The overarching expected outcomes are to:

- Build a framework for developing AI/ML solutions to be used in the context of official and European statistics.
- ESS staff and partners have access to established and proven AI/ML solutions/resources to be leveraged in the context of official statistics production.
- ESS organisations are encouraged to innovate and eventually realise benefits from the use of AI/ML approaches for statistical products.
- Economies of scales and resources are achieved through cooperation inside and outside the ESS and transition from ideas to production is accelerated.

2. The one-stop-shop on AI/ML for official statistics

2.1 Introduction of AIML4OS Project

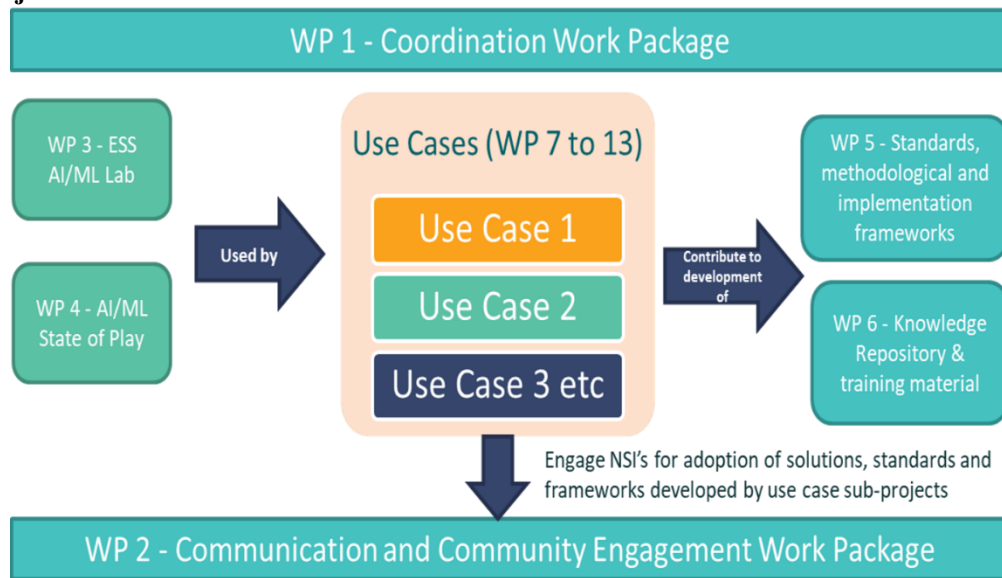
The innovation agenda endorsed by European Statistical System Committee (ESSC) in February 2023 recognised AI/ML as a key enabler for cross-cutting developments, for processing new data sources and supporting complex processes. The use of AI/ML will, for example, allow for the use of data from sensors, satellite images and web scraping, leading to new, more frequent, more timely and more granular statistical outputs and improved statistical processes.

The potential for AI/ML is still being developed and it is timely that a coordinated action to enable systematic learning, sharing of experiences, identification of good practices and reuse of solutions should be undertaken now, with the resultant economies of scale and reduction of costs at individual NSI level. In order to accelerate the adoption of AI/ML by National Statistical Institutes (NSI's) in the production of Official Statistics Eurostat issued a grant call to establish a One-Stop-Shop on AI/ML in Official Statistics.

The Central Statistics Office in Ireland brought together a consortium of 14 countries (see Appendix A for participating countries) to deliver the project. The coming together of the consortium will facilitate the scaling up or reuse of existing and new solutions, the standardising of methodology, understanding the AI/ML eco-system in statistics, the development of appropriate best practice and quality guidelines as well as providing access to shared resources such as a sandpit environment and training.

The project kicked off in April 2024 and will run for a period of 4 years. The rest of this paper will, therefore, describe the project and how it intends to consider quality at the heart of its work.

Figure 3-1
AIML4OS Project structure



3. Theoretical Framework

This section will describe the overall framework of the project and how it will be delivered. The design of the project delivery structure was based on standard project management approaches.

The project structure is centred round 2 broad concepts (Figure 3-1):

- **Use Case Work Packages** which are focused on looking at a specific problem which AI/ML may have a role in addressing or an area of AI/ML which may have a broader use within Official Statistics. These use cases will provide standards, methods, code and training which will be fed into the cross-cutting work packages so these can be collated centrally and provide some of the core resources being made available by the project, and
- **Cross-cutting Work Packages** which provide services on behalf of the overall project to both consortium members but also to the wider AI/ML community which the project is looking to both build and engage with.

The key method for delivering the project is through individual work packages as follows:

Cross-Cutting Work Packages

- WP1: Project management and coordination
- WP2: Communication and community engagement
- WP3: ESS AI/ML lab: Technical infrastructure and organisational setup
- WP 4: AI/ML state-of-play and ecosystem monitoring
- WP5: Standards, methodological and implementation frameworks
- WP6: Knowledge repository and training material

Use Cases Work Packages

- WP7: Use Case: AI/ML on earth observation data, satellite imagery
- WP8: Use Case: Editing focus - Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on editing
- WP9: Use Case: Imputation focus - Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on imputation
- WP10: Use Case: From text to code - Experiences and potential of the use of AI/ML for classifying and coding
- WP11: Use Case: Applying ML for estimating firm-level supply chain networks
- WP12: Use Case: Large language models
- WP13: Use Case: Generation of synthetic data in official statistics: techniques and applications

(See Appendix B for more details of each work package).

Quality will be considered in 2 main ways:

Quality Assurance of Work Packages

A review panel is being established to provide quality assurance for each individual use case. One or 2 reviewers with expertise in the particular topic of the use case will be allocated to the use case to carry out a peer review at the beginning, middle and end of each Use Case. The reviewers can also provide ad hoc advice and guidance during the delivery of the Use Case.

Development of Standards/Best Practice/Frameworks/Guidance

A key outcome of the project is to provide:

- Methodological and implementation guidelines for applying AI/ML in official statistics,
- Generalised knowledge and identified norms and best practices for developing and operating AI/ML-based solutions in official statistics, and
- Standards in applying AI/ML in official statistics. This will enable and facilitate the transition from development and experimentation of AI/ML-based solutions to industrialisation and production.

To achieve this the individual use cases will look to develop these in their individual work packages and feed these into a central repository to identify and provide overarching products and guidance to improve the overall quality of the use of AI/ML in official statistics as follows:

Standards

While artificial intelligence and machine learning methods have been tested in many statistical offices over the past ten years, the need for standardisation has become apparent with the increasing maturity of the applications on the one hand and the frequency of their use on the other.

This is not a one-size-fits-all for all AI/ML solutions, but standards clearly are valuable up to a certain level. Often, such standards do not have to be developed from scratch but can be derived from best practices in the statistical offices or from previous work (e.g., by the UNECE). The project will cover conceptual questions on quality aspects (including clarity, interpretability, transparency, data quality, auditability and ethics).

A standards manual for official statistics will be developed, refined and updated over the course of the project containing topics such as standard methods, quality aspects and interpretability. This is valuable for statisticians working in official statistics using new state-of-the-art ML and AI methods. Examples are confusion matrix basics and partial dependence plots.

Methodological Frameworks

Huge progress has been made in recent years in the area of AI/ML. As a result, a growing number of applications have emerged within the field of official statistics, that focus on the development of new products and services, on improving decision-making and on increasing the use of new data sources as rapid indicators.

However, blindly applying AI/ML techniques does not suffice when one wants to produce a high-quality output. It is clear that methods and instructions (a methodology) need to be developed to enable the application of AI/ML, in a trustworthy and reliable way. From a mathematical perspective, the methodology is aimed at producing unbiased, reproducible, and valid statistical outcomes for the target population for studies that usually start with so-called ‘found’ data. A methodological framework will be developed, refined and updated over the course of the project.

Implementation Framework

Available standards and methodological frameworks have the potential to greatly increase the sustainability of AI/ML applications in official statistics. For these to make an impact, it is critical that there are efficient and practical ways to go from theory to practical implementation.

The project will look to establishing efficient mechanisms to support the transition from experimentation to production. The success of this transition relies on streamlining and automating the machine learning lifecycle – a practice defined as Machine

Learning Operations (MLOps). To do this we will identify standard building blocks to support MLOps and will specifically provide guidance and templates on how to establish these in AI/ML labs. This will also describe the features AI/ML labs should have, in order to efficiently make use of best practices and standard building blocks.

Individual use-cases will be used to verify that the standards, methods and MLOps frameworks are practical and beneficial and that they help the transition from experimentation towards production, in a sustainable way whilst producing statistics of an appropriate standard of quality. The overall implementation framework will be developed, refined and updated over the course of the project.

4. Results

The project will look to present the results of its work at an annual conference for the AIML4OS as follows:

- NTTS 2025
- Conference 2026 (potentially Quality Conference)
- NTTS 2027
- Closing Conference Dublin 2028

This will be supplemented by production of papers/guides/frameworks, presentations/papers at relevant conferences, webinars and training. The material will be available via the Eurostat CROS platform which is currently being developed.

5. Conclusions

As a result of the delivery of the thirteen separate work packages identified by the consortium, the overarching expected outcomes are:

- the delivery of a framework for developing AI/ML solutions to be used in the context of producing official and European statistics,
- the provision of access for ESS staff and partners to established and proven AI/ML solutions/resources to be leveraged in the context of official statistics production,
- the encouragement of engagement from ESS organisations with AI/ML for innovation purposes and to facilitate their understanding and realisation of the benefits of AI/ML, and
- the delivery of economies of scale and resources through cooperation inside and outside the ESS and the acceleration from ideas regarding AI/ML to actual production.

If these outcomes are achieved it is expected that the project will provide a significant set of resources which will provide guidance, methods and standards to produce high-quality statistical products through the application of AI/ML.

Appendix A. Consortium countries

- Statistics Austria
- Statistical Service of Cyprus
- Statistics Denmark
- Institut National de la Statistique et des Etudes Economiques - INSEE, France
- Statistisches Bundesamt - DESTATIS, Germany
- Central Statistics Office - CSO, Ireland
- Istituto Nazionale di Statistica - Istat, Italy
- Statistiques - Luxembourg
- Statistics Netherlands - CBS
- Statistics Norway
- Statistics Poland
- Instituto Nacional De Estatistica, Portugal
- Statistical Office of the Republic of Slovenia
- Instituto Nacional de Estadística - INE, Spain
- Statistics Sweden - SCB

Appendix B. Work package description

Table A.1
Work package descriptions

Work Package Name	Work Package Country Lead	Work Package Description
Coordination	CSO	To support, coordinate and facilitate the consortium activities in the project, while also taking measures to ensure adequate quality Levels of the outputs that can be realised and remain sustainable across the whole ESS.
Communication	Istat	Building communities around open-source solutions developed and maintained by ESS members. Sharing ideas, experiences, success stories and lessons learned to stimulate innovation-based use of AI/ML External communication and dissemination of the results, in particular within the ESS.
Platform/Technical Infrastructure	INSEE	Deliver an AI/ML testing environment (sandpit) which can be used by the use case work packages (WP 7 to WP 13) for experimentation and development. Provide documentation and technical support for on boarding the use cases. Provide documentation and guidance to allow NSIs to create their own AI/ML sandpit ("platform as a package").
State-of-play & Ecosystem Monitoring	DESTATIS	Obtaining evidence of the state-of-play of the use of AI/ML in the ESS and beyond i.e., to landscape and maintain a comprehensive picture of AI/ML developments and activities for statistics. Identify the needs of the NSI's in relation to AI/ML.
Standards & Methodologies	DESTATIS & CBS	To develop methodological and implementation guidelines for applying AI/ML in official statistics. To generalise knowledge and identify norms and best practices for developing and operating AI/ML-based solutions in official statistics. To achieve standardisation in applying AI/ML in official statistics. This will enable and facilitate the transition from development and experimentation of AI/ML-based solutions to industrialisation and production.

Work Package Name	Work Package Country Lead	Work Package Description
Knowledge Sharing	STATISTICS POLAND	<p>Provide support and guidance for the integration and maintenance of relevant AI/ML-based solutions in ESS organisations, through the provision of training and active efficient support.</p> <p>Provide guidance for selecting/developing AI/ML models for specific statistical products.</p>
AI/ML on earth observation data, satellite imagery	CBS	<p>Can existing AI/ML models using Earth Observation be generalised over space (countries) and time (periods) and under what conditions?</p> <p>If feasible, methodological implementation guidelines will be developed.</p>
Editing focus - Statistically valid and efficient editing in official statistics by AI/ML	DESTATIS	<p>Discovering the opportunities of the use of AI/ML in terms of automation, efficiency, and quality improvement of the editing and imputation process (with a focus on editing).</p> <p>Exploring this in practical examples with a focus on editing, without forgetting the quality standards of official statistics.</p> <p>Investigating the effects on the corresponding workloads for editing in the statistical offices.</p>
Imputation focus - Statistically valid and efficient imputation in official statistics by AI/ML	INE	<p>To embrace different situations in statistical production where AI/ML techniques can be used for imputation.</p> <p>Early imputation/timeliness – imputation of values which will be known and measured after some process steps have been completed (e.g. after collection or after editing during collection).</p> <p>Post-collection imputation/accuracy – considers situations which are typical in both item and unit non-response in usual production conditions by covering imputation of values which will never be known, but related to units which are in the sample.</p> <p>Imputation beyond the sample/granularity – considers situations in which we shall impute population frames or datasets (e.g. administrative sources) which are not in a sample by reconstructing information (at least partially) for a whole population using data from a much smaller dataset.</p>
From text to code - Experiences and potential of the use of AI/ML for classifying and coding	DESTATIS	<p>The use of AI/ML to assign textual descriptions from potentially different sources to internationally agreed classification schemes (e.g. NACE, COICOP, ISCO, ...) of official statistics is investigated, including the challenge of texts in different languages as a barrier to exchanging data.</p> <p>A recommendation for a Text2code pipeline (including, e.g., pre-processing).</p> <p>First solutions for some of the special challenges such as different languages.</p>
Applying ML for estimating firm-level supply chain networks	CBS	<p>Develop Machine Learning (ML) models that can be used to estimate population-scale firm-level supply chain network datasets with only data on firm-level characteristics and variables</p> <p>To train the ML-models on population-scale network datasets based on detailed administrative data on user/supplier transaction</p> <p>To develop a methodology for estimating weights of the links, using the National Accounts' input/output-tables as a benchmark.</p>
Use of generative large language models in statistics	SCB	<p>Explore the 2024-2027 to-be existing LLM ability to integrate in production and support of official statistics.</p> <p>Explore what benefits fine-tuning of existing LLM will have on the use-cases.</p> <p>To support the above objectives, define the architectural enablers and constraints in relation to the use of LLM</p>
Generation of synthetic data in official statistics: techniques and applications	Istat & STATISTICS AUSTRIA	<p>Investigate different AI/ML algorithms to generate synthetic data in official statistics domains balancing utility and privacy.</p> <p>Deliver PoC to generate synthetic data in official statistics, applying AI/ML techniques.</p>